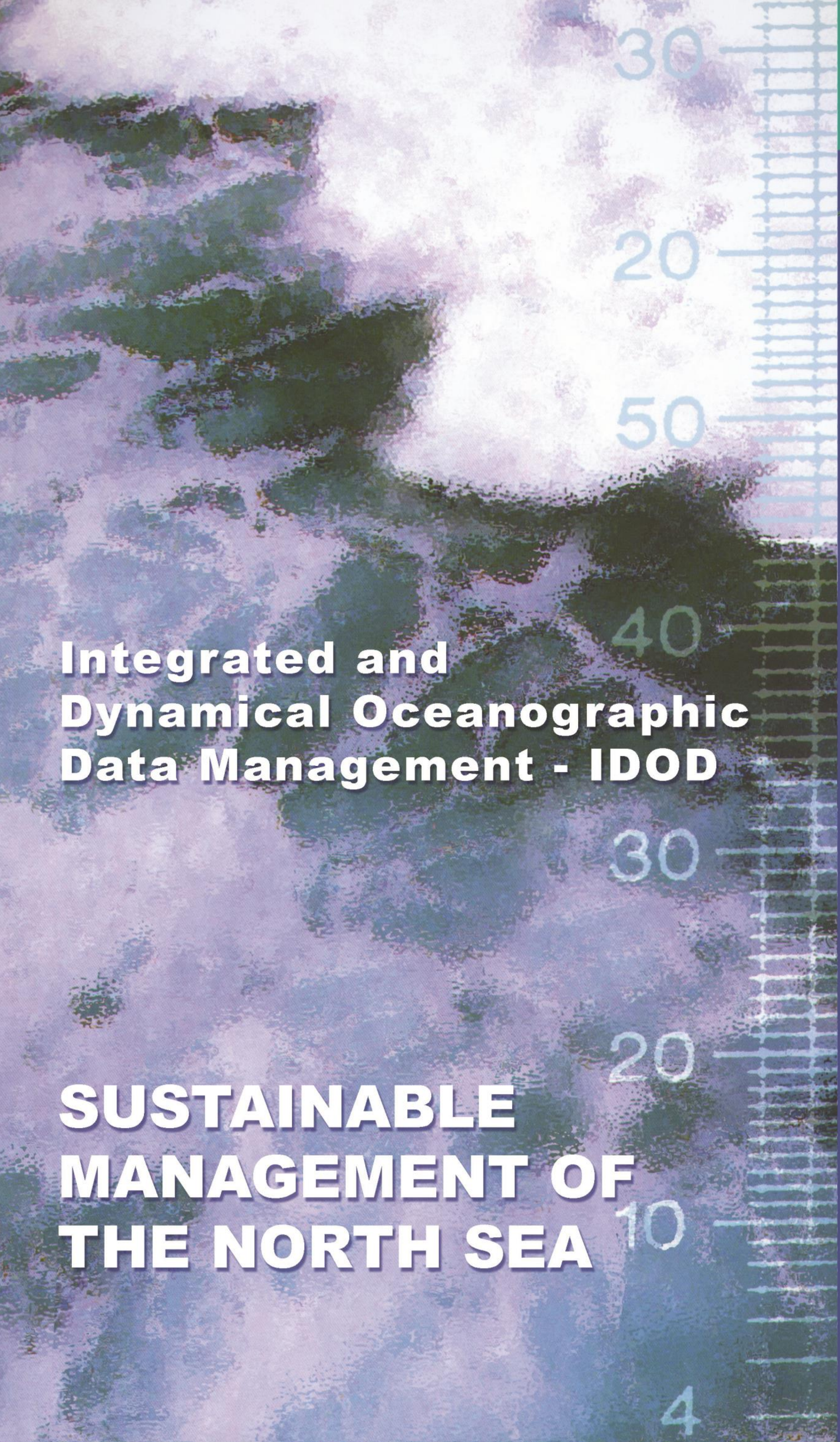


Integrated and Dynamical Oceanographic Data Management - IDOD

SUSTAINABLE MANAGEMENT OF THE NORTH SEA

SCIENTIFIC SUPPORT PLAN FOR A SUSTAINABLE DEVELOPMENT POLICY

BELGIAN SCIENCE POLICY

SCIENTIFIC SUPPORT PLAN FOR A SUSTAINABLE DEVELOPMENT POLICY (SPSD-I)

Programme: "Sustainable management of the North Sea"

Integrated and Dynamical Oceanographic Data Management
IDOD

Dr. ir. G. Pichot
Management Unit of the North Sea Mathematical Models and the Scheldt estuary (MUMM)
Gulledelle 100, 1200 Brussels

Prof. Dr. J.–P. Donnay
Université de Liège, Laboratoire "Surfaces"
allée du 6 Août 17 Bât B5 Sart Tilman, 4000 LIEGE

Dr. J. Van Dyck
Kathiolieke Universiteit Leuven, University Centre of Statistics (UCS)
De Croylaan 52B, 3001 Heverlee

D/2004/1191/45

ABSTRACT

The goal of the IDOD project was to provide the federal government, the scientific community and other users with an up-to-date tool for collecting, managing and analysing marine scientific data.

The resulting "marine information system" is hosted by the Belgian Marine Data Centre (BMDC), a team within the Management Unit of the Mathematical Models of the North Sea (MUMM). The BMDC committed itself to keep the IDOD information system alive and evolving. A remote user interface is available online at http://www.mumm.ac.be/datacentre.

The project faced all the aspects of modern scientific data management. A major challenge was to establish a fruitful dialog with the data providers. This has been done through extensive discussions in the *Users committee* and during bilateral meetings. The topics that have then been clarified range from the principles (in order to write down a standard common "Rights and duties" agreement) to the very technical and scientific details, specific to each data set.

A substantial effort has been put on the definition of guidelines for ensuring the data quality throughout their way from the field to the data centre. This has resulted, for instance, in the development of a "On-board registration of samples" computer programme, in a check list of meta–information to document the data or in the definition of a "common layout" for reporting data sets to the data centre.

On the technical side, the variety and complexity of the data to be stored and made available for further use lead us to elaborate a complex and robust data base scheme, after an in–depth conceptual analysis. The database itself is implemented using the ORACLE technology.

In parallel, a set of dedicated analysis and visualisation tools has been developed. They help the users and the data managers to tackle the data in their space and time dimensions and also allow cross–analysis between different parameters.

The primary providers of data for the information system were the various teams participating in the (first) Programme for a sustainable development of the North Sea, financed by the Belgian Federal Office for scientific, technical and cultural affairs. These data are being complemented by data from other (earlier) scientific programmes and from governmental surveys.

The project has been performed by MUMM, in partnership with the SURFACES laboratory (University of Liège) and the University Centre of Statistics (University of Leuven).

1. FOREWORD

The IDOD project was financed by the Belgian Federal Office for scientific, technical and cultural affairs. Its goal was to provide the federal government and other users with an up-to-date tool for collecting, managing and analysing marine scientific data. The first data to be considered were those resulting from sampling and measurements made by the various teams participating in the (first) Programme for a sustainable development of the North Sea.

The main scientific result of the IDOD project is thus the development of an information system for marine data, made of several data banking and data analysis tools. A remote user interface to it is accessible through the Internet at :

http://www.mumm.ac.be/datacentre.

We encourage the interested reader to browse through that website for a better understanding of the results obtained during the project.

2. PARTICIPANTS

Co–ordinator – Promoter : Dr. Ir. G. Pichot (MUMM, RBINS)

Co–workers : K. DECAUWER, M. DEVOLDER, S. JANS, J. LAZAROU, L. SCHWIND, S. SCORY

Promoter : Pr. Dr. J.–P. DONNAY (SURFACES, ULg)

Co–workers : M. BINARD, Y. CORNET, F. MULLER, J.–Ch. SAINTE

Promoter : Pr. J. BILLIET (UCS, KUL)

Co–workers : Pr. J. VAN DYCK, B. PLEVOETS, G. DIERCKX, F. VASTMANS

3. TIME LINE

The ultimate aim of the IDOD project was to develop, to manage and to promote a marine environmental database in order to ensure a smooth and scientifically sound data flow between the data producers and the end users.

To reach this aim, different tasks were consequently executed in order to bring the project to its successful conclusion. The time line of the core tasks reads as follows.

In 1998, the tasks were mainly concentrated on the elaboration of the conceptual scheme of the IDOD database and on the translation thereof into a prototype of the physical database.

In 1999, the continuous update of the inventory of data sets and the screening procedure for incoming data sets, including their documentation, began to be operational. The analysis of the first data sets received in the frame of the Programme "Sustainable Management of the North Sea" lead to updates and changes in the conceptual scheme for the seawater database, followed by their implementation in the prototype. Other data sets formed the basis for a first draft of the conceptual schemes for the plankton and sediment subsets of the database. Concurrently, thanks to the availability of actual data sets, the first developments of the spatial and statistical analysis tools could occur. In order to clarify the conditions of data transfer and access to the database, a convention between the data providers, the OSTC and the IDOD project was prepared.

In the beginning of 2000, it was decided to move from the Access prototype into an Oracle relational database that would best apply to the needs of the IDOD-project. Therefore, during that year, much effort focused on the in–depth analysis of the full specification of the system, on the training of the people dedicated to it and the physical implementation of the database with the related validation rules, an internal user interface and different In/Out modules. Considering the extent of the task MUMM also decided then to hire external assistance.

During the year 2001, the implementation was continued and part of the effort was devoted to the import of data in the database and consequently on the testing of the previously developed procedures for import and export of data. Another important, if not the most important, task in 2001 was the description and the development of the procedures for the interactive data access for the end-users (via the web site, using the access rules defined in the Convention). In parallel, the procedures and modules developed by ULg–SURFACES (spatial analysis and visualisation) and KUL–UCS (statistical analysis and quality control) were also implemented and tested.

A first full–scale demonstration was organised at the end of January 2002, during a two-day symposium devoted to the outcomes of the Programme "Sustainable

Management of the North Sea”. The last six months of the project were devoted to fine tuning and integration of the various modules –together with the as continuous as possible feeding of the data base–. The “IDOD information system” was finally officially launched in July 2002.

4. MATERIALS, METHODS AND SIGNIFICANT RESULTS

In this chapter we give a general view of the major steps performed to achieve the objectives and of the main technical and scientific results. Details are given in annex when necessary.

"IDOD" was an applied research project. Its major result is an information system "that works". The classical splitting of the reporting of a science project ("Materials/Methods/Results/Discussion") does not fit very well to what has been done and how. We therefore will follow the itemisation of the technical annex of the contracts. Percentages of workload devoted to each task and subtask also refer to the original figures.

4.1. Inventory of the data sets

(8% : MUMM, 90%; SURFACES, 5%; UCS : 5%)

A.1 Inventory of existing data sets

(MUMM)

As a starting point an exhaustive list of data sets or data sources available at MUMM was made (monitoring and input data, bathing water quality, NOWESP, mathematical models, airborne surveys, geographical data, data collected by the automatic *Oceanographic Data Acquisition System* –onboard the RV Belgica–, *etc*.). Results of previous national/federal research programmes were also identified ("Projet Mer/Projekt Zee", dedicated actions, Programme "Marine Sciences", ...)

As part of the project but also in the frame of a co–operative action at the European level, the data catalogues EDMED/EDMERP (European Directory of Marine Environmental Data/Research Projects) were also updated to reflect the current situation in Belgian institutes and universities. Up to 123 on–going or past research projects and their resulting data sets were identified.

A.2 Identification of future data sets

(MUMM)

"Future data sets", *i.e.* data sets to be generated during the course of the Programme "Sustainable Management of the North Sea", were *a priori* identified, starting from the information provided by the various involved research teams. See Annexes 1 & 2 of the Intermediate Joint Scientific Report #2 (Annex 1 to the present report) for the identification of the data sets and for a table with all parameters measured in the different projects.

A.3 Definition of terms and conditions of data transfer or access

(MUMM)

In order to clarify the conditions of data transfer and data access, a forum with all the concerned parties took place by the end of 1999. It resulted in an agreement on a formal convention between the OSTC, the IDOD data centre and the data providers. This convention is given in Annexes 8 & 9.

A.4 Inventory of relevant data required as meta-information

A.4.1 Definition of the needs regarding geo–referencing (SURFACES)

This aspect appeared to be less crucial than expected during the preparation of the project. Hence, the meta–data required by geo–referenced information, besides some basic information to identify the "spatial object" resume to the geographical reference system.

A.4.2 Definition of the needs regarding quality control (UCS)

The relevant meta–data for statistical quality control derive from the test to be applied to the data (see B.1.2) and have been defined accordingly.

A.4.3 Inventory of relevant data (MUMM)

Guidelines for data documentation have been developed starting from equivalent recommendations set up at the European level by the MAST Data Committee,. These guidelines define a succession of steps to be performed or items to be checked before data sets are reported to the data centre, in order to ensure they are consistent and adequately documented. The required meta–information covers, for instance, the description of the sampling and measurement methods, the environmental conditions during sampling, the quality controls performed by the data provider and so on. See Annex 1 for further details.

4.2. Set-up and improvement of the incoming flow of data

(19% : MUMM, 45%; SURFACES, 10%; UCS : 45%)

Apart from defining the practical aspects of actual data reporting to the data centre, the major task was to define and implement the quality control of the data, which is the foundation of the credibility of a database. Both practical aspects and QC procedures for the meta–information were handled by MUMM.

B.1 Design of the quality control procedures

(UCS)

UCS concentrated on the concept of the quality control procedures of the data themselves. The choice was made to implement a Bayesian combination of the results of a series of tests that can be applied to each data point. Furthermore, because of the different natures of the data types that may occur, it was decided to implement a general shell to support a variety of tests, rather than implementing a specific control procedure. Three general types of tests were selected:

1. verification of the data against the marginal distribution with parameters that may vary as a function of time and location,
2. verification of the data against other measurements of another nature at the same time and place,
3. verification of the data against the same type of data at another position within a limited time window.

Details on the methods used and on their actual implementation are to be found in Annex 11.

B.2 Development of the quality control procedures

B.2.1 Prototyping (UCS)

A prototype of the statistical quality control has been designed at an early stage of the project and was finished at the end of the year 1999. At that time the database was not yet operational and the linking of this prototype to it was thus impossible. Tests were therefore performed on an existing subset of data.

B.2.2 Full–scale implementation (MUMM)

The objective of this task was originally meant to be broader than the implementation of the statistical quality control procedures alone, hence the word "full–scale". It has been however agreed between the concerned partners that UCS would concentrate on these statistical tools (including their implementation as a "pluggable" module) while MUMM would manage all the other aspects (see B.3).

The statistical analysis programme was developed by UCS in a second stage of the project. The statistical quality control procedures making use of it, their integration into the system could only be addressed at the end of the project. Tuning of the whole is still on–going.

B.3 Implementation of the quality control procedures

(MUMM)

The improvement of the incoming flow of data was implemented on different levels :

- a system (in MS–Access) was developed for the registration of samples onboard the RV Belgica. This system allows the data originator to enter in an easy way all information about a sample and to automatically extract, once the sampling time is entered, the geographical co–ordinates and other relevant parameters measured by the automatic data acquisition system (ODAS) during the sampling event. This programme has been installed and was tested during routine monitoring campaigns. Due to reasons lying outside the scope of the IDOD project, it is not yet routinely used by all the teams using the RV Belgica.

- A "common layout" to be used by the data providers for the reporting of data to IDOD was developed in MS–Excel. The guidelines for data documentation were implemented in this common layout. This tool facilitates the work of the data manager when validating and importing the data in IDOD and reminds the data-provider of the information to be provided.
- The delivered data sets are manually checked for completeness and clearness by the IDOD data managers. Often the data providers still have to be contacted for missing or explanatory information.
- Based on the common layout, procedures were developed to import data in the IDOD database. These procedures start up a number of checks (validation rules : syntax checks, external value checks, combination of fields checks, position checks, …) at the entrance of data in the database.
- The last level is the statistical quality control (SQC). This tool is used when the data are in the database and is in the first place an external value check. After the values are tested in different test schemes (see B.1) and then flagged (acceptable, exceptional, ... : no rejection of values).

B.4 technical aspects bound to geo–referencing

(SURFACES)

As mentioned under A.4.1 geo–referencing appeared not to be a crucial issue with respect to the data under consideration. Quality of, for instance, co–ordinates of sampling stations is treated as any other measurement.

B.5 Upload of the data sets

(MUMM)

This task actually started when the core database was ready. Due to the lack of time and manpower, not all the data sets generated in the frame of the Programme could be incorporated by the end of the IDOD project. MUMM, however, took appropriate measures to keep this task running as a structural mission.

As mentioned under B.3, there is a long way before a data set reported to the data centre actually enters the data base. This is a consequence of the strict application of our requirements concerning documentation and quality control. But we are convinced (and think that most of the data providers now also are) that such high–level quality standards give an important added value to the data sets and hence to the scientific works based thereon.

B.6 Contacts with the data providers

(MUMM)

This task is summarised under G.2 (Co–ordination).

B.7 Feasibility study of connecting to and from the RV Belgica

(MUMM)

The feasibility as been performed at the early stage of the project. It concluded that full interactivity was not yet necessary for the purpose of "better science". An e-mail connection has been set–up that allow near real–time transfer of information. The parameters measured by ODAS, for instance, are presently transmitted to the data centre once a day and are automatically processed according to the needs (see, *e.g.*, http://www.mumm.ac.be/EN/Monitoring/Belgica/where.php).

4.3. Design and implementation of the database

(10% : MUMM, 45%; SURFACES, 40%; UCS, 15%)

N.B.: The tasks described in the technical annex to the contract, and the concepts and the structure used therein, correspond to a specific development methodology. In the course of the project, taking into account the skills and experience of the developers and, also, the production environment, we decided to follow another methodology.

The differences are mainly a question of terminology and of the phasing of the various subtasks. Using different words do not change the nature of the tasks to perform to achieve the goal...

Technical information given in the Annexes do not, therefore, always refer to the same subdivision of tasks as in the technical annex but reflect the way things were handled. Hereafter however, we keep the original structure, where possible.

C.1 Conceptual design

C.1.1 Methodology (SURFACES)

Some preparatory work on the conceptual scheme for the concentration values in sea water has been done using DB-Main. (http://www.info.fundp.ac.be/~dbm/). This lead to a preliminary visual clarification of the basic entities and relationships in the data base.

C.1.2 Definition and formalisation of the needs (MUMM)

This task has been performed through in–depth discussions with the data providers and the potential users. It has been an iterative process, the conceptual scheme being more and more refined as our knowledge of the various data types and data structures and of the needs increased.

Basically, we started with the knowledge on data banking existing at MUMM (former *Monit–B* database, holding the measurements made to fulfil Belgian duties on sea water quality monitoring at international level). This, together with the identified flaws of the existing system, lead to the definition of a first conceptual scheme of a relational data base, which was then implemented using MS–Access and fed by a subset of the 'monitoring' data set.

With the help of that prototype, we could, on one hand, give a starting point to the partners in charge of developing the spatial and the statistical modules and, on the other hand, further the analysis with, a.o., the data providers and the potential users. Finally, the latest major refinements occurred when the first actual data sets were transferred to the data centre and analysed.

This process focused on the conceptual design of a structure devoted to hold concentration values in sea water. Thanks to its completeness and careful design, it served then as a basis for developing similar schemes for, *e.g.*, data on sediments, on concentrations in sediments and biota, information on biodiversity, ...

C.2 Logical analysis

C.2.1 Methodology and technical aspects bound to geo–referencing (SURFACES)

As already mentioned above, this aspect rapidly appeared to be less crucial than expected during the preparation of the project. As a consequence, no needs for using a methodology specifically designed for geo–referenced data were identified.

C.2.2 Technical aspects of the statistical analysis of the data (UCS)

This task concerned primarily the design of the communication between the SQC program (statistical quality control), the SAT program (statistical analysis tool) and the IDOD database. Protocols for communication were therefore designed. For communication between SQC and SAT a separate databank (SATBUFFER) is used. Details on the links between the programs and the database are described in Annex 11.

C.2.3 Definition and formalisation of the needs (MUMM)

This phase dealt with the formal description of the concepts and the functionalities, defined under C.1, in order to ensure their completeness and their coherency. The data base structure was thus further described, together with the indispensable modules for importing, exporting, querying, checking, ..., the data. The integrity rules of the data base were also defined at this stage.

C.3 Implementation

C.3.1 Implementation of the techniques bound to geo–referencing (SURFACES)

As already mentioned above, this aspect rapidly appeared to be less crucial than expected during the preparation of the project. As a consequence, no requirements bound to geo–referenced data had to be taken into account for the implementation of the data base itself.

C.3.2 Implementation of the statistical techniques (UCS)

This task concerned primarily the implementation of the protocols earlier designed in C.2.2. The various links have been implemented and tested. Details on the various links between the programs and the database are described Annex 11.

C.3.3 Global implementation (MUMM)

The global implementation of the system could then occur. As it had been decided to move to an Oracle RDBMS for the implementation of the data base itself, the conversion of the concepts into modules and procedures has been made using associated tools (*e.g.* Oracle Designer). The documentation given in Annex 10, reflecting the status of the data base at the time of writing this report, are automatically generated by such development tools.

A full overview of the high-level analysis (process modelling, function hierarchy, entity relationship diagram, phases and tasks to perform and budgetary estimate), analysis phase and training courses is described in Annex 3.

C.4 Build phase (MUMM)

After the long-lasting task of describing the design of the IDOD database, the physical database was built. Accordingly, a whole package of the defined procedures was developed (validation rules, triggers, import and export procedures, …) and tested.

4.4. Data analysis tools

(19% : MUMM, 20%; SURFACES, 40%; UCS, 40%)

D.1 Selection of the tools

D.1.1 Aspects bound to geo–referenced data (SURFACES)

After screening of the needs and evaluation of the (commercial) products and tools, two paths were identified : ad hoc processing, using tools combined manually, and the development of a general (but simpler) visualisation tool, to be made available through the Web.

D.1.2 Aspects bound to statistical processing and quality control (UCS)

It has been decided to build a statistical analysis module presenting in a coherent way several statistical tools to the user. The core of that module had to be designed in such a way that it could also serve as 'computing engine' for the statistical quality control programme. The main characteristics of the data to process to be taken into account was their time–dimension and their spatial distribution (mainly 2D).

D.1.3 Justified choice of the tools (MUMM)

From the first inventory made by the partners, MUMM defined the directions to be followed by the design of the analysis tools :

- the tools have to be versatile, user–friendly and robust,
- they must connect easily with the data base,
- more complex analyses would be performed by the data centre, as a service.

D.2 Design of the tools

D.2.1 Tools pertaining to geo–referenced data (SURFACES)

A technical comparison showed that, at the time of selection, the functionalities offered by the ESRI products (ARcView, ArcIMS, ArcExplorer) were the best suited to fit the needs. It has then be decided to develop one module that allows the remote user to interpolate data extracted from the data base and another one to draw a value profile across that data subset.

D.2.2 Tools bound to statistical processing and quality control (UCS)

The choice has been made to develop the statistical analysis tools as modules in the *S-Plus* software. A number of analysis techniques were retained as of interest. Techniques that have been implemented are: "summary statistics", "summary plots", "correlation matrix", "scatterplot matrix", "trend fitting", "normality check", "multiple regression test", "subset regression", "variogram calculation", "SQC distribution check" and "SQC regression test". Some other functionalities that were thought of interest (*i.e.* principal component analysis, ...) could not be implemented because of lack of time.

The analysis techniques have been implemented in a user-friendly interface (that hides the underlying S-Plus software) that can be operated within a browser. The actual implementation of the various options has taken considerable effort because of the complications in feeding the input from the browser to S-Plus and showing the results of the operations. The SAT (statistical analysis tool) that followed from this effort is described in Annex 11.

D.3 Implementation of the tools

(MUMM)

Tools were already implemented and tested to some extent by the partners when they were delivered for incorporation into the information system. Smooth integration (from both the manager and the user's point of view) yet requested some changes and improvements.

4.5. Models

(MUMM : 60%, SURFACES : 15%, UCS : 25%)

This is definitely a point where the objectives of the project where not met. Although results of the relevant mathematical models (namely hydrodynamic models, providing for information on currents, elevation and sea state for the whole domain of interest) are available on request to the users, they are not yet integrated in the information system.

The two main reasons are :

- the shift in resources that occurred during the project, to take into account the available manpower (and the available skills) and the higher than foreseen workload to achieve the main objective of the project,
- a strategic decision that occurred in the course of the project to develop and implement at MUMM a new reference hydrodynamic model.

The second reason made it almost useless to develop and, even more apply, specific quality control procedures and data base structures without knowledge of the definitive output structure of the models.

We are still convinced however that offering access to synoptic hydrodynamic fields together with more 'common' oceanographic data is an asset for the information system. This task will undoubtedly remain a priority objective for the data centre.

4.6. Access to the data, derived products and valorisation

(MUMM : 50%, SURFACES : 40%, UCS : 10%)

F.1 Market analysis

(MUMM)

The very first users of the information system are the data providers themselves. The system must indeed provide them with functionalities that help them to give and added–value to their data, together with safekeeping facilities. We started therefore by listening to this category of users. The fulfilment of their needs was a prerequisite for the establishment of a fair relationship between them and the data centre.

A second category of users to be considered was the target audience of the results of the whole Programme, i.e. the administrations and the policy makers concerned by the sustainable management of the marine resources and, especially, those of the Belgian waters.

These potential users and the data providers were invited to participate in the accompanying committee of the project. Fruitful discussions lead to a list of requests that were later completed by a review of the requests for data MUMM had received before the period of the project and by questions asked by visitors of MUMM's general website.

F.2 Selection of the most adequate publishing means

F.2.1 Global strategy (MUMM)

The list of identified needs and requests were split into two categories : those of general interest that could be addressed by developing standard, multi–purposes, tools and those that would clearly need to be answered on *ad hoc* basis.

For the first category, it has been decided to give people an interactive, but secured, access to the data and to some high–level data analysis tools *via* the web.

F.2.2 On–line services and specifics of the geo–referenced data (SURFACES)

Two modules have been developed and made available online : one that allows the remote user to interpolate data extracted from the data base and another one to draw a value profile across that data subset.

F.3 Reporting to international bodies and contribution to international assessments

(MUMM)

Automatic procedures were developed for the yearly reporting of monitoring and other relevant data in the required format to ICES in the frame of the Belgian international commitments for the "Joint Assessment and Monitoring Programme" (JAMP) and the

"Nutrient Monitoring Programme" of the Oslo and Paris Commission (OSPAR). These procedures were progressively used in 2000, 2001 and 2002 to effectively report the data of the year before. They are now part of the routine procedures.

MUMM participated actively in the preparations for the Quality Status Report 2000 for Region II, published by the OSPAR Commission and the data gathered by the programme and the tools developed at that time in the frame of the project were extensively used.

The first actual test case for the newly born "IDOD Information System" was performed during the Summer 2002. For the first time, the efficiency of the IDOD database and the user tools were tested in practice, in collaboration with ULB-ESA, in order to provide the Belgian national contribution to the first application of the "Comprehensive Procedure", *i.e.* the application of the *Common Assessment Criteria for the Eutrophication Status* of the OSPAR Maritime Area as agreed by OSPAR in 2002. This report is given as Annex 12.

F.4 Regular issue of various kind of products

F.4.1 Global strategy (MUMM)

The strategy was to inform the concerned people on the progress of the IDOD project and to give them progressively access to data, products and tools :

- 4 newsletters informed them of the issues and the choices made during the development phase (Annexes 4 to 7);
- a presentation sheet has been printed and is being distributed at chosen events (Annex 13);
- a poster has been presented at the OSTC Symposium on the first Programme 'Sustainable Management of the North Sea' (January 2002 – Annex 13);
- The database and the information system are presented at relevant symposiums – *e.g.* CoastGIS'99 or 'Colour of the Ocean Data' (2000)– and in adequate forums – *e.g.* the European Sea–Search initiative–.
- A few monographs were issued (*e.g.* 'Statistical analysis of the 1977-1996 MUMM monitoring data for seawater', KUL–UCS in co–operation with MUMM). The publication of such monographs could only start when there were enough data in the information system and when the data management team could be freed from the data base development activities. It is thus to be expected that such an activity will be more intensive in the near future.
- The results of the assessment of the eutrophication level of the Belgian coastal zone

(using the OSPAR criteria) based on combined levels of oxygen, dissolved inorganic nitrogen and dissolved inorganic phosphorus together with the chlorophyll-a concentration, taking into account the season of the year;

- Different products were prepared for the website. Many of them will be updated regularly and new products will continuously be added.
  - Reference documents on information in the IDOD database (description of codes, output formats, …);
  - Interpolated maps on salinity and temperature for a series of field campaigns;
  - Processing and publishing of near-real time data, e.a. latest information on the position of the R/V Belgica and the sea temperature;
  - A series of "ready-to-use" thematic maps. These offer the user the possibility to have a quick look at basic parameters, such as salinity and temperature, and their evolution over time. Depending on the frequency of the available data, maps are made on a monthly, seasonal and/or yearly basis;
  - A series of "static" thematic maps (natural areas, human activities);
  - …

F.4.2 Specific aspects of geo-referenced data (SURFACES)

A few exploratory studies were performed, using more complex methods than those made available to the remote user. Their results were published as contribution to the intermediate scientific reports or, synthesised, in the Newsletters (Annexes 4–7).

F.4.3 Specific aspects bound to statistical analysis (UCS)

Manuals have been produced for the various programmes developed for supporting the statistical analysis and the statistical quality control. The methods have been applied to specific subset of data and the results were *published* as internal reports and as contribution to the intermediate scientific reports or, synthesised, in the Newsletters (Annexes 4–7).

4.7. Task G : Co–ordination

(10% : MUMM, 80%; SURFACES, 10%; UCS, 10%)

In this project the co–ordination was important at several levels. Not only had the partners to build, together, a coherent and integrated information system but there was a strong need to establish good relationships with the data providers and, also, with the future users of the information system.

G.1 Internal co–ordination

The internal co–operation materialised in numerous plenary and bilateral meetings, document exchanges and discussions around demonstrations of the prototypes of the various modules.

G.2 Co–ordination with the teams active in the Programme

There were also numerous contacts with the data providers. Initially, they aimed at :

- describing, as far as possible, the data and data structures to be generated during the programme and, therefore, to be transferred to the data centre,
- establishing the rights and duties of all the parties involved (OSTC, data provider, data centre), including the access rights to the data for third party users (Annexes 8 & 9).

This initial phase was successfully completed by the end of 1999.

Afterwards, contacts with the data providers reduced mainly to bilateral contacts between the data centre and the designated 'Project (or Team) Data Managers', in order to get data sets and to have them matching the quality standards.

The table below gives an overview of the state of data transfers from the teams participating in the Programme and the data centre at the end of the project. Contractually, teams were supposed to deliver their data by the end of March following the year were the measurements were made.

Table I : Data sets delivered to the data centre (Status June 2002)

PROJECTS	Laboratory	1997	1998	1999	2000	2001
Biogeochimestry						
	UA/UIA - Prof. Van Grieken	N.A.	√	√	√	√
	VUB - Prof. Baeyens	N.A.	√	√	√	—
	RUG - Prof. H. Van Langenhove	√	√	√	√	—
	ULB - Prof. R. Wollast	√	±	—	—	—
AMORE						
	ULB - Dr. Ch. Lancelot	√	√	√	±	—
	VUB - Prof. M.-H. Daro	N.A.	√	±	—	—
	MUMM - Dr. G. Pichot	√	√	√	√	±
ICAS						
	ULB - Dr. Ph. Dubois	N.A.	√	√	√	N.A.
	UMH - Prof. M. Jangoux	N.A.	√	√	√	N.A.
	UMH - Prof. R. Flammang	N.A.	√	√	√	N.A.
Biodiversity						
	RUG - Prof. M. Vincx	√	±	√	—	—
	IN - Prof. E. Kuijken	√	√	√	√	√
	KUL - Prof. F. Ollevier	√	√	√	—	—
Pathology and ecotoxicology						
	ULG - Prof. J.-M. Bouquegneau	√	√	√	±	±
	ULG - Prof. F. Coignoul	√	√	√	±	±
	VUB - Prof. C. Joiris	√	√	—	—	—
	IN - Prof. E. Kuijken	√	√	√	√	√
MARE-DASM		N.A.	N.A.	N.A.	N.A.	N.A.

—	Data set + meta-information not received
±	Data set + meta-information received incomplete
√	Data set + meta-information received complete

G.3 Co–ordination with stakeholders

On request of the OSTC the accompanying committee was soon transformed into a "Users Committee". This was not only a naming change but also, implicitly, a change in the "Terms of Reference" of that Committee. The Committee, made of representatives of the scientific teams participating in the project and of representatives of several public authorities (both at federal and regional levels) was convened several times in order to ensure that the developments of the information system were fitting the needs of its future users. The official release of the first operational version of the information system has been launched during an extraordinary session of the Committee in July 2002.

G.4 Viability of the information system after the end of the project

Our efforts went in two directions. The first one deals with the technical viability of the system. To meet this objective software and hardware were carefully chosen and the application–specific pieces of software well–documented.

The other direction deals with the people that run the information system. In order to keep the data banking and analysis activities running and improving, the so–called 'Belgian Marine Data Centre (BMDC)' has been created inside MUMM. It is made of a small team of scientists that is responsible for the data management activities, including the maintenance and the improvement of the existing tools. In this context, the "IDOD Information System" is, of course, a priority activity.

In order to ensure its viability on the long term, besides the maintenance and the upgrade of the software and hardware infrastructures which are now considered as being part of MUMM's structural duties, the BMDC needs to find external funding.

The team, therefore, took several initiatives. First it established its recognition as a reference centre at national and international levels : the BMDC is the *National Oceanographic Data Centre* designated by the Belgian federal government to the IOC/IODE. The BMDC is also the Belgian partner in the European programme "Sea–Search II" (as it was in "Sea–Search I"), a programme gathering more than 30 oceanographic data centres in order to set up improved common practices and expand data catalogues. The BMDC team is also active in the data committee of the ICES.

As a second step, the BMDC team promotes its skills by becoming a partner in selected projects, at national (follow-up's of the Programme 'Sustainable Management of the North Sea', digitisation of some collections of Royal Belgian Institute of Natural Sciences, ...) and international (*e.g.* "REVAMP", an European project on the validation of Envisat data for Chlorophyll products in North Sea Coastal waters, where the BMDC is "Project Data Manager") levels.

5. RESULTS

As underlined above, the major result of this project is the existence of an integrated marine information system and of its interface to it for the remote users.

The database itself is described in Annex 10.

The analysis modules are described in the various intermediate scientific reports (Annexes 1 to 3 and 11).

The interface can be accessed at http://www.mumm.ac.be/datacentre/.

6. CONCLUSIONS

Although the project suffered at its beginning from important delays due to the administrative changes affecting MUMM at that time, the final product meet most of the objectives laid down in the original proposal.

The 'Belgian Marine Data Centre' (BMDC), a group of skilled people within MUMM, now operates a marine information system that fulfil most of the needs expressed by the data providers and the (potential) users.

The fundamental investment has been done, in terms of money and of training. The database structure is operational and a series of data sets were already imported : the database is now ready for a continuous incoming flow of new marine data.

The BMDC will continue to maintain, upgrade and "feed" the system in the future. Some effort yet has to be requested from the data providers in order to obey they contractual duties and transfer their data in due time. They have to understand (and the BMDC has to let understand by providing concrete examples) that the data centre acts as a sort of 'peer–reviewer' for their data, giving therefore more value to them.

Another benefit, for the research teams and for the funds invested by their sponsors, is the commitment of the BMDC to take care of their data and of the related meta–information for a long period of time.

The BMDC, its tools and the people managing them, has now reached a level of skills that brings it to the level of the equivalent centres in Europe.

7. ACKNOWLEDGEMENTS

The IDOD project was financed by the Belgian Science Policy (OSTC), in the frame of the Programme 'Sustainable Management of the North Sea' (1997-2002), under contracts MN/DD/60 , 61 & 62.

The partners of the project (MUMM, KUL–UCS & ULg–SURFACES) wish to thanks all the people that gave them useful input during the project : members of MUMM and of the OSTC, scientists involved in the Programme, members of the Users Committee and database experts Jan VANOPPEN & Günther LEMMENS.

8. LIST OF ANNEXES

Annex 1 : Intermediate Joint Scientific Report #2 (Year1998)

Annex 2 : Intermediate Joint Scientific Report #3 (Year1999)

Annex 3 : Intermediate Joint Scientific Report #4 (Year2000)

Annex 4 : IDOD Newsletter #1

Annex 5 : IDOD Newsletter #2

Annex 6 : IDOD Newsletter #3

Annex 7 : IDOD Newsletter #4

Annex 8 : Convention "IDOD"

Annex 9 : "IDOD" Overeenkomst

Annex 10 : Data base documentation

Annex 11 : KUL–UCS scientific report 2002

Annex 12 : Belgian national contribution to the OSPAR "Comprehensive Procedure" on the Common Assessment for the Eutrophication Status of the OSPAR Maritime Area.

Annex 13 : "Marketing tools" : information sheet, poster, Presentation at the Symposium on the Sustainable Management of the North Sea (Brussels, January 21–22, 2002)

PRIME MINISTER'S SERVICES
Federal Office for Scientific, Technical and Cultural Affairs

SCIENTIFIC SUPPORT PLAN
FOR A SUSTAINABLE DEVELOPMENT POLICY

«SUSTAINABLE MANAGEMENT OF THE NORTH SEA»

RESEARCH CONTRACTS MN/DD/60, 61 & 62

INTEGRATED AND DYNAMICAL OCEANOGRAPHIC DATA MANAGEMENT

JOINT SCIENTIFIC REPORTS
for the year 1998

January 1999

The present document gathers the yearly scientific reports of the three partners to the *Integrated and Dynamical Oceanographic Data Management* project, performed on behalf of the Federal Office for Scientific, Technical and Cultural Affairs.

It covers the activities performed during the year 1998. It contains, in sequence :

- the scientific report of the Management Unit of the Mathematical Model of the North Sea (MUMM), and its annex;
- the scientific report of the SURFACES laboratory (ULg), with its five annexes;
- the scientific report of the *Universitair Centrum voor Statistiek* (UCS, KUL), with its two annexes;
- the dictionary of the IDOD data base;
- the *IDOD Newsletter #1*;

and, for the original copies of the report only,

- the conceptual and logical schemes of the IDOD data base, on separate sheets.

January 1999

MANAGEMENT UNIT OF THE
NORTH SEA MATHEMATICAL MODELS

DATA MANAGEMENT GROUP

Integrated and Dynamical Oceanographic Data Management - IDOD

Karien De Cauwer, Serge Scory

DMG/1/KDC/199808/EN/IR

Joint Scientific Report n°2 (January-December 1998),
prepared for OSTC, Contract MN/DD/60

MUMM
100 Gulledelle
B–1200 Brussels
Belgium

Table of contents

1. Introduction

The tasks performed this year have mainly concentrated on the elaboration of the conceptual scheme of the IDOD database and on the translation thereof into a prototype of the physical database. This physical database is of prime importance for other tasks requiring knowledge of the structure or of the outputs of the database (*e.g.* quality control and statistical analysis tools).

The definition of the data structure and the elaboration of the conceptual scheme are shortly described in Section 5. In Section 6 the physical implementation is presented. The functionality of the database prototype is tested, based on a preliminary market analysis as listed in Section 4. ***N.B.*** : this report only gives a synthesis of the conceptual scheme of the IDOD database. As the concept and the application result from a joint effort of ULg–SURFACES and MUMM, the details are only given once, in the scientific report written by ULg–SURFACES (*aka* "The IDOD Database User's Guide").

Besides this, two long-lasting activities continued: the inventory of existing and future data sets (see Section 2) and the validation of data held in the old database at MUMM. The latter activity has resulted in a list of tests to be performed when these data will be transferred into the new database. The layout of the quality control, with a list of checks, is described in Section 3.

2. Inventory of data sets (Tasks A.1 & A.2)

The inventory of data sets has continued at different levels. The purpose of the inventory is to give an overview on the existence, the availability and the relevancy for the IDOD database of data sets concerning the Northwest European shelf (the IDOD zone of interest is bounded by the 200 m depth contour, corresponding approximately to the zone between 44°N–62°N & 10°W–12°E).

The inventory also aims at distinguishing the different data types. The data available at MUMM have been inventoried in a more detailed way in order to characterise the most important data types. This characterisation is the basis for the conceptual scheme (see Section 5). Considering the 'future' data sets, a start has been made by the identification of the parameters measured in the frame of the programme 'Sustainable Management of the North Sea'. The descriptions of all the data sets will finally be incorporated in an updated and reorganised EDMED (European Directory of Marine Environmental Data) inventory.

2.1. Existing data sets

2.1.1. Data collected by or available at MUMM

Several days have been spent to discuss with ULg-SURFACES the different data sets available at MUMM. The following list of data sets has then been set up. It can be considered as an exhaustive list of data sets or data sources available at MUMM.

1. Monitoring data

Monitoring data are concentration data measured during campaigns in the Belgian Continental Shelf area and in the Scheldt estuary. The parameters routinely measured are:

- For seawater (years 1977-1997) : salinity, temperature, pH, dissolved oxygen, suspended matter, nutrients, pigments, metals, organic contaminants;
- For sediment (years 1985-1994) : general sediment parameters, metals, organic contaminants;
- For biota (years 1978-1994): metals and organic contaminants in fishes; metals and organic contaminants in birds, sea mammals and invertebrates.

MUMM is only responsible for the measurements made on seawater samples. The Sea Fisheries Department disposes over the recent data on biota and sediments.

2. “Input”

Measurements made on samples collected in effluents and rivers are used to calculate the total quantity of pollutants entering the North Sea. For this purpose the discharge is calculated using flow rates, either measured or estimated. The loads are calculated for direct effluents (industrial and sewage), riverine input (Scheldt and Ijzer) and tributary rivers (canals and polders). The *Vlaamse Milieumaatschappij* (VMM) is responsible for the analyses, MUMM for the calculation of the total input. The results are then reported to the Secretariat of the Convention of Oslo and Paris.

The parameters analysed since 1990 are: nutrients, metals, suspended matter, organic contaminants, pH, dissolved oxygen, chemical oxygen demand (COD), biological oxygen demand (BOD), conductivity and sulphates (SO_4).

3. Bathing water quality

Water samples are regularly taken in the 39 official bathing zones along the Belgian coast during the tourist season, in order to identify possible risks for the human health. The concentrations of faecal coliforms, total coliforms, total *Streptococci* and *Salmonellae* are measured. Since 1995 the *Vlaamse Milieumaatschappij* (VMM) is responsible for these analyses while MUMM was in charge of this monitoring programme the years before.

4. NOWESP (Northwest European Shelf Programme)

The NOWESP database contains the results provided by nearly 50 institutes and data centres concerning the area of the Northwest European Shelf. Results are stored for: nutrients, salinity, temperature, suspended matter, chlorophyll and particulate organic carbon for the period 1960-1993. As a contributor to this project, MUMM has the right to use the content of the whole database.

5. Mathematical models

Four models were identified as relevant for incorporating their results in the IDOD database:

- STORM : computation of tides and storms,
- OMNECS : new model (1998) for the computation of tides and storms,
- HYPAS and REFRAC (together: the *DEINING* system) : computation of waves.

These models will be run on long periods of time, with or without (for tidal models only) meteorological forcing, in order to define characteristic current and wave patterns corresponding either to true meteorological conditions or to a "characteristic year".

6. BELMEC (BELgian Marine Environmental Control)

BELMEC is the name of the aerial surveillance programme of the pollution in the Belgian Zone of Interest in the North Sea. For each flight, the following information is available: date, time, equipment used, routing, visibility, sea state, wind speed and direction. For each observation, following items are, amongst others, recorded: date, time, position, length of pollution slick, width of pollution slick, colour, mission type, wind speed and direction, sea state, visibility, administrative information.

7. Geographical data

A list of the geographical features of interest has been set up. It includes a. o. (in no special order):

- the bathymetry,
- the coastlines, the estuaries and the rivers,
- the limits of the national continental shelves and the land boundaries,
- the limits of territorial waters (12 miles zone),
- the responsibility zones in the frame of the Bonn agreement,
- the ICES zones,
- the zones of activity of the RV BELGICA,
- the delimitation of the navigation routes,
- the sand and gravel extraction zones,
- the pipelines and the communication cables,
- the anchoring zones,
- the military exercise areas,
- the monitoring stations.

A first inventory of the data available for the moment shows that they mainly concern the Belgian coastal region and not the complete zone of interest for IDOD.

8. ODASII (Oceanographic Data Acquisition System II)

ODASII is a data acquisition and online data processing system for oceanographic, navigational, meteorological and hydrographic data on board of the RV BELGICA (and for meteorological parameters measured at MUMM–Ostend). 400 different parameters are stored. Since 1996, results are stored at least every 10 seconds when the BELGICA navigates.

Besides the results of continuous measurements during the cruises, current meter, tidal, meteorological, SCTD vertical profiles, sand and gravel automatic surveillance data ("black box") and data from an automatic station located along the Scheldt will be entered in ODASII.

2.1.2. Data available in other institutes and university laboratories

So far, only the data sets kept at the Sea Fisheries Department (DVZ) have been clearly identified. Further contacts need to be made to assess the existence and readability of possibly useful data sets in other state institutions and universities (see 2.1.3).

- Sea Fisheries Department

  As the DVZ is responsible for the analyses in sediment and biota, important data sets complementary to the monitoring data in seawater, a technical reunion took place to discuss the characteristics. The department disposes over the following data:

  - Sediment (since early in the seventies, digital since 1990 [under *Access*]): contaminant concentration measured on grab samples,
  - Biota (since early in the seventies, digital since 1978 [under *Access*]): heavy metals and organochlorines in fishes. Other kinds of data are available outside the database like the age of fishes and biological effect measurements.

2.1.3. Catalogue of existing data sets : EDMED

The European Directory of Marine Environmental Data (EDMED) was compiled in the frame of the Marine Science and Technology Programme (MAST) of the European Commission. The list of Belgian data sets (almost 100) was last updated in 1995 and covers most of the past and present oceanographic studies. In the frame of EURONODIM, an EC Concerted Action in which MUMM participates, the EDMED inventory will be updated and extended. Actions are ongoing to achieve a uniform format taking into account other international approaches and present experienced shortcomings, including multilingual aspects.

In the frame of IDOD extra fields are added giving information on the data type and the availability and readability of the data. Some fields need systematisation. This update is currently ongoing for the most recent data sets.

2.2. *Future data sets - Data collected by the programme 'Sustainable Development of the North Sea'.*

To start with the identification of the future data sets, the research groups involved in the programme 'Sustainable Development of the North Sea' were contacted. Reunions took place the 16th of June and 4th of December 1998. The list of parameters measured by each laboratory is given in Annex. In the coming months each team is expected to send one sample data set to IDOD, in order to help the project team to characterise the data and to check and, possibly, adapt the actual conceptual scheme (see Section 5 and 6).

3. Quality control (Task B.2.2 & B.3)

3.1. *Levels of quality control*

Quality control is the foundation of the credibility of a database. The control process takes place at different levels. For the IDOD database following levels were identified:

1. **Acceptation of data and data documentation**

   First, it should be verified that the delivered data set contains:

   - values expressed in significant figures,
   - complete position (including depth) and time reference,
   - clear description of parameter and measurement unit,
   - description of methodology used, the detection limit and the expected accuracy,
   - information on quality control and assurance.

   More details on the expected data documentation are given in chapter 5 of the *IDOD Newsletter 1* (see annex to the joint scientific report).

2. **Control of integrity**

   Level two will be performed at the entrance of the data in the database. Following checks are envisaged:

   - syntax check : agreement of value with a pre-defined type
   - external value check : values checked against a defined list of valid values
   - combination of fields check : some fields can only be filled or are required if one or more other fields are given, a field's value can be predefined depending on the value of other fields (*e.g.* if qualifier flag is "smaller than", value should equal detection limit as mentioned with the analysis method, *e.g.* date should comply with the period of the campaign during which the sample was collected)
   - check of replicate data (*e.g.* the same values for salinity at a sampling occasion are reported by different laboratories, same parameters stored every 10 seconds and every minute for continuous data)

- position checks by means of sampling station names, speed calculations and comparisons with track plots
- internal value check : coherence of the values of several related fields in the same record (*e.g.* ecosystem code can be checked by the values for latitude and longitude)

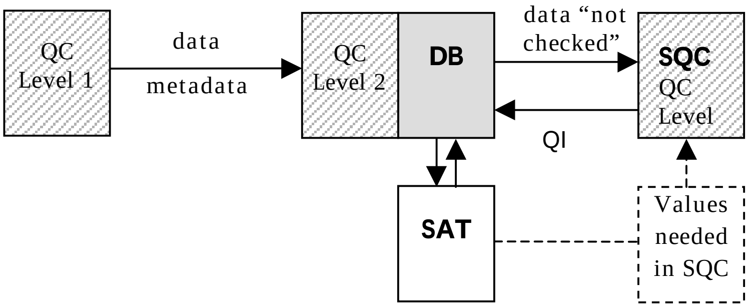
3. **Statistical quality control**
   - external value check: values must fall within a range (*e.g.* salinity approaching zero only possible at positions in rivers). Typing errors and errors in measurement unit can be detected (*e.g.* the units of dissolved oxygen are normally given in g/l but were also entered in the old data base as l/l according to ICES format). Values will be checked against values made at a small distance or checked against values of related parameters at the same place and time. Remarkably high or low values will be flagged as 'exceptional'.

The tests of level 2 will be elaborated by ULg-SURFACES and MUMM and will be implemented when entering the data in the database. The list of checks, as given above, will be updated continuously.

For the elaboration of the statistical tests, designed by KUL-UCS, bilateral reunions will take place with the data providers when data sets are delivered. For the moment, the statistical quality control procedure is implemented with 3 specific tests.

3.2. *Data link: Quality Control Programme – Database – Statistical Analysis Programme*

The following model is envisaged to organise the data stream between the different modules of the IDOD database:



After acceptation, the data and meta information are entered in the database. When entered, the data are checked for format inconsistencies and integrity (QC Level 2). The Statistical Quality Control Program (SQC) receives data from database and sends the quality information, back (QC Level 3). The meta information is not used in SQC. The statistically defined relationships used in SQC will be determined by using the Statistical Analysis Tool (SAT).

4. Market analysis (Task F.1)

A wide range of users will request for information stored in the database. The type of question can differ from user to user. An inventory of different types of questions will aid in the conceptual analysis of the database and the choice of the analysis tools supplied with the database. The requests for data in the old database at MUMM are used as starting point for the inventory of the user needs. A list of typical questions, especially for monitoring data in seawater, is presented below.

TYPICAL DATA REQUESTS

Data measured in seawater for

- *a project and time period*

  e.g. MUMM monitoring data for 1996
- *a cruise*

  e.g. Belgica cruise 97/3
- *certain parameters and/or parameter categories at sampling stations, season and time period*

  e.g. Nutrients and general inorganic parameters at certain stations measured during spring and autumn 1995-1997.
- *certain parameters at or near a described place, season and time period*

  e.g. Organic contaminants in seawater of Westerscheldt during last 5 years.

  e.g. Nutrients, pigments and phytoplankton in seawater near the Belgian coast measured in autumn in the period 1980-1997.

Derived data measured in seawater

- *Grouped in time*

  e.g. Monthly or seasonal averages of temperature, salinity and water depth for certain stations since 1977.

  e.g. Seasonal or yearly overviews with averages, minimum and maximum values and number of measurements of suspended solids in the entrance routes to a harbour.
- *Grouped in time and place*

  e.g. Monthly averages of depth, temperature and suspended particles in squares of 5'.

Scientifically interpreted data measured in seawater:

e.g. Comparison of contaminant concentration with concentrations in 'non' polluted sea- or brackish water.

e.g. Extreme salinity and temperature analysis for the Belgian Continental Shelf (period 1977-1990)

Questionnaires/inventories:

e.g. Frequency of sampling and reporting of certain parameters

e.g. List of parameters measured by service and year

Data resulting from continuous measurements

e.g. 30 minutes averaged data for navigation and meteorological parameters for a certain campaign

This list is certainly not complete. Potential users will be asked whether these questions cover their general needs. For this purpose, the list will be published in the IDOD Newsletter.

On a similar topic, a first experience with market analysis, led by the Data Management Group of MUMM, took place during the second semester of 1998 in the frame of the European project MAUVE (MAST programme). The purpose was to collect and analyse the expectations of the potential users of an autonomous underwater vehicle for the acquisition of oceanographic data. The survey was performed by means of an electronic form made available on the Web and announced by means of an electronic message sent to more than 500 addressees. It allowed us to put in evidence the pros and the cons of such an “electronic survey” (Scory & Pichot, 1998).

5. Conceptual design of the database (Task C.1.2)

The conceptual design of the database has been done by MUMM and ULg-SURFACES. A general structure was presented for the most important data types:

- point measurements of concentrations in
    - seawater : *e.g.* monitoring data, input, bathing water quality and NOWESP (see 2.1.1),
    - biota : *e.g.* monitoring data (see 2.1.1 and 2.1.2),
    - sediment : *e.g.* monitoring data (see 2.1.1 and 2.1.2);
- continuous measurements in seawater : *e.g.* ODAS (see 2.1.1).

So far, the conceptual model was further elaborated for seawater. The conceptual schemes, realised with the graphical tool of DB-Main[1], are presented as annex to the joint scientific report (only included in the original copies). The data dictionary (see annex to the joint scientific report) lists the entities represented in the scheme together with their attributes. It also gives the full details of the attributes: description, format, example value, *etc.* A general description of the conceptual model is given in section 2.2 of the *IDOD Newsletter 1*.

For the characterisation of the continuous measurements, a reunion took place at the data acquisition centre of MUMM at Oostende. Besides the higher rate of sampling, the main difference with the discontinuous point measurements is the absence of the concept of “samples” and “subsamples”. Continuous data are automatically acquired using sensors and can be related directly to the measurement campaign.

For sediment and biota, the changes to the basic scheme presented here above will mainly consist in the addition of new attributes. Information has been gathered in collaboration with the Sea Fisheries Department. When measuring contaminant concentrations in sediment, additional information is

[1] *DB–Main* is a generic case tool for database conception and was developed by the *Institut d'Informatique* of the *Facultés Universitaires Notre-Dame de la Paix* in Namur. It is available on the Web at *http://www.info.fundpc.ac.be/~dbm/*.

recorded: grain size fraction subject to analysis, extraction method of the sediment, structural analysis, upper and lower depth of the core in case of core samples, *etc.* For concentrations measured in biota, data about the species, number of individuals, analysed tissue, average weight and length of collected individuals, *etc.* are required.

The model and the data dictionary were presented, in the *IDOD Newsletter 1* and during reunions, to all persons working at MUMM, to the person responsible for the monitoring database at the Sea Fisheries Department and to the teams involved in the Programme 'Sustainable Management of the North Sea'. All their comments are presented on the IDOD website and were taken into account whenever possible and necessary.

The conceptual model underwent the normalisation process up to the third form. The resulting changes are given in Annex A of the scientific report written by ULg.

6. Logical design and physical implementation (Tasks C.2.3 & C.3.3)

The transformation of the conceptual model into a logical model required the addition of attributes to establish links between the different tables. This task has been done by ULg-SURFACES and MUMM.

The implementation of the physical prototype database has been done in *Access97*. More details are given in the contribution of the ULg.

6.1. Input of data

In order to test the prototype of the database, the data corresponding to the monitoring campaigns of 1996 have been prepared for inclusion. This involved the following steps:

- conversion from RBase to *Access97*,
- addition of new fields (*e.g.* sample code, subsample code, units),
- addition of new tables (*e.g.* descriptions of methods, cruises, projects, services and institutes, as this information did not exist in the old database at MUMM),
- establishment of links and queries to store the information in the appropriate tables,
- definition of lists for certain attributes (parameter names, parameter categories, ICES ship codes, country codes, units, and substrates).

6.2. Queries and forms

The functionality of the prototype has been tested in close collaboration with ULg-SURFACES. Based on the list of requests (see Section 4), different queries have been performed using the *Access query design grid*.

An example of input form has been made, allowing the addition of new records for an entity and the addition of new records in a linked table.

7. Reporting to international bodies & international evaluations (Task F.3)

MUMM has fulfilled the Belgian obligations in the frame of *the Joint Assessment Monitoring Programme* of the Oslo and Paris Commissions by reporting the monitoring data for 1997 to ICES. National comments (De Cauwer, 1998) have been issued to describe the monitoring activities and to present the results.

The reporting procedure involves the gathering of data from different sources including method descriptions and units, formatting of the data, checking with the ICES screening programme and the analysis and interpretation of the monitoring results.

8. Other tasks

1. **Definition of the terms and conditions of the data transfer and access to the database (Task A.3)**
2. **Inventory of required meta information (Task A.4.3)**

On both subjects general proposals have been formulated in the *IDOD Newsletter 1* which was distributed to all the network co-ordinators of the 'Sustainable Management of the North Sea' programme. These topics have been discussed during the reunion of the 4th of December 1998.

Concerning the access rules, a more specific proposal of agreement will be written by MUMM.

The 'Guideline for data documentation' (from the MAST Data Committee) will serve as a checklist for the required meta–information. The network co-ordinators of the 'Sustainable Management of the North Sea' programme were asked to provide that information together with the first set of data they will send to IDOD. Depending on the difficulties encountered and the possible comments, the list of required meta–information will be evaluated.

3. **Contact with data originators (Task B.6)**

As mentioned in 2.2, two co-ordination reunions took place in the frame of the programme 'Sustainable Management of the North Sea'. The Sea Fisheries Department and the data acquisition centre of MUMM were also involved in several aspects of the conception of the final database. As stated above, news on the project was presented in the *IDOD Newsletter 1*. The contacts with the data providers contributed to the preliminary transfer and access rules, to the data inventory and structure definition and to the design of the conceptual scheme. In the future, the involvement of data providers will be of continuing importance in the above-mentioned tasks and in several aspects of quality control and statistical analysis tools.

4. Statistical analysis tools (Tasks D.1.3, D.3)

MUMM contributed to the elaboration of a first set of statistical checks by KUL–UCS. These functions will serve as a basis for the design of standard statistical analysis tools. For more details, see the Scientific Report for the year 1998 by KUL–UCS.

5. Mathematical models : validation and incorporation of results (Tasks E.1, E.3, E.4)

A procedure has been set up to verify the results of the most recent operational hydrodynamic model. Actually, actions are undertaken to acquire observation series long enough to validate these results.

9. Conclusion

A prototype of the IDOD database is implemented in *Access97*. Several queries, based on the list of requests sent to MUMM, have been tried out successfully. The database structure might be subject to minor changes, as a final evaluation has not yet taken place.

Although the actual prototype is only valid for concentration measurements in seawater, information has been gathered in the course of the year about other data types. In a short time it should be possible to make the necessary adaptations for sediment and biota. After this, the less standard data types will be addressed. The data set samples expected shortly from the teams of the programme will be of utmost importance to check and extend the database structure.

The inventory of existing data sets will be continued following the extended and updated EDMED format. Priority is given to the most recent data sets.

Regarding the quality control, the global organisation is now defined. A list of required checks exists which will be updated continuously taking into account the specifications expressed by the data providers.

10. References

Anonymous, 1999. *The IDOD Database User’s Guide. In* “IDOD – Scientific report for the year 1998”, F. Muller, ULg, Laboratory SURFACES.

De Cauwer, K., 1998. *National comments to the 1997 Belgian monitoring data for seawater*. Submitted by Belgium for ICES and PARCOM, Ref. DMG/1/KDC/199808/EN/IR.

KUL–UCS, 1999. *IDOD – Scientific report for the year 1998*.

De Cauwer, K. and Scory, S. (Eds), 1998. *IDOD Newsletter 1*. October 1998. By KUL-UCS, MUMM & ULg-SURFACES.

Scory, S., and G. Pichot, 1998. *MAUVE: User requirements survey. A contribution to MUMM's MAUVE final report*, December 1998, 24 pp.

COLOPHON

This report was issued by MUMM in January 1999.

Its reference code is DMG/1/KDC/199808/EN/IR.

Status ☐ draft
☒ final version
☐ revised version of document
☐ confidential

Available in ☒ English
☐ Dutch
☐ French

If you have any questions or wish to receive additional copies of this document, please send an e-mail to *IDOD@mumm.ac.be*, quoting the reference, or write to:

MUMM
100 Gulledelle
B–1200 Brussels
Belgium
Phone: +32 2 773 2111
Fax: +32 2 770 6972
http://www.mumm.ac.be/

MANAGEMENT UNIT OF THE
NORTH SEA MATHEMATICAL MODELS

MODELLING GROUP

The typefaces used in this document are Gudrun Zapf-von Hesse's *Carmina Medium* at 10/14 for body text, and Frederic Goudy's *Goudy Sans Medium* for headings and captions.

UCS
Universitair Centrum voor Statistiek

Integrated and Dynamical Oceanographic Data Management - IDOD

Scientific Report

Contribution of KUL-UCS

KATHOLIEKE UNIVERSITEIT
LEUVEN

University Center of Statistics
Katholieke Universiteit Leuven
de Croylaan 52b
B-3000 Leuven
Belgium

1. Introduction

During this second year of the IDOD project the work of UCS has concentrated on the actual implementation of the statistical quality control procedure (SQC), of which the conceptual framework has been developed during the first year. In addition, a statistical study of actual data has been executed with the purpose of identifying general characteristics that could be used in the quality control procedures. The results of this study are summarized in Section 2. Section 3 briefly describes the functionalities that are supported by the current version of the SQC program. The work accomplished is summarized in Section 4.

2. Statistical study of the 1977-1996 monitoring data for seawater

In a first step, an exploratory analysis has been executed of the 1996 Belgian monitoring data for seawater (see Appendix 1, for details of the study). In that study it was concluded that the dispersion of the data is such that a simple lower- and upperbound check is unlikely to detect erroneous measurements and that somehow location dependency must be accounted for. In this exploratory study, salinity rather than the actual spatial coordinates were used to account for such dependency. It was furthermore found that at least for some of the measured variables there is a high correlation between the different measurements (after removal of the location effect) and that by using this correlation, the dispersion around predicted values could be reduced.

In a second step, a more complete study has been made of the contaminant data gathered from 1977 until 1996 of the Belgian monitoring program with the specific objective of identifying statistical characteristics that could be used for quality control testing. In order not to confound the conclusions of that study with problems associated with a too small data size, the analysis has been limited to consider those measurement variables for which ample data have been measured. Three statistical characteristics have been examined: 1. the univariate distribution, 2. the spatial correlation, and 3. the correlation with other variables. The nature of each of these characteristics, their possible use in a quality control check and problems still to be resolved are summarized next (see Appendix 2, for the details of the study).

The univariate distribution could be used to define lower- and upper-bound values that are expected to occur very rarely and against which the measurement values could be compared. The analysis shows that it is important to first of all transform the variable of interest in order to satisfy the assumption of normality and that after transformation the mean value of the univariate distribution is for most of the variables a function of location (in particular for the case of the river Scheldt) as was also found in the exploratory study. Differentiation between open sea and the Scheldt river and differentiation with respect to seasons is also found to be important in several cases. A procedure to automatically select the appropriate model has been devised and could be used to automatically extract this relation from the database (then to be used in the quality control program). Not investigated in this study is whether also the standard deviation would vary with the same variables. We decided that such a study is better postponed after the present algorithm has been incorporated into the statistical analysis tool (SAT) program that will be integrated into the IDOD database, after which it should be easy to apply the same algorithm in this case to the squared residuals of the model to obtain thus a model for the standard deviation. It is also found that for practically all of the variables the lower-bound value is physically equal to 0 and, in practice, corresponds to the detection limit of the instrument or method used. Thus, for practical applications, the quality control check will be typically based on the comparison against a statistically derived upper-bound value, while the lower-bound value would be physically based.

The study shows that for most of the measurement variables there is a reasonable amount of spatial correlation and this is a second characteristic that could be used in the quality control program. The spatial correlation can be quantified through the variogram and in the study a first estimation procedure has been applied to the data. Again, it is planned that this procedure will be part of the SAT program so that estimates can be automatically obtained. The distinction between open sea and the river Scheldt is found to be significant. It is also found that for metal observations in open sea, spatial correlation is small or non-existent and in such a case, a quality control based on the variogram value will be of little use. The study has been limited to a particular model for the variogram (a power-law model) which is found to lead in most of the cases to a reasonable fit. Other models and other estimation methods for the variogram will be eventually considered in a final implementation in the SAT program.

Finally, the study has considered the possibility to predict a measurement value on the basis of a linear combination of other simultaneous measurements (in the study no transformation is applied in this case). It is shown that one can construct a ranked list of possible predictions that depend on the availability of the other measurements and that, for most of the variables, the top-ranked combinations (the best fits) have a high R^2 goodness-of-fit statistic. In practice this would mean that a quite accurate independent measurement value can be estimated on the basis of the other measurements and thus a reliable quality control can be performed. The study shows that this does not apply for metals where correlation with other variables is found to be relatively small. An issue that has not been addressed in this study is the type of model: in all cases, a linear relationship between the measurement variables has been assumed. Again, this may be an issue to be considered in further depth in the final implementation of the algorithm in the SAT program. In this study, the maximum number of regressor variables has been limited to 3 and the best model is always found to use this maximum number. For this reason, one might also consider increasing this number, even though in practice it will become of course less likely that all such measurements are made at the same time.

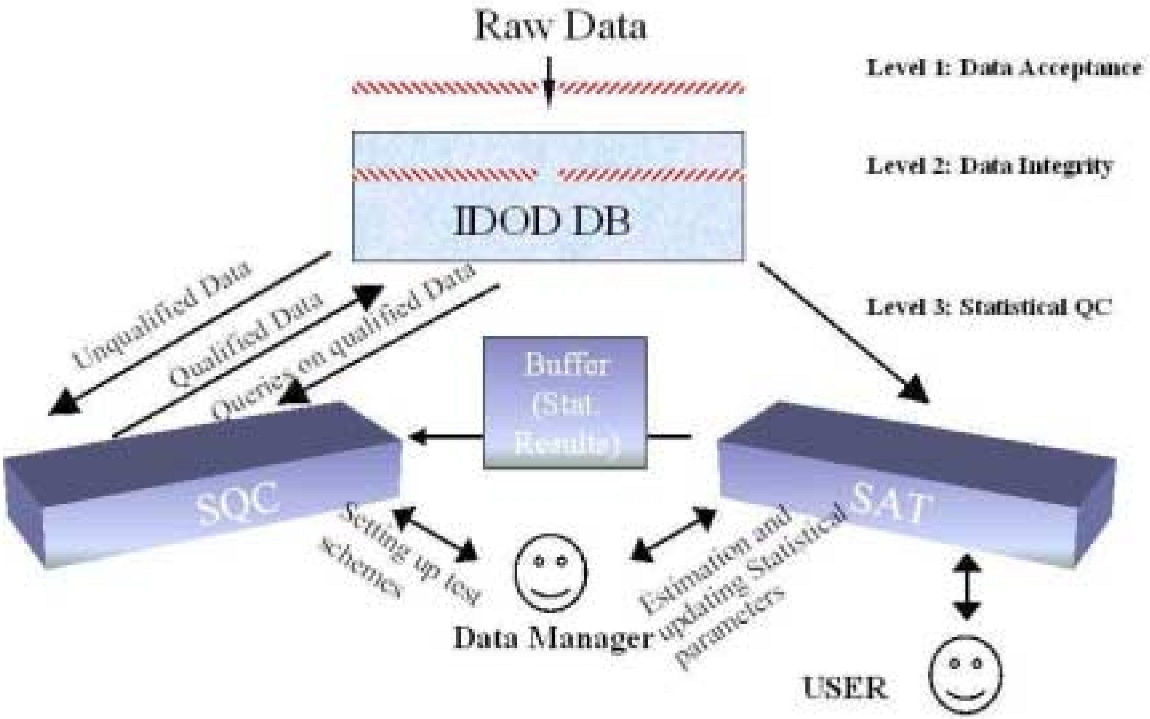
Perhaps the most important overall conclusion of this study is that it will be extremely difficult and potentially dangerous to develop fully automated algorithms that result in setting the coefficients for a quality control test. In each and every case, some assumptions are fixed (normality, or linearity, or dependence only on certain variables) and it is only through visual examination of the data and consideration of the results that the validity of these assumptions can be (to some extend) verified. Practically speaking, this means that we advise that each quality control check, that is used in practice, is derived with the help of the SAT program, but that transfer of the results of this program to the application of a specific quality test requires the intervention of the data base manager. Such a human interaction appears to us essential to safeguard the system against happenstance relationships or the overlooking of quite strong relationships that are not detected within the (necessarily) fixed framework of any model estimation.

3. The Statistical Quality Control Program

In this second year of the IDOD project an operational prototype version of the statistical quality control program has been implemented in the form of the SQC program. In the following we first describe the function of the SQC program in the overall quality assurance approach and its integration with the IDOD database. Next, we briefly describe the main functionalities of the program.

3.1 The overall QA-approach

The overall dataflow and the function of the SQC program in this dataflow are illustrated in Figure 1. The raw data delivered by the data providers is directly put into the IDOD database. At this stage, a first level of quality assurance is executed which consists in the check of whether all data acceptance requirements are fulfilled (i.e. proper documentation of the instruments, completeness of the data, etcetera). BMM, the leader of the project has assumed responsibility for setting up the data acceptance requirements.



Figu

Figure 1: Overall QA-approach

Immediately upon entrance into the database a second level of quality assurance is executed which consists of checking the data integrity of the incoming data (i.e. do reference locations exist, do references to measurement methods exist, are the data of a valid numeric form, etcetera). At this level, simple deterministic checks may be also included (i.e. is the ship speed that is implied by the data a reasonable one, etcetera). ULg, as designer of the database, has assumed responsibility for this step.

Once the data are accepted into the database, they are considered valid but "unqualified" data, meaning that their statistical consistency with other data or with expected distributions has not been verified. This is the task executed by the SQC program which is activated by the data manager. As explained in Section 2, this program will for instance check the value of the different measurements against upper- and lower-bound values that are temporal and spatially dependent, it will check the values against similar measurements made at a distance and it will check the values against measurements of another type made at the same location and time. The relationships used in this checking are statistically defined and will be determined by use of the SAT (statistical analysis tool) program. This program which is to be developed in the upcoming year will include general statistical tools tailored to the IDOD database and the problems associated with oceanographic data, but will also include specific algorithms that can be used to estimate the relationships needed by the quality control program. While UCS has already estimated such relationships from the data made available and will continue to do so during the course of this project, the final responsibility of updating these relationships as more data become available will rest with the data manager. This provision of a human intervention (someone critically examining the results of the analyses) has been specifically made in view of the conclusions of the statistical study (see Section 2). The derived relationships are stored in a buffer, partly to shorten the CPU time of execution of the quality control procedure and partly to enforce the examination of such results and prevent the fully automated estimation of relationships.

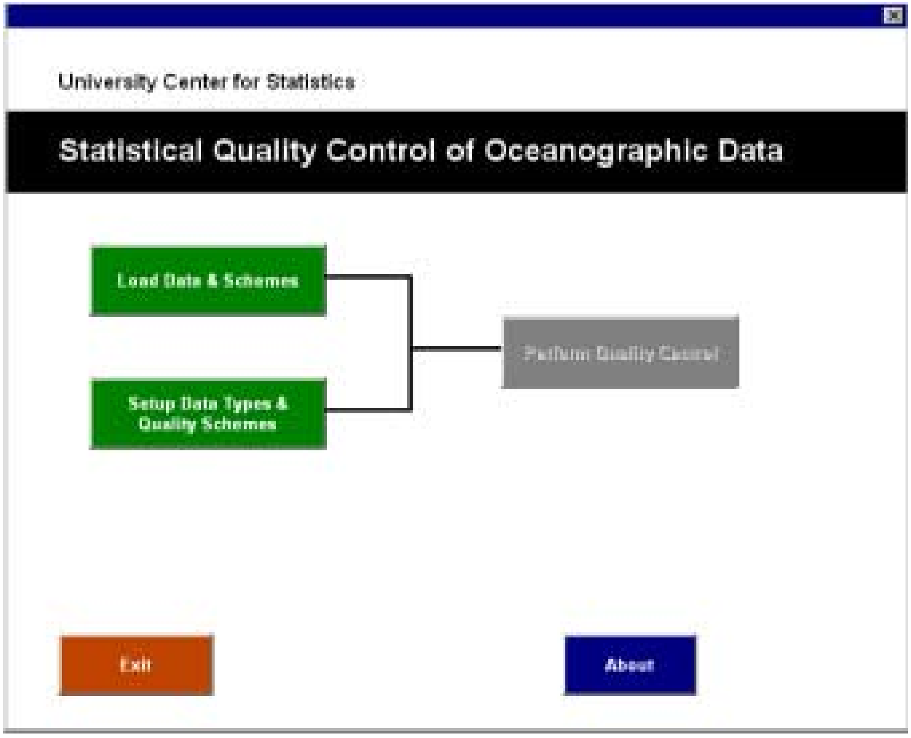
3.2 The conceptual framework of the SQC program

The conceptual framework of the SQC program has been explained in the previous scientific report and has remained largely unchanged. For details, we therefore refer to that report. In essence, a dataset (that may consist of multiple measurements of different type) is entered into the SQC program and each datapoint (or a set of datapoints within a given window) is subjected to a series of tests which are grouped into different stages (i.e. one stage might consider tests on individual datapoints, while another stage might apply tests that consider a group of consecutive measurements). The definition of the different stages and the different tests that are part of the stage is the responsibility of the datamanager, but UCS will predefine such tests for specific data types as part of this project and the SQC program is specifically designed to facilitate this work of defining the stages and the consecutive tests.

As anticipated in the conceptual framework, the application of the tests and the combination of the results of these tests into a single numeric quality label is done using a Bayesian algorithm (i.e. the a-posteriori probability that a datapoint is an "exceptional" measurement is calculated). It is this algorithm that is supported by the developed program. The application of this algorithm to actual data does require however the assignment of the probability that a "normal" measurement is classified as "exceptional" by an individual test and vice-versa. We expect that this input may be difficult to generate in practice and, it is possible, that the combination of the different test results in a final version of the program will be done by combining the significance level of the different test results. This is an issue that will be considered in further detail in the next year of research.

3.3 Functionalities of the SQC program

The main functionalities of the SQC program are shown in Figure 2, which corresponds to the main menu of the SQC program.



University Center for Statistics

Statistical Quality Control of Oceanographic Data

Figure 2: Main Menu of SQC-program

"Setup Data Types and Schemes" is an option that can only be activated by the datamanager and refers to the ability of the SQC program to define in an easy and user-friendly manner:

- data types, which consist of a group of oceanographic measurements from a data provider. In its current version, the SQC program maintains its own list of measurement types (this list may at a later stage be eventually linked directly to the IDOD database) and allows the declaration of new measurement types. SQC is further conceived to allow for the possibility that not all measurements that should or could be measured as part of a given data type are actually provided. As a consequence, data types can be rather generic descriptions of data that may fit more than one data provider or data campaign;
- quality control schemes which represent a sequence of tests (organized into different stages) that should be applied to a specific data type. As part of this functionality it is possible to define new tests by choosing a generic test (i.e. univariate regression test, variogram test, ...) and entering a limited number of parameters that make the test specific to its application for a given measurement or data types. The definition of a scheme then consists of a selection from the available tests and the organization of these into a sequential chain, possibly subdivided into different stages.

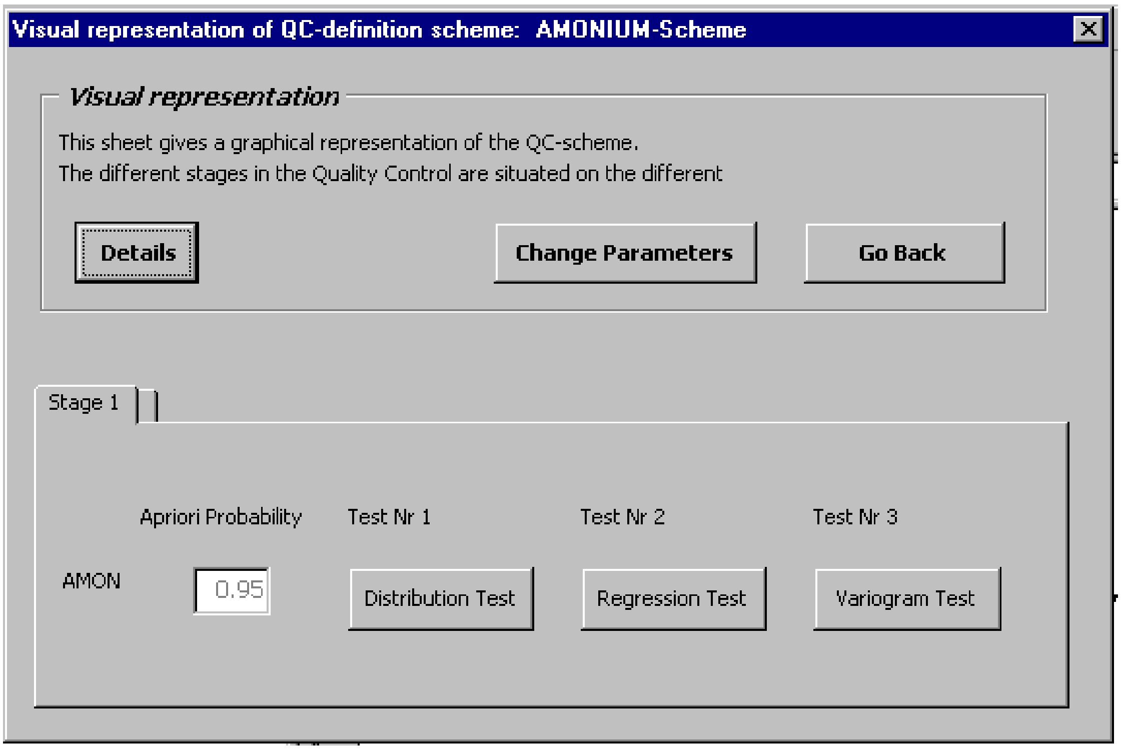
"Load Data & Schemes" is a 4-step initialization process which includes the following steps:

1. indication of the file from which the measurements should be imported (in a final version of the program, this will be directly linked to the IDOD database);
2. indication of which data type applies and which measurements of the data type are actually provided (at individual instances, some of these measurements may still be missing);
3. the order of the measurements in the data record (not necessary if linked to the IDOD database);
4. the quality control scheme that should be applied.

"Perform Quality Control" executes the scheme on the provided data and determines for each measurement a numeric quality value varying between 0 and 1 (0 indicating that the measurement is highly exceptional while a 1 indicates that the measurement is extremely normal). Based on a user-specified treshold level, also a simple quality label is assigned (“E” for exceptional, “N” for normal). In the case of “E” measurements, the results of the individual tests are stored and archived as part of the quality information.

Upon execution of the quality control, it is possible to view the results of the quality assignments, to "backtrack" the more detailled results of the tests for “E” measurements and to export the quality information (including the backtracking information) to the IDOD database.

Figure 3 shows as example the visual representation (that is generated by the SQC program) of a simple scheme that can be used to check amonium measurements. In this case the scheme consists of a single stage (each measurement is examined individually) that includes the 3 tests that have been earlier identified in Section 2 as reasonable quality control procedures: 1. a distribution test, where the measurement is checked against the distribution of amonium values at the measurement location and time of measurement, 2. a regression test, where the measurement is checked against other measurement values made at the same place and time, and 3. a variogram test, where the measurement is checked against other amonium measurements at the same time (or within a short timespan of that time) but at other locations. The a-priori probability shown in this scheme refers to the quality value that would be assigned if none of the tests can be applied (i.e. because values needed in that test are missing).



Visual representation of QC-definition scheme: AMONIUM-Scheme

Visual representation

This sheet gives a graphical representation of the QC-scheme.
The different stages in the Quality Control are situated on the different

	Apriori Probability	Test Nr 1	Test Nr 2	Test Nr 3
AMON	0.95	Distribution Test	Regression Test	Variogram Test

Figure 3: Example of a Quality Control Scheme for Amonium

4. Summary

The implementation of the statistical quality control of the oceanographic data in the form of an operational computer program has been practically completed and 3 specific tests that are generally applicable have been implemented as part of that program. In the upcoming year the program will be further tested, applied to real data and, when necessary, adjusted to meet specific problems encountered with the data.

Furthermore, a statistical study of a subset of actual data has been executed and this study has served as input to the development of the quality control program. The same study has also allowed UCS to familiarize itself with the data and will be useful in the development of a more general statistical tool to execute similar and more extensive studies in a routine manner. The development of such a general statistical analysis tool that is fully integrated with the IDOD database will be the main objective of the third year of the project.

References

UCS rapport, "Exploratory Statistical Analysis of the 1996 Belgian Monitoring Data for Sea Water", juni 1998.

UCS rapport, "Statistical Analysis of the 1977-1996 Contaminant Data for Seawater", september 1998.

UCS

Universitair Centrum voor Statistiek

APPENDIX 1:
EXPLORATORY STATISTICAL ANALYSIS OF THE 1996 BELGIAN MONITORING DATA FOR SEAWATER

June 1998

KATHOLIEKE UNIVERSITEIT
LEUVEN

University Center of Statistics
Katholieke Universiteit Leuven
de Croylaan 52b
B-3000 Leuven
Belgium

Table of Contents

1 Introduction

In this report an exploratory analysis of the 1996 Belgian monitoring data is described. The objective of this study is to consider how the different data obtained during the measurement campaign are correlated one to the other and to extract on this basis potential rules for quality control. The report is structured as follows:

- first the different data obtained during the campaign are described;
- next summary statistics, exploratory plots and the results of a correlation analysis of the raw data are discussed;
- based on these data it is decided to correct first of all the data for the effect of salinity both for the mean value and for the standard deviation. The results of such analysis are described in the following section;
- after removal of the effect of salinity it is evident that some spatial correlation and correlation with the other variables remains. The degree of this correlation is investigated in the subsequent section.

Conclusions of this study that are relevant to the statistical quality control of the data are summarized in the last section.

2 Description of the data

The dataset analyzed consists of contaminant data from the Belgian monitoring program of 1996 in the river Scheldt and the North Sea. A description of the different variables, the category they belong to and their units are given in table 1.

Table 1: Description of the data analyzed

Code	CAT	Parameter	Units	ICES reference
AMON	NUT	Ammonium	Moles/l	Annex 6 p. 41
NTRA	NUT	Nitrate	Moles/l	
NTRI	NUT	Nitrite	Moles/l	
NTRZ	NUT	Nitrate + Nitrite	Moles/l	
PHOS	NUT	Phosphate	Moles/l	
SLCA	NUT	Silicate	Moles/l	
DOXY	HYD	Dissolved oxygen	l/l	Annex 6 p. 38
PSALA	HYD	Salinity. Conductivity measurement by lab salinometer. From water sample	PS	Annex 6 p. 38
PSALD	HYD	Salinity. Conductivity/salinity measurement from CTD/STD (in situ)	PS	Annex 6 p. 39
SUSP	HYD	Suspended solids	g/l	Annex 6 p. 38
TEMP	HYD	Temperature	°C	Annex 6 p. 38
PHATL	PIG	Phaeophytin-a (Lorenz method)	g/l	Annex 6 p. 42
CPHLL	PIG	Chlorophyll-a (Lorenz method)	g/l	Annex 6 p. 42

The measurements are taken during 4 different cruises of the oceanographic vessel "Belgica" during the year 1996. These are given in table 2.

Table 2: Time period of the cruises in monitoring year 1996

Cruise number	Period
1	30/01/1996-01/02/1996
3	13/02/1996-13/02/1996
12	20/05/1996-23/05/1996
29	04/12/1996-10/12/1996

The data are measured in several stations located either in the North Sea or in the river Scheldt.

3 Summary Statistics, Exploratory Plots and Correlation Analysis of Raw Data

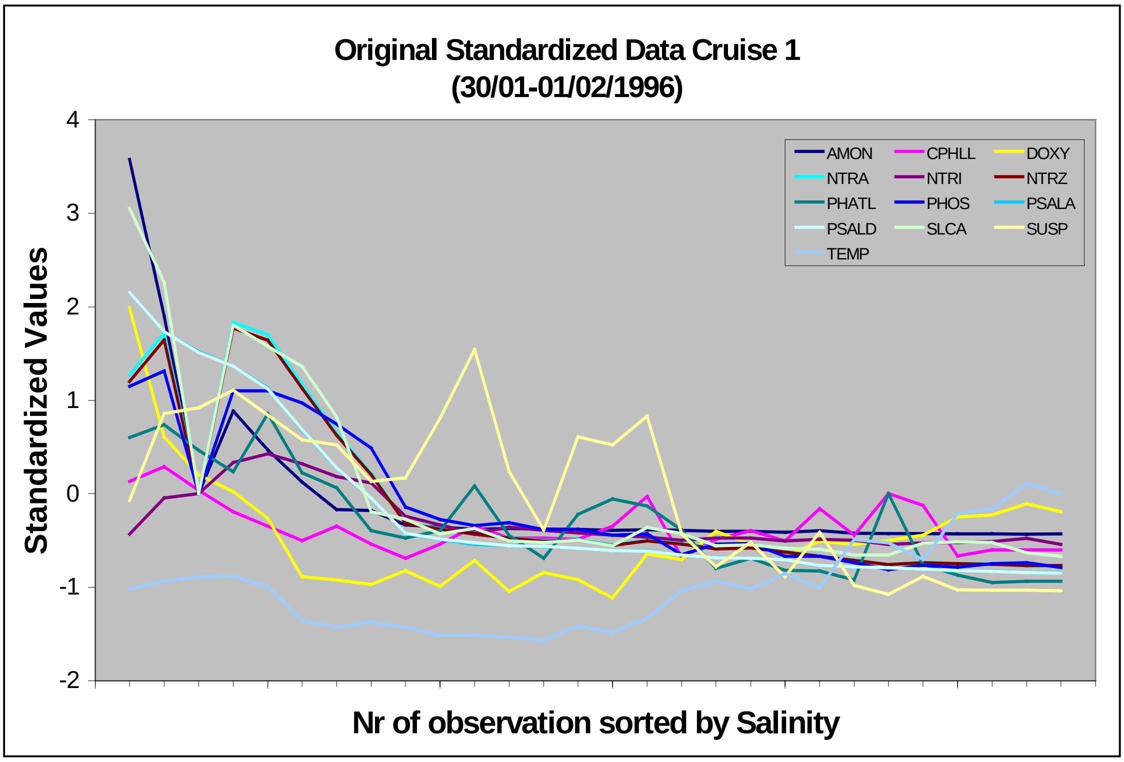
Table 3 summarizes various statistics (number of measurements, mean value, standard deviation, sum of all measurements, minimum and maximum) of the different measurements.

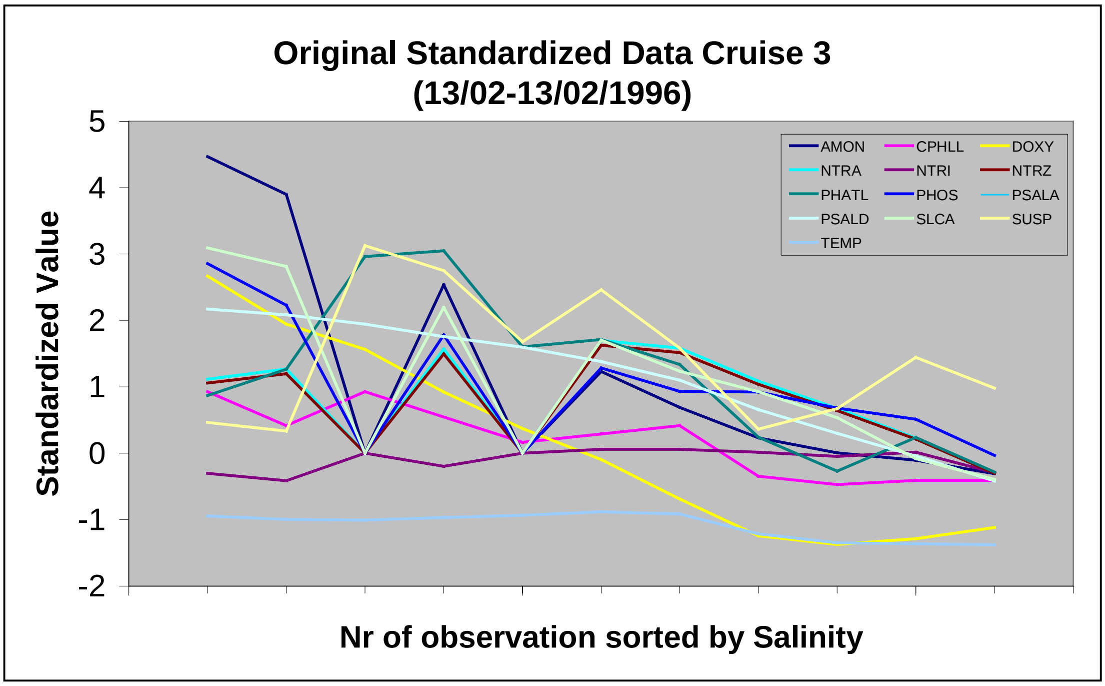
Table 3: Summary Statistics

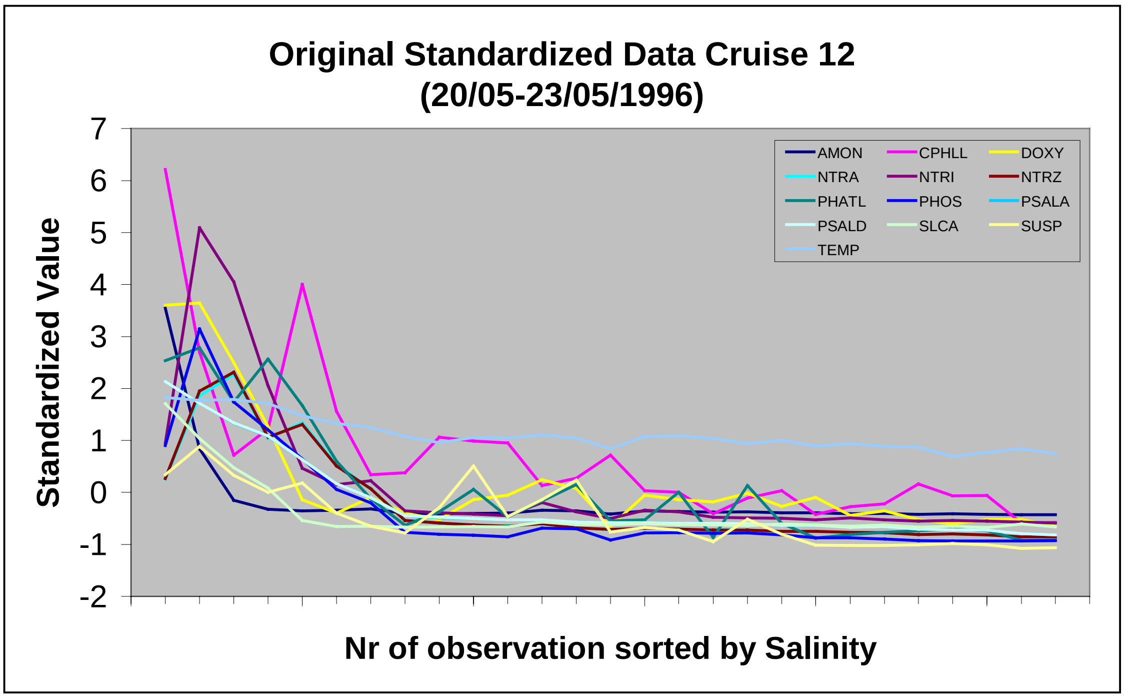
```
Simple Statistics

Variable            N          Mean       Std Dev           Sum       Minimum       Maximum  Label

AMON               87    0.00004792    0.00011046       0.00417          7E-8    0.00054129  AMON
CPHLL              88    0.00000195    0.00000268    0.00017152          1E-7    0.00001857  CPHLL
DOXY               90       0.00689       0.00177       0.62007    0.00045100       0.00932  DOXY
NTRA               87    0.00009323    0.00010889       0.00811          1E-7    0.00041981  NTRA
NTRI               87    0.00000230    0.00000383    0.00020024          5E-8    0.00002182  NTRI
NTRZ               87    0.00009668    0.00011289       0.00841          1E-7    0.00043048  NTRZ
PHATL              88    0.00000334    0.00000341    0.00029352          1E-7    0.00001373  PHATL
PHOS               87    0.00000229    0.00000245    0.00019938          1E-8    0.00000999  PHOS
SLCA               87    0.00005309    0.00007328       0.00462    0.00000382    0.00027981  SLCA
SUSP               90       0.03338       0.02932       3.00460       0.00180       0.12510  SUSP
PSALA              90      25.44579      11.27950    2290.12101       0.50000      34.98500  PSALA
PSALD              90      25.44861      11.26479    2290.37501       0.49500      34.98500  PSALD
```

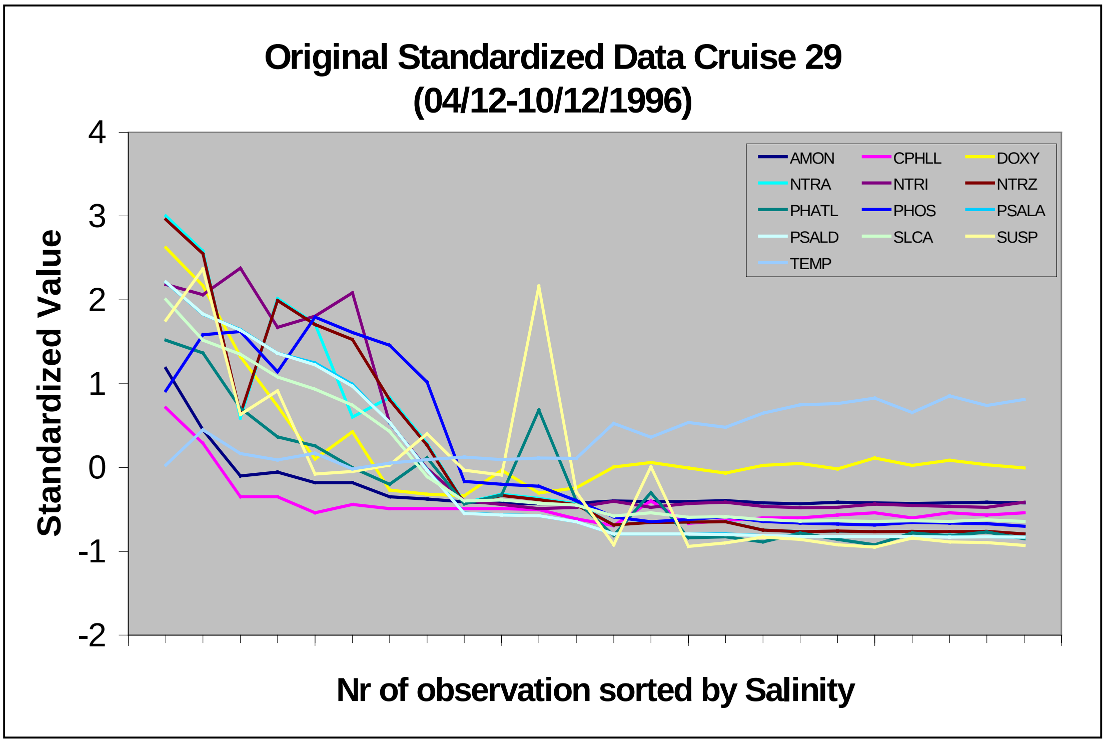
It is clear that for most of the variables the standard deviation is very high relative to the mean value of the variable. Because of this large dispersion, one would therefore expect that a simple check against a statistical upper-bound value will not be very effective in a quality control.

The absolute value of many of the measurements is very small. To facilitate the further interpretation of the results, the original data for each variable have been therefore standardized by subtracting the mean value and scaling the residual with the sample standard deviation. The following figures show for each cruise the standardized measurements where the observations have been ordered for increasing salinity. In these plots, the sign of the variables DOXY, PSALA and PSALD have been inverted.



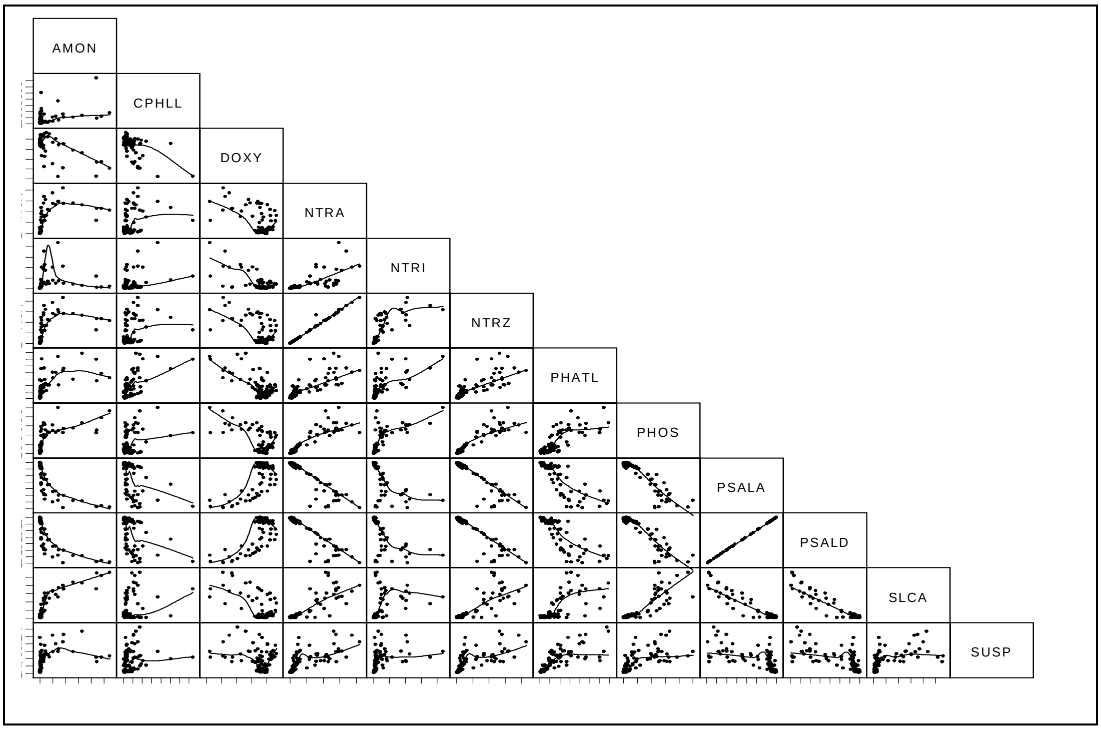

Original Standardized Data Cruise 3
(13/02-13/02/1996)
Standardized Value
5
4
3
2
1
0
-1
-2
Nr of observation sorted by Salinity
AMON
CPHLL
DOXY
NTRA
NTRI
NTRZ
PHATL
PHOS
PSALA
PSALD
SLCA
SUSP
TEMP


Original Standardized Data Cruise 12
(20/05-23/05/1996)
Standardized Value
7
6
5
4
3
2
1
0
-1
-2
Nr of observation sorted by Salinity
AMON
CPHLL
DOXY
NTRA
NTRI
NTRZ
PHATL
PHOS
PSALA
PSALD
SLCA
SUSP
TEMP



The previous plots show very clearly that all of the variables have a similar trend with a nearly constant value within a cruise for high salinity values (in the North Sea) and decreasing standardized values as one moves within the river Scheldt towards the sea (increasing salinity values). The dispersion of the measurements within the river Scheldt is clearly much higher than that in the North Sea and, also from cruise to cruise, there are notable differences which would indicate that there are important variations depending on the timeperiod of the measurement.

The figure below shows a scatterplot matrix of the different variables. The solid line in the different subplots corresponds to a nonparametric regression line of the Y-variable (indicated at the top) versus the X-variable (indicated at the right side) obtained through loess smoothing (the loess smoothing uses for different values X a weighted average of the Y-variables within a local neighborhood). .



The scatterplot matrix shows that there is a strong correlation of each of the variables to SALINITY (PSALA or PSALD) as has been already noted in the earlier exploratory plots.

The correlation value, together with the p-value of the Pearson correlation coefficient, are shown in the following table for each combination of the variables. The p-value when very low (e.g. <1%) indicates that the correlation is statistically significant.

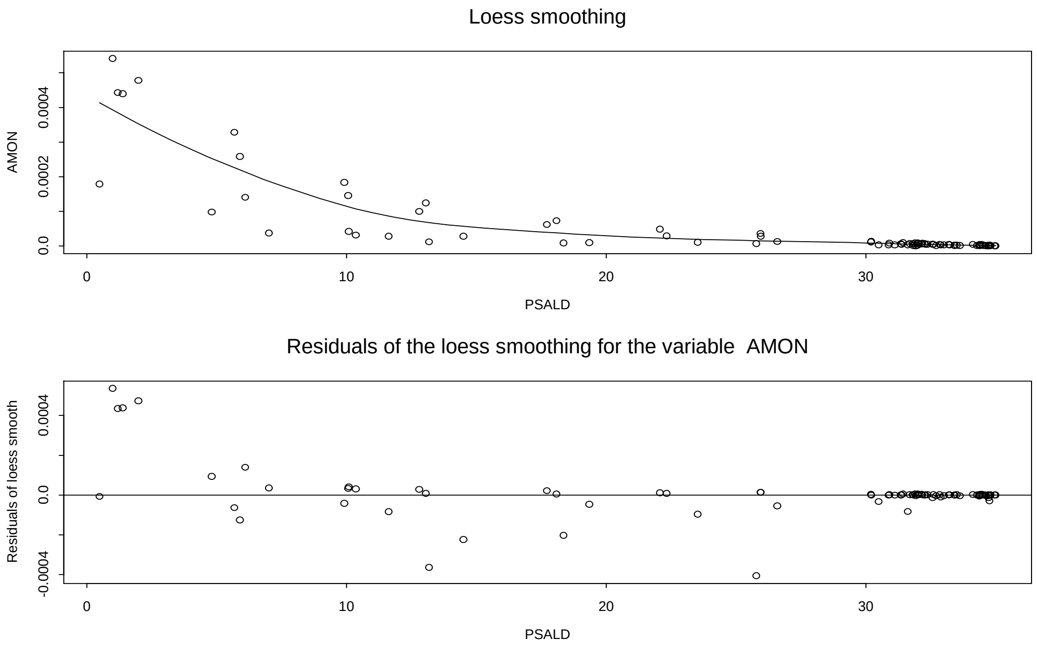
```
   Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / Number of Observations
                 AMON         CPHLL          DOXY          NTRA          NTRI          NTRZ
  AMON        1.00000       0.44601      -0.65547       0.54934       0.16578       0.53364
  AMON            0.0        0.0001        0.0001        0.0001        0.1249        0.0001
                   87            85            87            87            87            87
  CPHLL       0.44601       1.00000      -0.57117       0.32443       0.37583       0.32132
  CPHLL        0.0001           0.0        0.0001        0.0025        0.0004        0.0027
                   85            88            88            85            85            85
  DOXY       -0.65547      -0.57117       1.00000      -0.53235      -0.64424      -0.53999
  DOXY         0.0001        0.0001           0.0        0.0001        0.0001        0.0001
                   87            88            90            87            87            87
  NTRA        0.54934       0.32443      -0.53235       1.00000       0.70239       0.99462
  NTRA         0.0001        0.0025        0.0001           0.0        0.0001        0.0001
                   87            85            87            87            87            87
  NTRI        0.16578       0.37583      -0.64424       0.70239       1.00000       0.73294
  NTRI         0.1249        0.0004        0.0001        0.0001           0.0        0.0001
                   87            85            87            87            87            87
  NTRZ        0.53364       0.32132      -0.53999       0.99462       0.73294       1.00000
  NTRZ         0.0001        0.0027        0.0001        0.0001        0.0001           0.0
                   87            85            87            87            87            87
  PHATL       0.60574       0.63604      -0.65508       0.79368       0.65395       0.78852
  PHATL        0.0001        0.0001        0.0001        0.0001        0.0001        0.0001
                   85            88            88            85            85            85
  PHOS        0.65430       0.33156      -0.59708       0.88986       0.70831       0.89895
  PHOS         0.0001        0.0019        0.0001        0.0001        0.0001        0.0001
                   87            85            87            87            87            87
  SLCA        0.86212       0.27882      -0.60090       0.83527       0.43483       0.82808
  SLCA         0.0001        0.0098        0.0001        0.0001        0.0001        0.0001
                   87            85            87            87            87            87
  SUSP        0.41369       0.22657      -0.27908       0.70627       0.38980       0.69468
  SUSP         0.0001        0.0338        0.0077        0.0001        0.0002        0.0001
                   87            88            90            87            87            87
  PSALA      -0.77100      -0.46903       0.69283      -0.92200      -0.65373      -0.92276
  PSALA        0.0001        0.0001        0.0001        0.0001        0.0001        0.0001
                   87            88            90            87            87            87
  PSALD      -0.77174      -0.47004       0.69318      -0.92203      -0.65307      -0.92251
  PSALD        0.0001        0.0001        0.0001        0.0001        0.0001        0.0001
                   87            88            90            87            87            87

                PHATL          PHOS          SLCA          SUSP         PSALA         PSALD
  AMON        0.60574       0.65430       0.86212       0.41369      -0.77100      -0.77174
  AMON         0.0001        0.0001        0.0001        0.0001        0.0001        0.0001
                   85            87            87            87            87            87
  CPHLL       0.63604       0.33156       0.27882       0.22657      -0.46903      -0.47004
  CPHLL        0.0001        0.0019        0.0098        0.0338        0.0001        0.0001
                   88            85            85            88            88            88
  DOXY       -0.65508      -0.59708      -0.60090      -0.27908       0.69283       0.69318
  DOXY         0.0001        0.0001        0.0001        0.0077        0.0001        0.0001
                   88            87            87            90            90            90
  NTRA        0.79368       0.88986       0.83527       0.70627      -0.92200      -0.92203
  NTRA         0.0001        0.0001        0.0001        0.0001        0.0001        0.0001
                   85            87            87            87            87            87
  NTRI        0.65395       0.70831       0.43483       0.38980      -0.65373      -0.65307
  NTRI         0.0001        0.0001        0.0001        0.0002        0.0001        0.0001
                   85            87            87            87            87            87
  NTRZ        0.78852       0.89895       0.82808       0.69468      -0.92276      -0.92251
  NTRZ         0.0001        0.0001        0.0001        0.0001        0.0001        0.0001
                   85            87            87            87            87            87
  PHATL       1.00000       0.79346       0.69867       0.76848      -0.84549      -0.84618
  PHATL           0.0        0.0001        0.0001        0.0001        0.0001        0.0001
                   88            85            85            88            88            88
  PHOS        0.79346       1.00000       0.86129       0.65233      -0.92103      -0.92059
  PHOS         0.0001           0.0        0.0001        0.0001        0.0001        0.0001
                   85            87            87            87            87            87
  SLCA        0.69867       0.86129       1.00000       0.62376      -0.93752      -0.93767
  SLCA         0.0001        0.0001           0.0        0.0001        0.0001        0.0001
                   85            87            87            87            87            87
  SUSP        0.76848       0.65233       0.62376       1.00000      -0.68238      -0.68300
  SUSP         0.0001        0.0001        0.0001           0.0        0.0001        0.0001
                   88            87            87            90            90            90
  PSALA      -0.84549      -0.92103      -0.93752      -0.68238       1.00000       0.99999
  PSALA        0.0001        0.0001        0.0001        0.0001           0.0        0.0001
                   88            87            87            90            90            90
  PSALD      -0.84618      -0.92059      -0.93767      -0.68300       0.99999       1.00000
  PSALD        0.0001        0.0001        0.0001        0.0001        0.0001           0.0
                   88            87            87            90            90            90
```

The previous results show that the correlation is significant and quite large between many of the different variables. Still, the correlation with salinity values is for most of the variables one of the most outspoken. Since these salinity values are also indirectly an indication of the spatial position of the measurement (in the sea or at a given distance within the Scheldt) it is decided to further investigate the measurements after removal of this dependence to search for significant correlations. This analysis is explained in the following section.

4 Analysis of the Data after Correction for Salinity Effects

4.1 Removal of the Influence of Salinity on the Mean Value

To remove the influence of salinity on the mean value of the observations, again a nonparametric regression (in this case obtained through a loess smoothing with span 0.75) is used. The next figure shows the loess regression line for the case of AMON as well as the residuals of the original measurements of AMON with respect to this loess smoothing. Similar plots for the other variables can found in the full report.

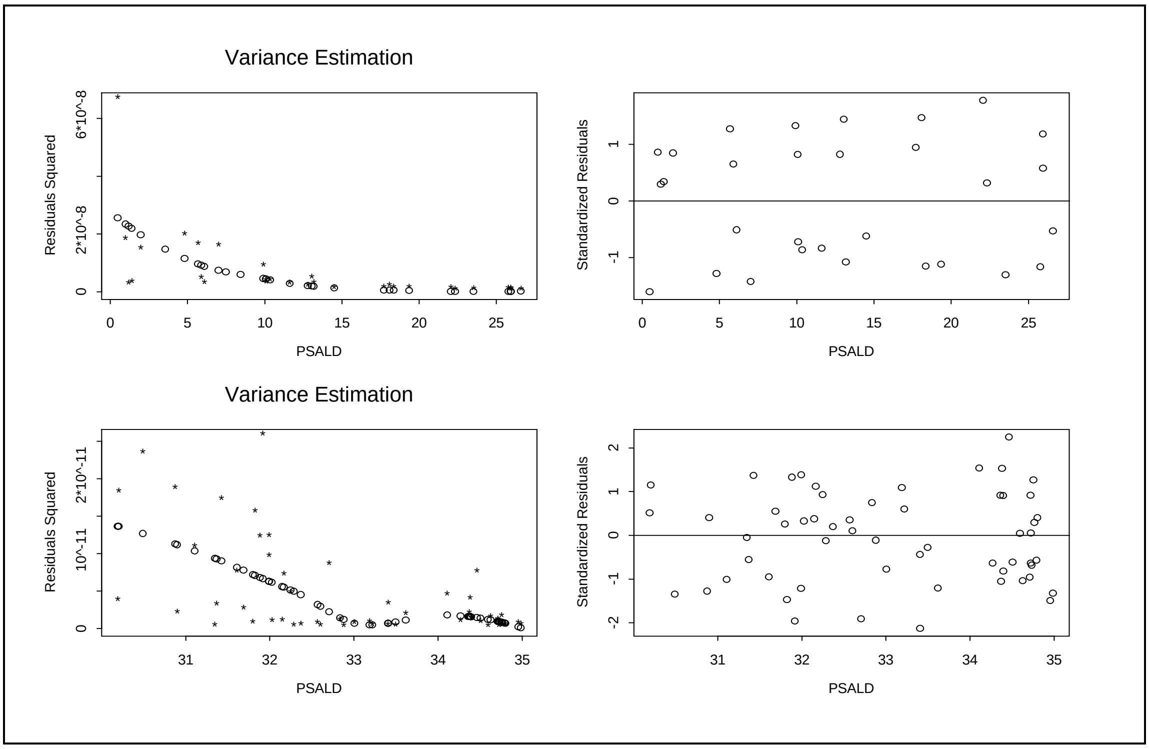


4.2 Removal of the Influence of Salinity on the Standard Deviation

The plots of the loess residuals clearly indicate that the dispersion of the residuals is much higher for low salinity values (measurements in the river Scheldt) than for high salinity values (measurements in open sea). The fact that the variance of the residuals is not constant is referred to as heteroscedasticity and it is found that this characteristic introduces spurious correlations between the different variables. For this reason it is decided to estimate the influence of the salinity on the variance and to standardize the residuals with the corresponding estimate of the standard deviation.

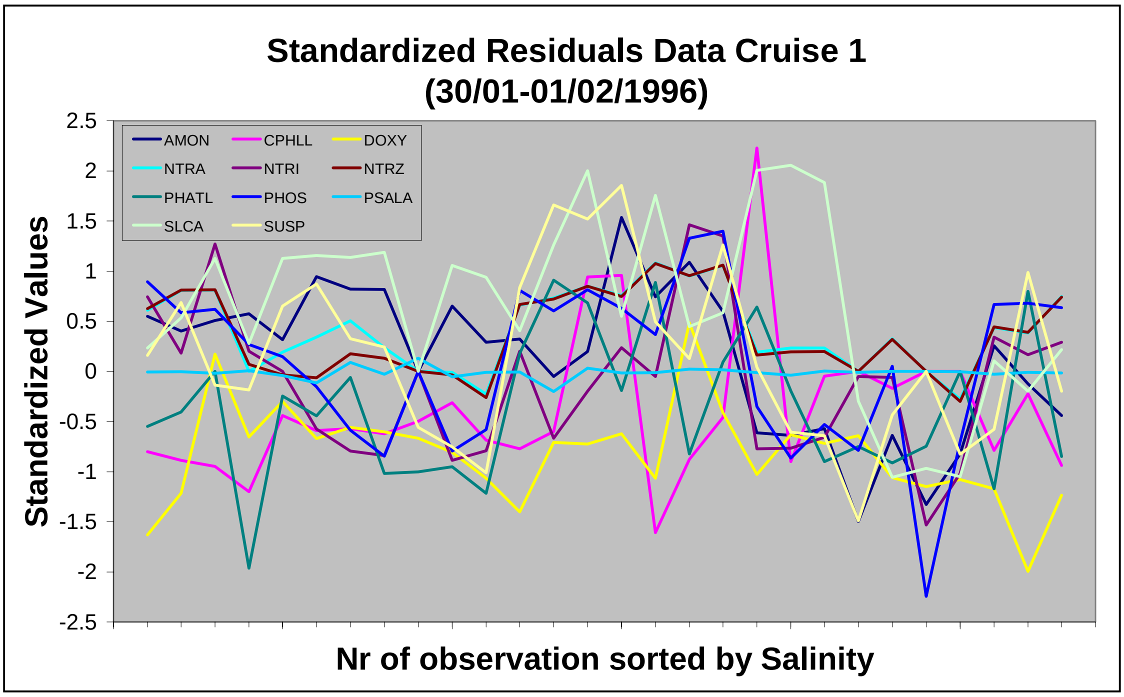
The figure below illustrates how this has been done for the case of the AMON variable. Similar plots for the other variables can be found in the full report.

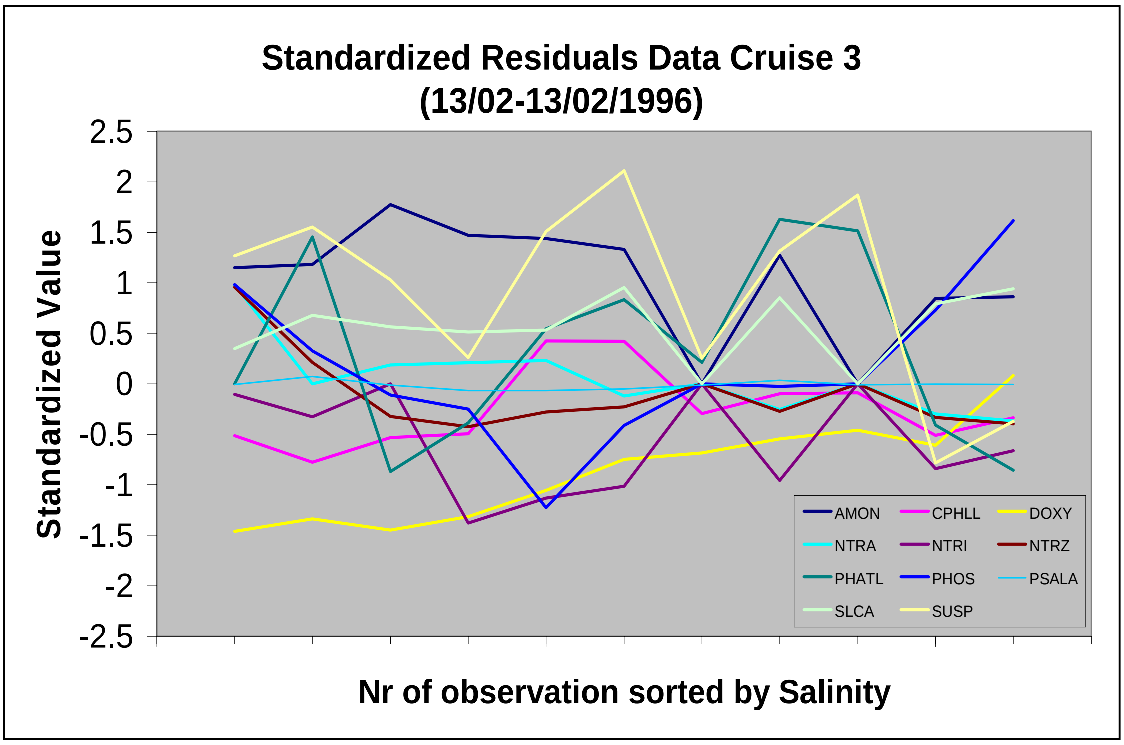
The top left figure refers to a loess smoothing regression of the squared residuals of AMON for salinity values less than 30 (the loess regression line is indicated by the larger circles, while the squared residuals are indicated by the smaller dots). At the right side of this plot the standardized residuals (the residual scaled with the corresponding estimate of the standard deviation) is shown. The figure below shows similar results for salinity values larger than 30. In this last figure one notices that the variance decreases for salinity values in the range of 30 to 33 and then converges to a constant level. This might indicate that the salinity threshold of 30 that is used to separate the Scheldt' measurements from measurements in the open sea should be increased. One may remark that for the open sea condition (if in the case of AMON redefined for a 33 threshold value of salinity), the application of the loess smoothing is perhaps superfluous and a constant value might have been alternatively assumed.

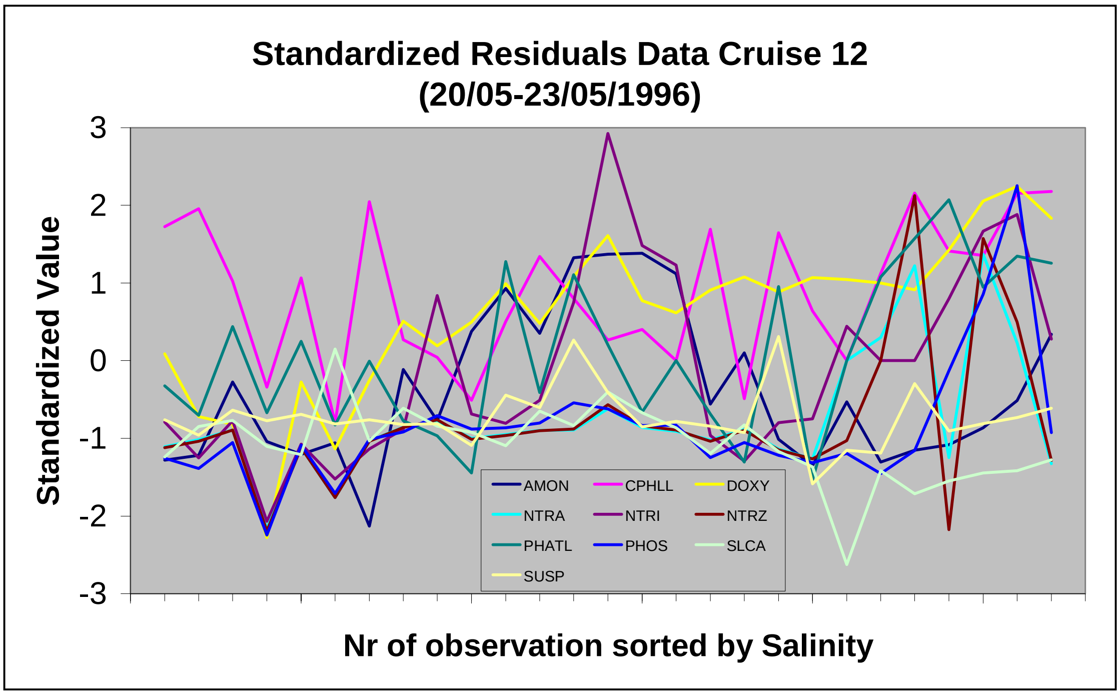


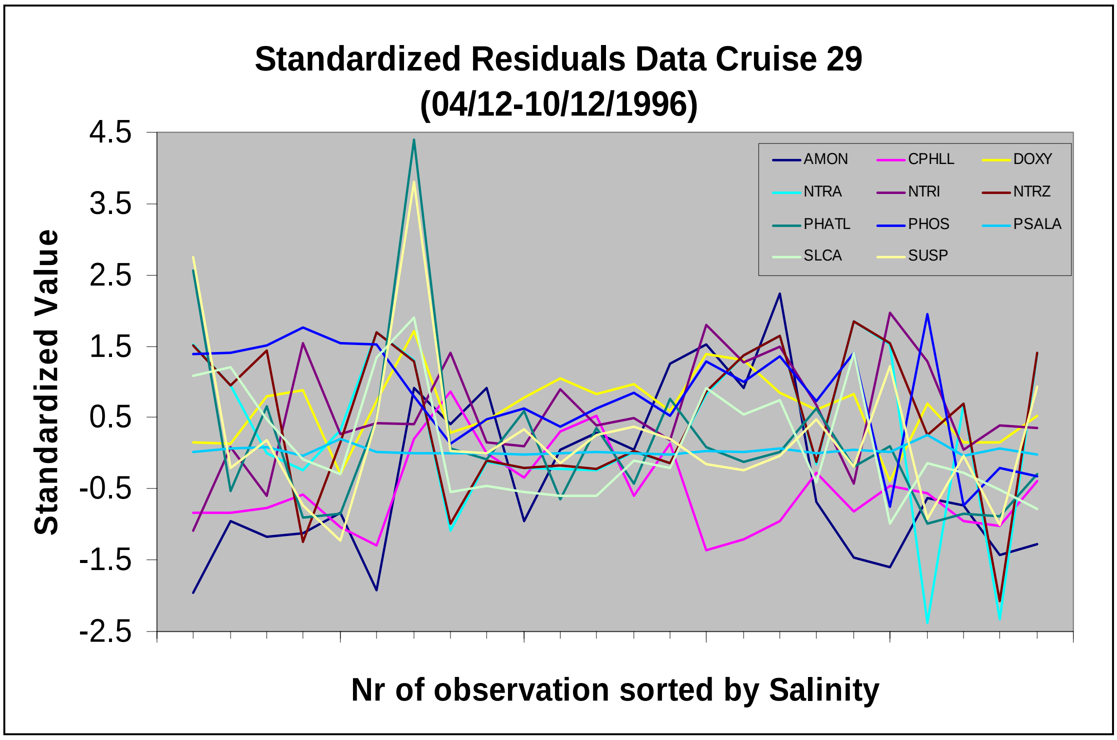
4.3 Plot of Standardized Residuals after Removal of the Influence of Salinity

The standardized residuals after removal of the salinity effect (on mean value and standard deviation) are shown below for each cruise separately.









The previous plots show that, after removal of the influence of salinity, the standardized residuals within a given cruise tend to be either above or below the mean value (of zero). This feature is especially apparent for Data Cruise 3. Such a behavior is indicative of correlation between the same variable at different locations. Supposedly this correlation is induced by temporal effects: i.e. the standardized residuals tend to have the same sign for measurements made within the same cruise. For some other variables, there appears to be some evidence of correlation between the different variables at the same measurement point (e.g. DOXY and NITRI in Data Cruise 1 show a similar profile).

Both the degree of spatial dependency and the possible correlation of the standardized residuals with other variables is investigated in the following sections.

5 Spatial Correlation and Correlation with Other Variables of the Standardized Residuals

5.1 *Spatial Dependency*

Most geographical data show some spatial dependency. If the spatial dependency is very pronounced, this feature can be used to check one measurement value against another at a nearby location. One way to characterize this dependency is to determine the variance of differences between measurements of the same variable as a function of the distance between these measurements; Such a curve is referred to as a semi-variogram, which mathematically corresponds to one half of this variance, and for N measurements at distance h of each other can be estimated as

$$\gamma(h) = \frac{1}{2N(h)} \sum_{|s_i - s_j| = h} (y_i - y_j)^2$$

where

γ(h) is the semi-variance at distance h,
h is the lag (this is the distance between sample locations),
y_I is the value of variable y at location s_I,
N(h) is the number of pairs of observed data points.

For the purpose of this analysis, the Euclidean distance between two sample locations is calculated using the following approximate formula:

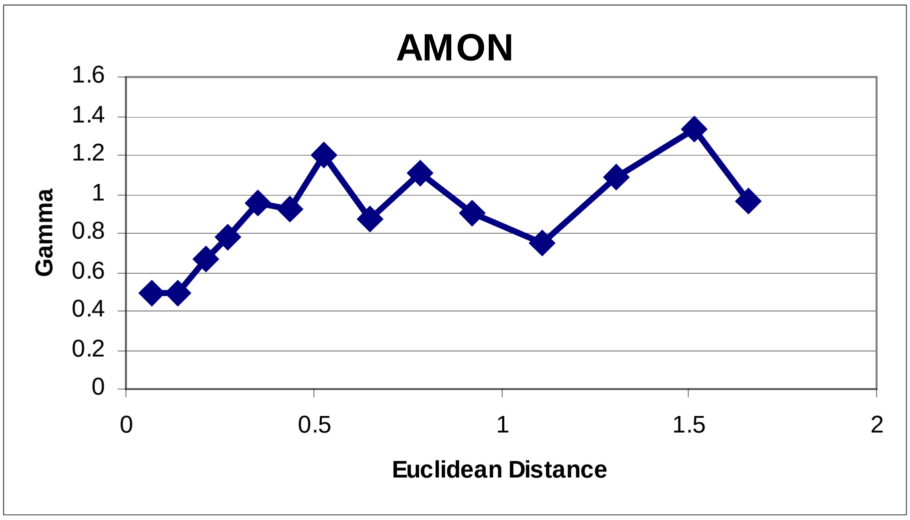
$$h = \frac{111.1}{100} \sqrt{(x_2 - x_1)^2 + \left[\cos\left(\frac{x_2 + x_1}{2}\right)(y_2 - y_1)\right]^2}$$

where

h is the Euclidean distance,
x_I is the latitude of the location i,
y_I is the longitude of the location i,

To obtain reliable results the number of data N(h) used in the estimation of the semi-variogram must be sufficiently large. For this reason, the distances between the different measurement points are grouped into classes of increasing distance which contain each 65 measurement differences. Measurement differences are considered only within the same cruise (e.g. for data where the time differences is less than 4 days) to exclude the comparison of measurements that are affected by seasonal changes.

The following figure shows the result that is obtained for the AMON variable. Similar results for the other variables are shown in the full report.



Excluding random fluctuations, the semi-variogram shows a clear upwards trend as a function of the Euclidean distance. At a distance of 0.1 (i.e. 10 km) the variance is about one half of that noted at large distances. The fact that the curve converges to a value of about 1 is to be expected since the semi-variogram is applied to the standardized values (which have a mean value of 0 and standard deviation of 1) and thus one would expect to find a value of 1 if the residuals are uncorrelated.

The previous results show convincingly that for most of the variables predictions on the basis of neighboring measurements can significantly reduce the variance (and thus lead to a better quality control). Nevertheless, we expect that after removal of this dependency there may remain an even stronger or equally important correlation with other variables measured at the same sampling point. This point is further investigated in the next section.

5.2 Stepwise regression

To investigate whether the standardized residuals are correlated one against the other after accounting for spatial dependency with neighboring measurements of the same variable, it is decided to use a stepwise regression technique: in this analysis the variable of interest (the standardized residual) is regressed against the nearest measurement of the same variable within the same campaign as well as against all of the other variables (the standardized residuals thereof). Within this set of potential regressor candidates, the method then iteratively searches for those variables that are significantly correlated, by stepwise inclusion or exclusion of potential regressor candidates until a stable point is found. Specifically, this means that at this point all regressor variables are found to be statistically significant (at a significance level of 5%) and none of the excluded variables (if added one by one) improves the regression result (at a significance level of 5%).

The results below show the end result for the case of the AMON variable. Results for the other variables are given in the full report.

```
                              Analysis of Variance
                                 Sum of         Mean
         Source          DF     Squares       Square      F Value       Prob>F

         Model            2    20.33553     10.16776       13.295       0.0001
         Error           63    48.18208      0.76479
         C Total         65    68.51760

             Root MSE       0.87453     R-square       0.2968
             Dep Mean      -0.10039     Adj R-sq       0.2745
             C.V.        -871.16883

                              Parameter Estimates

                  Parameter      Standard    T for H0:                Standardized
Variable  DF       Estimate         Error   Parameter=0    Prob > |T|      Estimate

INTERCEP   1      -0.095394    0.10828652        -0.881        0.3817    0.00000000
AMON2      1       0.471945    0.10600039         4.452        0.0001    0.47100667
NTRIRES    1       0.252479    0.10655589         2.369        0.0209    0.25066393
```

In the case of the AMON variable, it is found that two variables are retained in the model: the standardized residual of the nearest measurement of AMON (AMON2) and the standardized residual of the NTRI measurement at the same location. While the former variable is highly significant, the p-value of the T-statistic of the regression coefficient of the second regressor variable is 2%.

The table below summarizes the different regressor variables that are thus retained for the different variables:

variable	regressor variables
AMON	AMON2, NTRI
CPHLL	CPHLL2, NTRA, PHATL
DOXY	DOXY2, NTRI
NTRA	NTRZ
NTRI	NTRI2, DOXY, PHOS
NTRZ	NTRA, PHOS
PHATL	CPHLL, DOXY, SUSP
PHOS	PHOS2, NTRA, NTRI, NTRZ
SLCA	SLCA2
SUSP	DOXY, NTRA, PHATL, AMON

6 Conclusions with respect to Quality Control

The measurement data (when considered over different cruises and different locations) show a very high dispersion and therefore a simple check against an upper- and lower-bound based on the marginal distribution is not expected to be very effective.

Exploratory analysis shows that part of the dispersion can be removed by considering the salinity value. Since this value is primarily an indicator of location, one could alternatively consider the marginal distribution at a specific location.

The dispersion of the values varies however strongly with salinity (or alternatively location) with higher dispersions noted in the Scheldt river, in particular at low salinity values.

After elimination of the effect of salinity (or location) it is found that for most of the variables some further reduction of the dispersion can be made either by considering similar measurements at a nearby location (which is equivalent to consider a seasonal or time effect) or by considering the value of other measurements at the same location. The list of significant variables that are thus retained are summarized in the last table of the previous section and could form a starting point for empirical relationships that should be used in the statistical quality control.

The relationships thus found are however based on a limited set of data. Therefore it is planned to repeat the same analysis on a larger dataset. Furthermore, the physical significance of the different relationships should be considered and for this reason, presentation and discussion of the results with BMM is planned.

UCS
Universitair Centrum voor Statistiek

APPENDIX 2:
STATISTICAL ANALYSIS OF THE 1977-1996 MONITORING DATA FOR SEAWATER

September 1998

KATHOLIEKE UNIVERSITEIT
LEUVEN

University Center of Statistics
Katholieke Universiteit Leuven
de Croylaan 52b
B-3000 Leuven
Belgium

Table of Contents

1 Introduction

In this report the contaminant data gathered from 1977 until 1996 of the Belgian monitoring program are analyzed. The objective of this analysis is to examine characteristics of the data could be used in a statistical quality control and to identify eventual problems that may be posed in extracting and/or applying such characteristics.

Specifically, following characteristics will be considered:

- the univariate distribution of the measured variable. Such a distribution could be potentially used in a quality control check by comparing a measurement value with lower- and upper-bounds that are unlikely to be exceeded;
- the spatial correlation of the measured variable. Such correlation could be potentially used in a quality control check by comparing a measurement with nearby measurements, if any;
- the correlation of one measured variable with one or several other measurements. Such correlation could be potentially used in a quality control check by comparing a measurement value with the value that is expected on the basis of another measurement.

It should be noted that this study certainly does not aim to make a final and exhaustive study of all of the data. Instead, the study should be considered a search for guidelines for such a more definitive study, which may be performed for each variable separately with the help of the software still to be developed under the IDOD research project.

The results of this study are presented in this report as follows:

- In a first chapter the data obtained during the monitoring years are briefly described and a selection is made of the data considered in more detail in this study;

- in the following chapter, various summary statistics of the data are presented and the need to transform the basic variables is considered;

- in the next chapter it is explained how results for each of the three statistical characteristics are obtained. The analysis is illustrated for one of the measurement variables (AMON). Complete results for all different measurement variables considered in this study are shown in the full report. In the present chapter these results are only summarized.

We conclude this report with a summary of the main results of interest for the implementation of a statistical quality control.

2 Description of the data

2.1 Details on the variables

The dataset analyzed consists of the contaminant data of the monitoring program from 1977 until 1996 in the river Scheldt and the North Sea. The different types of measurements are listed the full report together with a short description, the type of variable, the units used and the ICES reference.

2.2 Overview of the monitoring program

While the number of variables measured is very large, further analysis shows that few of these variables have been consistently measured over the period of 19 years covered by the monitoring program. An overview of the number of measurements in each year and for each variable is shown also given in the full report.

2.3 Choice of the data to be analyzed

In the list of measurements in the report it is clearly shown that a majority of the variables are only sparingly measured. As previously mentioned, the objective of this study is not to analyze each of the variables, but rather to verify to which extent different statistical characteristics can be typically used for quality control checking and to demonstrate how such characteristics can be extracted. Because of this exploratory nature, the present study limits itself to the consideration of only those variables for which a substantial number of measurements have been made so that conclusions reached in this study are less sensitive to random error. Specifically following measurement variables are considered: AMON, CD, CDP, CDPS, CPHLC, CPHLH, CPHLL, CU, CUPS, DOXY, HG, HGPL, NTRA, NTRI, NTRZ, PB, PBP, PHATL, PHOS, SAL, SUSP, ZN, ZNP and ZNPS.

In the original data salinity is indicated by 5 different variables (PSAL, PSAL?, PSALA, PSALC, PSALD), which are apparently associated with different measurement techniques. In this study, the different variables are all considered measurements of the same basic variable SAL and, if more than one measurement is available at the same time and location, the average of the measured values is used.

In this study only those data for which the location and the time of measurement is clearly defined are retained.

2.4 Overview of the locations

Each measurement of the monitoring program is associated with a particular station. Maps of the different stations that occur in the campaign are in the report.

3 Summary Statistics and Exploratory Data Analysis

The Table below summarizes a variety of elementary statistics (minimum value, lower quartile, mean, median, upper quartile, maximum, standard deviation and number of measurements) for each of the variables considered in this study.

Table 4: Summary Statistics

	AMON	CD	CDP	CDPS	CPHLC	CPHLL	CPHLH
Minimum:	0.00E+00	2.00E-09	2.50E-10	5.00E-08	5.00E-08	0.00E+00	1.94E-07
1st Quantile:	2.40E-06	3.20E-08	1.38E-08	8.30E-07	1.13E-06	8.50E-07	1.78E-06
Mean:	1.84E-05	1.54E-07	2.15E-07	5.45E-06	4.87E-06	4.39E-06	8.36E-06
Median:	5.40E-06	6.40E-08	3.86E-08	1.40E-06	2.50E-06	2.14E-06	3.94E-06
3rd Quantile:	1.18E-05	1.51E-07	9.44E-08	2.90E-06	5.44E-06	4.98E-06	9.79E-06
Maximum:	5.94E-04	3.50E-06	2.17E-05	2.19E-04	7.40E-05	8.96E-05	9.76E-05
St.Dev	5.05E-05	2.89E-07	1.45E-06	1.76E-05	7.01E-06	6.50E-06	1.17E-05
Number:	2460	980	230	250	2060	2270	590

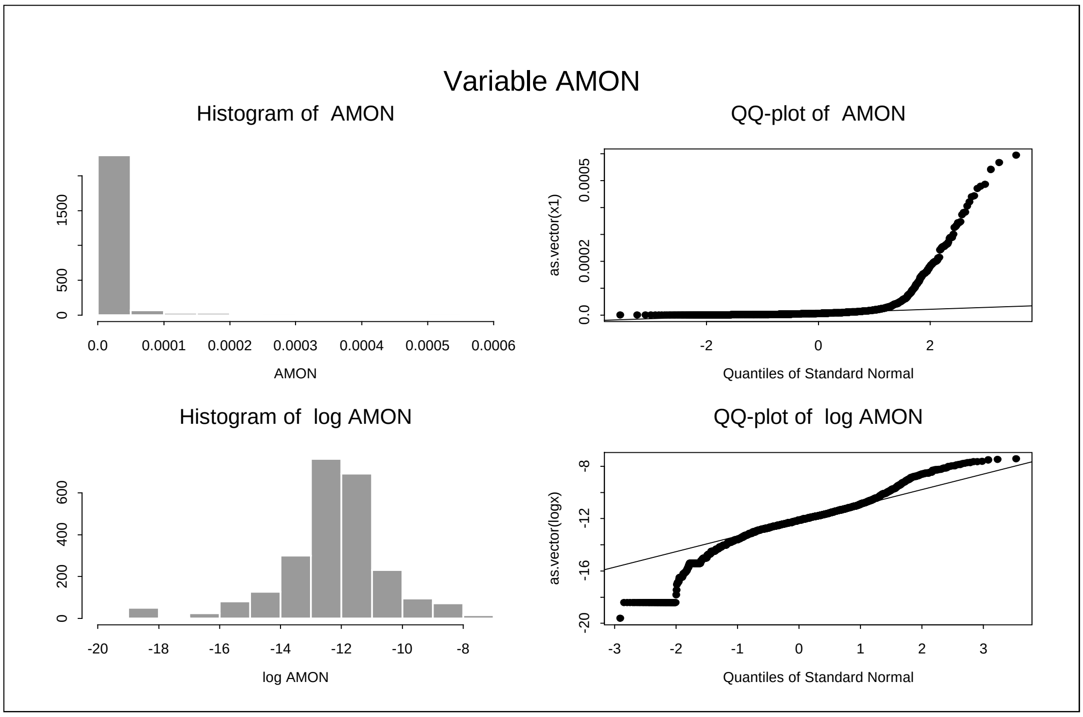
	CU	CUP	CUPS	DEPH	DOXY	HG	HGPL
Minimum:	3.00E-08	1.99E-08	5.60E-06	0	0.00E+00	2.10E-10	4.00E-09
1st Quantile:	6.80E-07	3.25E-07	1.80E-05	3	4.40E+00	1.40E-08	1.70E-08
Mean:	2.90E-06	2.05E-06	1.17E-04	3.35	7.55E+00	6.86E-08	1.15E-07
Median:	1.20E-06	9.79E-07	3.00E-05	3	9.00E+00	3.80E-08	5.00E-08
3rd Quantile:	2.48E-06	2.12E-06	6.70E-05	4	1.05E+01	8.80E-08	1.15E-07
Maximum:	2.19E-04	2.31E-05	5.65E-03	15	1.98E+01	7.87E-07	1.04E-06
St.Dev	8.76E-06	3.45E-06	4.26E-04	1.27	4.51E+00	8.51E-08	1.95E-07
Number:	910	210	230	3877	890	810	150

	NTRA	NTRI	NTRZ	PB	PBP	PBPS	PHATL
Minimum:	0.00E+00	0.00E+00	6.00E-08	1.80E-08	8.00E-09	5.00E-06	-6.90E-05
1st Quantile:	6.59E-06	4.10E-07	5.30E-06	1.40E-07	3.76E-07	2.50E-05	4.49E-07
Mean:	8.15E-05	2.65E-06	4.94E-05	2.33E-06	2.89E-06	7.46E-05	2.92E-06
Median:	2.13E-05	8.00E-07	1.60E-05	3.50E-07	9.80E-07	4.40E-05	1.43E-06
3rd Quantile:	1.08E-04	2.57E-06	3.90E-05	1.90E-06	2.94E-06	8.40E-05	3.70E-06
Maximum:	5.38E-04	6.02E-05	6.58E-04	8.48E-05	7.17E-05	1.03E-03	3.65E-05
St.Dev	1.20E-04	5.08E-06	8.96E-05	6.59E-06	6.37E-06	1.03E-04	4.65E-06
Number:	580	580	2610	490	210	230	1940

	PHOS	SLCA	SUSP	ZN	ZNP	ZNPS	SAL
Minimum:	0.00E+00	0.00E+00	1.00E-03	1.30E-07	7.00E-09	1.00E-06	0.17
1st Quantile:	6.30E-07	3.29E-06	1.48E-02	2.57E-06	1.16E-06	5.90E-05	30.67
Mean:	2.46E-06	2.94E-05	4.84E-02	2.67E-05	5.64E-06	1.94E-04	29.37
Median:	1.24E-06	1.10E-05	3.08E-02	6.31E-06	2.44E-06	1.02E-04	32.92
3rd Quantile:	3.14E-06	2.67E-05	5.96E-02	1.94E-05	6.57E-06	1.75E-04	34.2
Maximum:	9.20E-05	4.57E-04	7.66E-01	1.56E-03	5.34E-05	2.67E-03	37.97
St.Dev	3.69E-06	5.07E-05	5.82E-02	8.76E-05	8.18E-06	3.18E-04	8.92
Number:	1570	1640	3030	810	240	250	3330

The univariate statistics readily show that for many of the variables the assumption of a normal distribution is not appropriate: the median value differs substantially from the mean value and the upper- and lower-quartile as well as the minimum and maximum are highly asymmetric relative to the mean. Since in most of these cases the data are clearly skewed to the right and the measurements are known to be always positive, it is natural to consider as a potential alternative modeling assumption a lognormal distribution (the logarithm of the measured value is assumed to follow a normal distribution).

To investigate whether such a logarithmic transformation improves the "normality" of the data a histogram and a QQ-normality plot has been made for the data, before and after applying a logarithmic transformation. The plots that are thus obtained for the variable AMON are shown below. Graphs for the other measurement variables are shown in the report.



Normality of the data can be verified either by recognizing the typical bell-shape in the histogram or by verifying that the QQ-plot follows (approximately) a straight line.

In the case of the AMON variable it is quite clear that the logarithmic transformation is essential to obtain a distribution similar to a normal one. Even after the logarithmic transformation one can note in the QQ plot that the normal assumption does not strictly apply: the deviation in the lower tail is explained by the fact that some data are measured near the detection limit and are only approximately reported. The deviation noted in the upper tail is less obvious to explain but may not be statistically significant.

Also for the other variables (in the full report) the same plots invariably suggest that the logarithmic transformation improves the similarity to a normal distribution. Only for salinity (variable SAL) and for dissolved OXYGEN (DOXY) exceptions must be made and in these cases, the original variables without transformation have been retained for further analysis.
It is worth noting that the condition of normality must not be strictly satisfied for the original measurement variables is nonstationarity (ie. variations of the mean value or its standarddeviation) are allowed. In principle it would be better to incorporate the test of normality together with the fitting of such nonstationarities (as is done in the next section). The possibility to do so will be investigated at a later stage.

4 Statistical Characterization of the Variables

In this chapter we explain how the different variables are characterized with respect to 1. their univariate distribution, 2. their spatial correlation, and 3. their correlation with other regression variables. We also indicate how such characterization could be used in a statistical quality control test. Detailed results of the various characterizations are shown here only for the AMON variable. Detailed results for the other variables can be found in the full report and their findings will are summarized each time.

Before discussing the characteristics, the definition of some additional location and time variables that will be used in the following analysis should be explained:

- SEA is a categorical location variable that indicates whether a measurement is taken in the North Sea or in the river Scheldt. (value 0 for measurements in the river Scheldt and 1 for measurements in the North Sea);

- DISTANCE is a continuous location variable that measures the distance (in km) to the station S01, located in the mouth of the river Scheldt. The values are defined to be negative for measurements located along the river Scheldt and correspond in this case to the distance measured along the river banks. For measurements located in the open sea DISTANCE corresponds to the Euclidean distance to station S01, which is approximated by the formula:

$$h = \frac{111.1}{100}\sqrt{(x_2 - x_1)^2 + \left[\cos\left(\frac{x_2 + x_1}{2}\right)(y_2 - y_1)\right]^2} \qquad (1)$$

  where x_1 and x_2 correspond to the latitude coordinates (in degrees) of the two stations while y_1 and y_2 correspond to the longitude coordinates (in degrees);

- SEASON is a categorical time variable which indicates the time of measurement within the year. SEASON equals 1 for measurements between January and March, equals 2 for data collected between April and June, equals 3 for data between July and September and equals 4 for data in the time interval October-December.

4.1 *Univariate characteristics as a function of location and time*

A first way of characterizing the data is to consider the univariate distribution of a measurement. Such a characterization could potentially be used in a quality check by comparing the measurement value with quantile values that are very unlikely to occur (in the lower or upper range of the measurements).

In order to be discriminative, such a test should however recognize that the univariate distribution may vary as a function of location and of time of the year. In this analysis we have therefore considered the automated estimation of the mean of the variable (after transformation if deemed necessary) as a function of the location variables DISTANCE and SEA and as a function of the time variable SEASON. An interaction term SEA*DISTANCE is also allowed. The interaction terms allows to model differences between the relation of the variable to the DISTANCE in the open sea and in the river Scheldt.

Not each of these terms needs however to be of significance and therefore a stepwise regression method is used to select the subset of possible regressor candidates (DISTANCE, SEA, SEASON and DISTANCE*SEA) that leads to the most accurate prediction of the measured variable (i.e. the smallest R^2 goodness-of-fit statistic). To choose the best model the Mallows' C_p statistic rather than the R^2 goodness-of-fit statistic is used. While the R^2 goodness-of-fit statistic is a very useful and direct measure of the goodness-of-fit (it indicates the fraction of the variance that is explained by the model) in its original form it is less suited for the selection of the best model, because the statistic always increases as more terms are added to the model. The Mallows' C_p statistic is closely related to the R^2 goodness-of-fit statistic, but is constructed such that it is known that its expected value should equal the number of parameters in the model for a good fit. The best model is then the model for which this condition is satisfied and which has the smallest number of parameters.

The results of such analysis for the variable AMON are detailed below. Results for the other variables are reproduced in the report.

```
The results of variable AMON

The Standard Deviation of the Data is given by:
sigma = 1.74353957920037

Stepwise Regression Results are given by:
Call:   lm(formula   =   dataset[,   1]   ~   dataset$DIST   +   dataset$SEA   +
dataset$SEASON + dataset$SEA:dataset$DIST, na.action = na.omit)
Residuals:
      Min       1Q   Median       3Q      Max
 -7.17042 -0.59635 0.227161 0.912914 5.40388

Coefficients:
                              Value Std. Error    t value   Pr(>|t|)
             (Intercept) -11.749208   0.178646 -65.768266   0.000000
            dataset$DIST  -0.038826   0.003300 -11.765514   0.000000
             dataset$SEA  -0.044324   0.171977  -0.257733   0.796635
          dataset$SEASON  -0.120450   0.024586  -4.899034   0.000001
dataset$SEA:dataset$DIST   0.028386   0.003405   8.336347   0.000000

Residual standard error: 1.48271 on 2418 degrees of freedom
Multiple R-Squared: 0.283898
F-statistic: 239.654 on 4 and 2418 degrees of freedom, the p-value is 0
```

The previous results show that for AMON the stepwise regression method retains all terms. Details on the significance of the estimated coefficients for the different terms show on the other hand that the estimate for SEA is not significant (the t-value is only -0.26 and there is nearly 80% chance that such a t-value would be produced by random error in a model without the SEA term). Nevertheless the SEA term is retained in the stepwise regression method, because the interaction term SEA*DISTANCE is found to be highly significant (the t-value equals 8.3 and it is highly unlikely that such value would be produced by random error). It should be noted that the requirement that the term SEA is retained when the interaction term is significant, is specific to the algorithm used and only applies to categorical variables (such as the SEA variable). It is possible that the stepwise regression method finds the interaction term SEA*DISTANCE to be significant and yet excludes the DISTANCE term.

The table below summarizes the results of the analysis for the other measurement variables. Details of the analysis are found in the report.

Table 5: Results of univariate regression

Response	**Regressor**			
	SEA	**SEASON**	**DISTANCE**	**Interaction**
AMON	X*	X	X	X
CD	X	X	X	X
CDP			X	
CDPS	X*	X	X	X
CPHLL		X	X	
CPHLH	X*	X	X**	X
CPHLC	X*	X	X	X
CU	X	X		
CUP	X*		X	X**
CUPS			X	
DOXY	X*	X*	X	X
HG			X	
HGPL		X	X	
NTRA	X	X	X	X
NTRI	X		X	X
NTRZ	X	X	X	X
PB	X	X		
PBP			X	
PBPS			X	
PHATL	X		X	X
PHOS	X	X	X	X
SAL	X		X	X
SLCA	X*	X	X	X
SUSP	X	X	X	
ZN	X	X	X	X
ZNP			X	
ZNPS			X	

Note: In the table given above, an * is used if the estimate of the parameter is non-significant at a 5% level, but retained as mentioned above. A ** is used for significance of the parameter estimate between 5% en 10 %.

It is clear from the above analysis that the DISTANCE variable is in most cases significant. Also the effect of SEA, SEASON and the interaction term is often found to be significant.

As for the case of the AMON variable, the results of the method are not always evident to interpret. For instance, for the DOXY variable, the offset of SEASON is retained the stepwise regression, whereas the estimate itself is found to be not significant. Such apparent inconsistencies are due to the fact that in the stepwise regression method uses the Mallows' C_p statistic to select the best model without conditions on the significance of the various terms. In this respect, the method should be improved.

4.2 Characterization of spatial correlation

Geographically distributed data typically show some form of spatial continuity and hence some correlation between neighboring values. Such dependency can be used to advantage in the quality control by verifying one measurement value against another measured at a nearby location.

In order to perform such a test, it must be known how the difference between two measurements varies as a function of the distance between the two measurements. This dependence is expressed through the variogram that models how the expected value of the squared difference of two measurements varies with the distance h between the two measurements. The semi-variogram corresponds to 1/2 of this value and can be estimated on the basis of data as follows:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{|s_i - s_j| = h} (y_i - y_j)^2 \qquad (2)$$

where

γ(h) is the estimated semi-variance at distance h,
h is the lag (that is the distance between sample locations),
y_I is the value of variable y at location s_I,
N(h) is the number of pairs of observed data points.

The previous estimation formula only works well if a sufficient number of data is available for each distance h considered. In the present data set this is not the case and therefore a continuous variation with the distance DIST is assumed as follows:

$$\gamma = a + b^*\mathrm{DIST}^c. \qquad (3)$$

A power model of this type allows for the fact that the data would show a (unknown) trend. For small values of c, the previous model is however indeterminate (i.e. the same model is obtained if a is increased while b is decreased by the same amount). In such a case, an alternative linear model is used that does not pose this problem. The model is in this case of the following simplified form:

$$\gamma = a + b^*\mathrm{DIST} \qquad (4)$$

When in this last model, b is estimated to be negative, the semi-variogram is assumed to be simply constant as follows:

$$\gamma = a \qquad (5)$$

This last model corresponds to a case where there is no spatial correlation between neighboring stations. In practice this means that the measurement value cannot be checked against another measurement, even if such measurement is taken at a short distance.

In this study, different variograms are estimated for observations in the river Scheldt and observations in the North Sea. Also the estimation method differs between these two cases as is explained next.

For observations in the river Scheldt the distances are measured along the river banks: it is natural to believe that differences between two measurements will increase as a function of this 1-D spatial coordinate. In the case of the open sea, no such preferential direction is immediately evident and, in this study the distance between two stations is defined to correspond to the Euclidean distance between the two stations which is approximated as follows:

$$h = \frac{111.1}{100}\sqrt{(x_2 - x_1)^2 + \left[\cos\left(\frac{x_2 + x_1}{2}\right)(y_2 - y_1)\right]^2} \qquad (6)$$

where

h is the Euclidean distance,
x_i is the latitude of the location i,
y_i is the longitude of the location i.

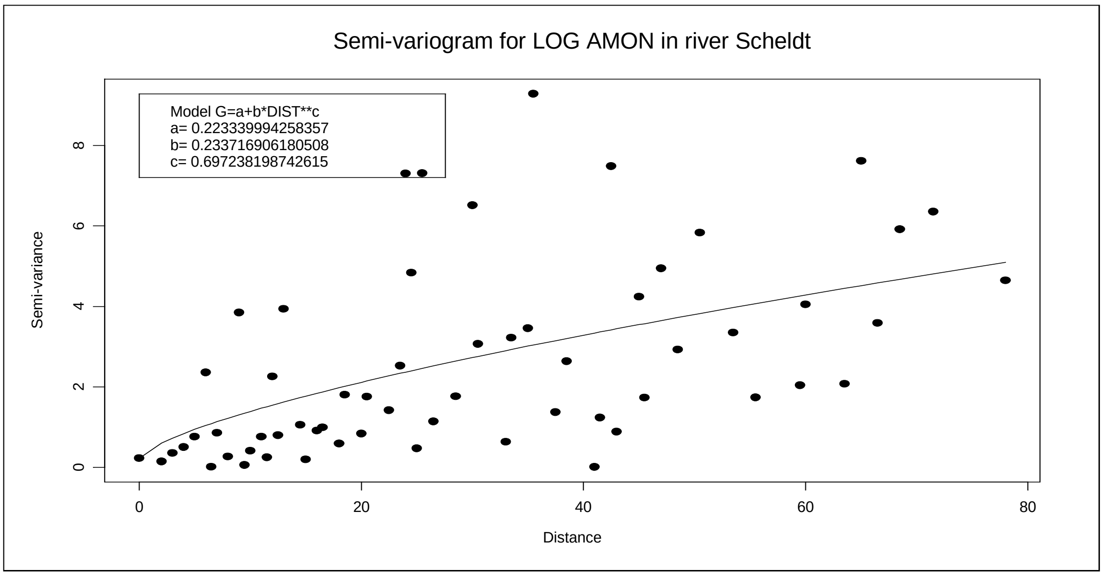
Later on in the analysis some evidence was found that perhaps better fits could be obtained by considering the distance to the coast (or the distance to the nearest river) as a preferential direction. Within the context of this study, this alternative modeling option has not been further pursued, because of the difficulty in calculating such distances without the support of a GIS-system.

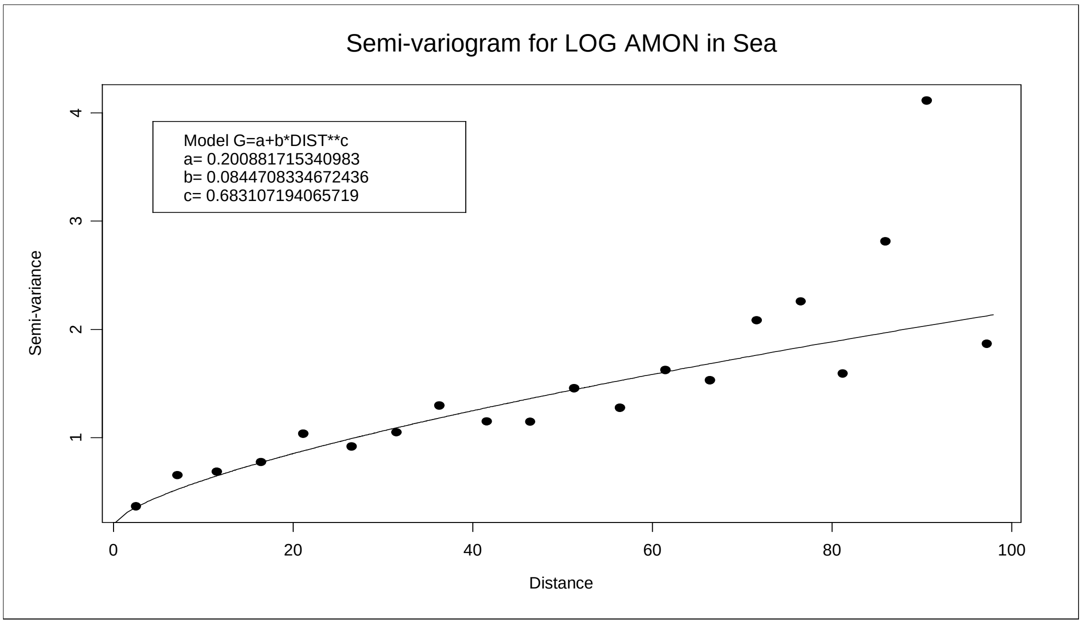
To estimate the semi-variogram, first of all Equation 2 is applied to obtain values of $\gamma(h)$ at those distances h that are found for different pairs of stations. Note that because the measurements are made at fixed stations, the same distance lag h will typically be repeatedly found if one considers the data set over a longer time period. Differences y_i-y_j in Equation 2 are used only if those measurements are taken within a timespan of 5 days. This constraint is added, to exclude the comparison of measurements that are affected by seasonal changes.

For the river Scheldt, the semi-variogram model of Equation 3 is then fitted to the estimates of $\gamma(h)$ for the observed distance lags h. In the case of the North Sea, a very large number of distance lags h may however result and therefore it is has been decided to apply further grouping of the different lags. This grouping will have little effect on the final estimates (in fact the estimate of the model is based on the original data without such grouping applied), but the averaging operation removes some of the random scatter about the model so that the variation of $\gamma(h)$ with distance h is visually easier to interpret. Specifically, the grouping intervals have been chosen such that the total distance span is divided into 20 groups with an (approximately) equal number of observations.

To fit the model of Equation 3 a non-linear least-squares method is used. Thus, the estimated coefficients are those that minimize the sum of the squared residuals (the difference between the model fit and the nonparametric estimate of $\gamma(h)$). The decision to replace the fit with the simpler models of Equation 4 or 5 has been made in this study on a case by case basis. Automatic decision rules, if necessary, will be considered at a later stage.

The results of this analysis are shown below for the variable AMON. Results for the other variables are shown in the full report.





For the AMON variable the semi-variogram fit shows both for observations in open sea and for the river Scheldt a clear upward trend as function of the distance. However, the variogram is clearly higher in the river Scheldt indicating that in this case more variation between the measurements is to be expected for the same distance measure. Visually, the fit appears to be in both cases a satisfactory representation of the trend observed in the original data (scatter of the nonparametric estimates $\gamma(h)$ about the fitted model is to be expected and can be attributed to random variation; that this scatter would be larger for the Scheldt river is also logical because the estimates $\gamma(h)$ are in this case based on fewer observations).

The following table shows the estimates of the coefficients a, b and c for each of the variables. Detailed results and graphs are shown in the report.

Table 6: Results of modeling the semi-variogram

Response	**Parameter estimates in model $\gamma = a + b*DIST^c$.**					
	River Scheldt			**Sea**		
	a	**b**	**c**	**a**	**b**	**c**
AMON	0.22	0.23	0.70	0.20	0.08	0.68
CD	0.22	0.03	0.71	0.72	4E-4	
CDP	2.42	0.06		1.10	8E-3	
CDPS	0.47	0.10	0.71	1.11	8E-3	
CPHLL	0.56	7E-3	1.22	0.27	0.05	0.46
CPHLH	0.07	7E-3		0.42	8E-8	3.80
CPHLC	0.22	0.02	1.08	0.16	0.04	0.57
CU	0.23	7E-10	4.74	0.65		
CUP	0.51	1E-4	2.46	0.57	5E-3	
CUPS	0.25	5E-4	2.03	0.30	1E-3	
DOXY	0.85	0.02	1.40	0.71	5E-3	
HG	0.25	1E-3	1.24	0.56		
HGPL	0.14	2E-3	1.41	0.90	8E-3	
NTRA	0.43	2E-5	2.62	0.70	0.02	
NTRI	0.02	0.12	0.71	0.13	9E-3	
NTRZ	0.08	9E-3	1.16	0.33	0.12	0.54
PB	0.05	8E-3	0.95	0.60		
PBP	0.52	7E-3		0.91	6E-3	
PBPS	0.25	1.12	4.13	0.44		
PHATL	0.23	6E-3	1.27	0.85	0.01	
PHOS	0.13	1E-5	2.71	0.33	3E-4	1.79
SAL	1.94	0.53	1.51	0.43	0.45	0.47
SLCA	0.07	0.02	1.28	0.37	1E-4	1.95
SUSP	0.12	6E-3	0.98	0.28	0.08	0.48
ZN	0.13	6E-5	2.24	0.66		
ZNP	0.49	0.04		1.37	2E-3	
ZNPS	0.69	7E-3		0.73		

In general one finds that for practically all cases, the variograms estimated in the river Scheldt show an upward trend (either the power model or the linear model is found to lead to the best fit). For the North Sea on the other hand, the upward trend is found to be less significant and, in particular for metals, it is found that a constant semi-variogram will typically lead to the best fit. As noted earlier, in this last case, spatial correlation would be of little use in a quality control check. In the other cases however, the variogram often shows a very steep increase and, if measurements exist at a small distance, it is clear that these measurements could be extremely useful to check the validity of another value.

Note also that in a few cases the estimate of the power coefficient c exceeds the value of 2. In theory such values can not occur for stationary, isotropic processes. Evidently this problem can be always avoided by fitting the variogram over a shorter distance range (if the range is sufficiently small a linear fit will be for all practical purposes sufficient). We plan to address this issue and more in general the estimation of the variogram in more depth at some later stage.

4.3 Correlation with other variables

One would expect that the different measurements, if taken at the same location and time, are not entirely uncorrelated. This is evidently true for variables that are by definition related (i.e. NTRZ=NTRI+NTRA), but may also apply to other variables that are indirectly related. Correlation of this type, if it exists, is obviously of interest in a quality control, since it would allow to predict the measured value (with a given accuracy) on the basis of other measurements. In this chapter, we examine to which extent such correlations exist and we estimate their form for the different variables.

A-priori one may expect that a measurement variable could be related to each of the other variables or a combination of such variables. In this study, we have limited the analysis to consider only relationships of the measurement variable of interest (the response variable) to 1, 2 or 3 of the other measurement variables (the predictor variables). The possible relationships that can be fitted are further constrained by the fact that a sufficient number of simultaneous readings of the potential subset of measurement variables should be available. In this study, "a sufficient number" has been defined to correspond to approximately 10% of the total number of observations of the response variables. Typically, the number of records is either close to 2500 or 1000 records, so that the estimates are based on at least either 250 or 100 records.

For each measurement variable and for each subset of predictor variables that satisfies the previous condition, a linear regression model is fitted using least-squares. The significance of the estimated regression coefficients is then considered (using a sequential t-test at a 5% significance level) and if one (or more) of the terms is not significant the model is rejected (the simpler model with the non-significant term removed will be tested separately in one of the other combinations). The models that thus remain are then ordered according to decreasing adjusted R^2 goodness-of-fit statistic. This statistic indicates the variance that is explained by the model (a high value corresponds to a good fit), but adjusts this estimate for the number of parameters used to explain that variance (i.e. if the same variance is explained by a model with 1 measurement variable and another model using 3 measurement variables, the first model will be ranked higher according to the adjusted R^2 statistic).

The results thus obtained for the AMON variable are listed below. Results for the other measurement variables are shown in the report.

Table 7: Results of the analysis for AMON

R^2adj	Reg1	Reg3	Reg4
0.7337135	CPHLL	NTRA	SLCA
0.7127315	NTRA	PHATL	SLCA
0.5760580	NTRA	NTRI	SLCA
0.5740157	NTRA	SLCA	SAL
0.5688520	NTRA	SLCA	0
0.5540013	NTRI	NTRZ	SLCA
0.5399463	NTRZ	PHATL	SAL
0.4961977	CPHLL	SLCA	SAL
0.4850517	PHATL	SLCA	SUSP
0.4812561	NTRZ	SUSP	SAL

In the case of the AMON variable the ‘best’ model uses the predictor variables CPHLL, NTRA and SLCA. The adjusted R^2 statistic corresponds to 73%. An alternative model of nearly equal performance uses the predictor variables NTRA, PHATL en SLCA.
In a quality control check, one might maintain for each measurement variables such ranked lists of possible regression models and then use the highest ranked model that can be applied (in view of the availability of other measurements).

The best model that is obtained for the other measurement variables are summarized in Table 5. A list of alternative models for each measurement variable is shown in the full report.

Table 8: Best model obtained from regression analysis

Response	R^2adj	Regressors		
AMON	0.73	CPHLL	NTRA	SLCA
CD	0.54	CU	DOXY	SUSP
CDP	0.77	CDPS	CU	CUP
CDPS	0.75	CDP	CU	ZNPS
CPHLL	0.98	CPHLC	CU	PHATL
CPHLH	0.88	CPHLC	CPHLL	DEPH
CPHLC	0.99	CPHLL	CU	PHATL
CU	0.34	AMON	CD	CPHLC
CUP	0.72	CDPS	ZNP	SAL
CUPS	0.82	PBP	PBPS	ZN
DOXY	0.57	CPHLC	TEMP	SAL
HG	0.42	AMON	CPHLL	PHATL
HGPL	0.47	NTRZ	SUSP	SAL
NTRA	0.997	DEPH	NTRI	NTRZ
NTRI	0.62	NTRA	NTRZ	PHATL
NTRZ	0.997	DEPH	NTRA	NTRI
PB	0.49	CD	CPHLC	ZNP
PBP	0.74	CUP	NTRZ	SUSP
PBPS	0.85	CUPS	PBP	ZNPS
PHATL	0.86	CPHLC	CPHLL	SAL
PHOS	0.66	DOXY	SLCA	SAL
SAL	0.87	DOXY	NTRZ	PHATL
SUSP	0.87	AMON	CPHLL	NTRA
ZN	0.51	DOXY	PHATL	TEMP
ZNP	0.49	PHATL	ZNP	ZNPS
ZNPS	0.70	SUSP	ZNPS	SAL

The previous results show that in most cases very high values of the adjusted R^2 statistic are obtained. Only for overall metal concentrations (CD, CU, HG, PB, ZN), the goodness-of-fit statistic is relatively low.

The results for NTRA, NTRI en NTRZ are of particular interest since in this case the relationship is theoretically known: NTRZ=NTRI+NTRA. The table above shows that indeed a model is selected that includes the correct predictor variables but in addition another variable (for NTRA en NTRZ depth DEPH, for NTRI phaeophytin-a PHATL) is included. The detailed results show that the models without these variables are ranked as second best and have an adjusted R2-statistic that is very close to the optimal. On the other hand, the additional variable included is found to be statistically significant (albeit of little practical importance). Whether these variables may in fact have a physical influence (i.e. depth may indirectly have an influence through the measurement procedure?) remains to be seen. More in general, it stresses the fact that when using such lists in a quality control, the composition of such a list should not be entirely automatic, but should be at least verified by an expert user that is knowledgeable about the physo-chemical background of the relationships.

5 SUMMARY AND CONCLUSIONS

In this study, the contaminant data gathered from 1977 until 1996 of the Belgian monitoring program have been analyzed within the perspective of using statistical characteristics for quality control testing. In order not to confound the conclusions of this study with problems associated with a too small data size, the analysis has been limited to consider those measurement variables for which ample data have been measured.

Three statistical characteristics have been examined: 1. the univariate distribution, 2. the spatial correlation, and 3. the correlation with other variables. The nature of each of these characteristics, their possible use in a quality control check and problems still to be resolved are summarized next.

The univariate distribution could be used to define lower- and upper-bound values that are expected to occur very rarely and against which the measurement values could be compared. The analysis shows that it is important to first of all transform the variable of interest in order to satisfy the assumption of normality and that after transformation the mean value of the univariate distribution is for most of the variables a function of location (in particular for the case of the river Scheldt). Differentiation between open sea and the Scheldt river and differentiation with respect to seasons is also found to be important in several cases. A procedure to automatically select the appropriate model has been devised and could be used to automatically extract this relation from the database (then to be used in the quality control program). Not investigated in this study is whether also the standard deviation would vary with the same variables. We propose that such a study is postponed after the present algorithm has been incorporated into the SAP program, after which it should be easy to apply the same algorithm in this case to the squared residuals of the model to obtain thus a model for the standard deviation. Furthermore it is found that for practically all of the variables the lower-bound value is physically equal to 0 and, in practice, corresponds to the detection limit of the instrument or method used. Thus, for practical applications, the quality control check will be typically based on the comparison against a statistically derived upper-bound value, while the lower-bound value would be physically based.

The present study shows that for most of the measurement variables there is indeed some spatial correlation (based on the transformed variables). In this study it is shown how this spatial correlation could be quantified through the variogram and how the variogram value could be estimated from the data. Again, it is planned that this procedure will be part of the SAP program so that estimates can be automatically obtained. The distinction between open sea and the river Scheldt is found to be significant. It is also found that for metal observations in open sea, spatial correlation is small or non-existent and in such a case, a quality control based on the variogram value will be of little use. The present study has been limited to a particular model for the variogram (a power-law model) which is found to lead in most of the cases to a reasonable fit. Other models may be eventually considered in a final implementation in the SAP program.

Finally, the study has considered the possibility to predict a measurement value on the basis of a linear combination of other simultaneous measurements (no transformation is applied in this case). It is shown that one can construct a ranked list of possible predictions that depend on the availability of the other measurements and that, for most of the variables, the top-ranked combinations (the best fits) have a very high R^2 goodness-of-fit statistic. In practice this would mean that a quite accurate independent measurement value can be estimated on the basis of the other measurements and thus a reliable quality control can be performed. The study shows that this does not apply for metals where correlation with other variables is found to be relatively small. An issue that has not been addressed in this study is the type of model: in all cases, a linear relationship between the measurement variables has been assumed. Again, this may be an issue to be considered in further depth in the final implementation of the algorithm in the SAP program. In this study, the maximum number of regressor variables has been limited to 3 and the best model is always found to use this maximum number. For this reason, one might also consider increasing this number, even though in practice it will become of course less likely that all such measurements are made at the same time.

Perhaps the most important overall conclusion of this study is that it will be extremely difficult and potentially dangerous to develop fully automated algorithms that result in setting the coefficients for a quality control test. In each and every case, some assumptions are fixed (normality, or linearity, or dependence only on certain variables) and it is only through visual examination of the data and consideration of the results that the validity of these assumptions can be (to some extent) verified. Practically speaking, this means that we advise that each quality control check that is used in practice is derived with the help of the SAP program, but that transfer of the results of this program to the application of a specific quality test requires the intervention of the data base manager. Such a human interaction appears to us essential to safeguard the system against happenstance relationships or the overlooking of quite strong relationships that are not detected within the (neccessarily) fixed framework of any model estimation.

Integrated and Dynamical Oceanographic Data Management

IDOD

IDOD Database User's Guide

Scientific report

Fabrice MULLER

MN/DD/61

December 1998

SURFACES

Service Universitaire de Recherches Fondamentales et Appliquées
en Cartographie et en Études Spatiales

TABLE OF CONTENTS

IDOD Database

User's Guide

1. Introduction

The design of a GIS database requires a methodological approach to the problem. The most common scheme used to manage hybrid data sets is based on a couple of databases, one for non-spatial alphanumerical data, and another reserved for spatial or geographical data. These two data sets constitute the data-side of the GIS.

The design of a powerful GIS database must be elaborated on the basis of a reliable methodology. The MERISE method is one of the most sophisticated technique for the development of such an IS database [NANCI & ESPINASSE, 1996]. Therefore, in the framework of the IDOD project, it was decided to follow the construction line suggested by the MERISE method for the elaboration of the non-spatial alphanumerical database.

The MERISE method uses the Entity/Relationship formalism.

At the present time, a prototype of the IDOD database has been realised and is operational. It is mainly devoted to the parameters and measured values concerning the quality of sea water.

This prototype of the database was designed by the Laboratory SURFACES of the University of Liège in strong collaboration with the Brussels MUMM team.

2. The conception of a database.

2.1. Definition of a database

A database is a collection of data describing a specific subject. These data are recorded on storage devices and can be accessed and manipulated through a computer. This collection has to verify the following properties:

- To be organised as a structured set of data:

The storage of the data is organised to optimise the data access. For example, in the IDOD database, the measured values are more often used than the information relative to projects. For this reason, the two files will not be organised similarly.

- To constitute an integrated data set:

All the files of the database are in the same system. It is required for easier transactions and links between all the different types of information inside the database.

- To represent the reality:

It is the most important rule of a database. The stored data have to represent the exact reality. To force the data to respect the reality, we define the integrity constraints in the database. These constraints are the operating rules of the database.

- To contain operational data about a specific subject:

The information inside the database is organised with entities and relations or associations. In the framework of the IDOD project, the design is elaborated using the Entity/Relationship formalism. This formalism will be described further in this document.

- To be usable by several users and/or applications simultaneously:

A database is a tool that has to be shared by many users simultaneously. For example, a user can access the database to introduce or modify data, while another user is consulting existing data. It is required that the database system allows more than one user, each user running a specific application for consulting or updating the database. The system has also to assure that only one user has write access to a record of data at a time and that other users have only read access to this record during that time.

- To avoid redundancy:

The non-redundancy implies that each datum has a single occurrence into the database. It is required to insure the coherence of the database. If the datum is single, it can be correct or incorrect but cannot be contradictory. If it is repeated, there can be some discordance between the different occurrences due to the update of one without the same process applied to the other ones.

A database can be viewed as a self-describing collection of integrated records. It is self-describing in that it contains, in addition to application data, a description of its own structure. This description is called the *data dictionary* or *meta-data*, meaning data about data. The data dictionary allows program/data independence.

For example, the database of a library is a self-describing collection of books. In addition to books, the library contains a card catalogue describing the books in the library. In the same way, the data dictionary, which is part of the database just as the card catalogue is part of the library, contains descriptions of the data contained in the database.

Why is this self-describing characteristic of a database so important ? First, the structure and contents of the database can be determined by examining the database itself. Second, all changes to the structure of the data in the database (such as adding a new data item to an existing record) is resumed to a simple change in the data dictionary. Few, if any, programs will need to be changed.

2.2. Database Management System (DBMS)

The Database Management System (DBMS) is a program or collection of programs used to establish, maintain, and process a database. The DBMS offers to the database administrator some functionality such as to create the database, to define its physical parameters and the objects it contains. A database is designed at different levels, as it will be explained further.

In a database it is important to have a complete independence of the manipulated data and the application programs that manipulate them. The structure of the database can be modified without having to rewrite all the programs that access to the database. The structure of a database is always growing in complexity with the adjunction of new entities or relationships, the splitting of some entities into sub-entities, etc. A database has to evolve physically to optimise the access times of the different applications: it is the *tuning* of the database. All these reasons show why the independence of data and applications is recommended.

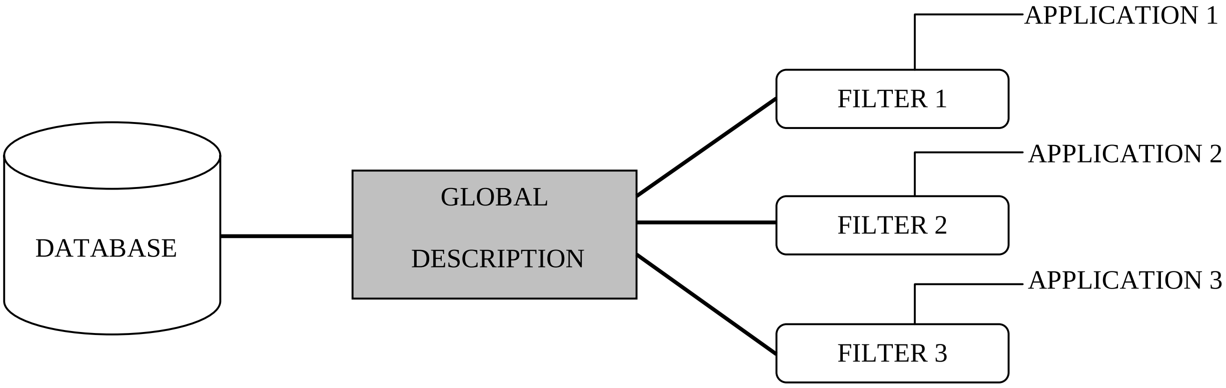


Figure 1. Structure of a database**.**

2.3. Architecture of a database

A database is represented by three main different levels that describe the connections between the real world and the physical implementation of the database into a computer. The architecture can be divided into the following levels: internal, conceptual and external levels. This scheme corresponds to the figure 2.

- **The external level.**

This level represents a part of the information system. In a relational database, it is a subset of the conceptual design.

- **The conceptual level.**

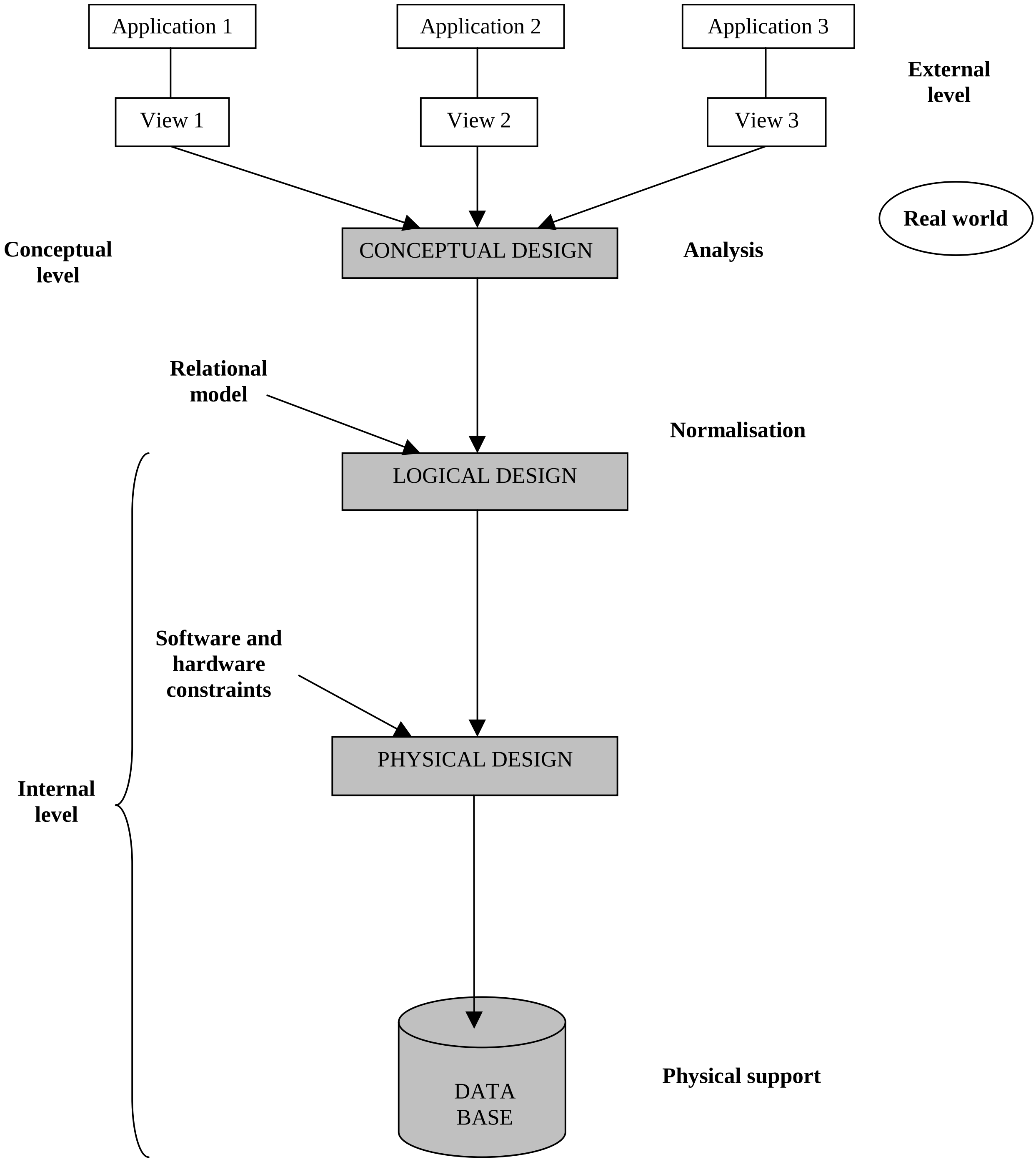


Figure 2. The representation levels of a database.

The conceptual schema is derived from an analysis method used to produce a model of the real world. At this level, there is no technical constraints or access time optimisation.

- **The internal level.**

This level requires the choice of a type of DBMS: relational, network or files database. In the case of the IDOD database, a relational DBMS has been chosen and the model used to elaborate the logical design is the relational model. This level describes the implementation of the information system and assures the independence between the software and the hardware, this level is subdivided into the logical and the physical levels.

The physical level determines the storage media and formats on basis of several parameters such as: the volume size of the data, the number and frequency of transactions, the number of users, etc.

The design of a database can be resume as in the figure 3.

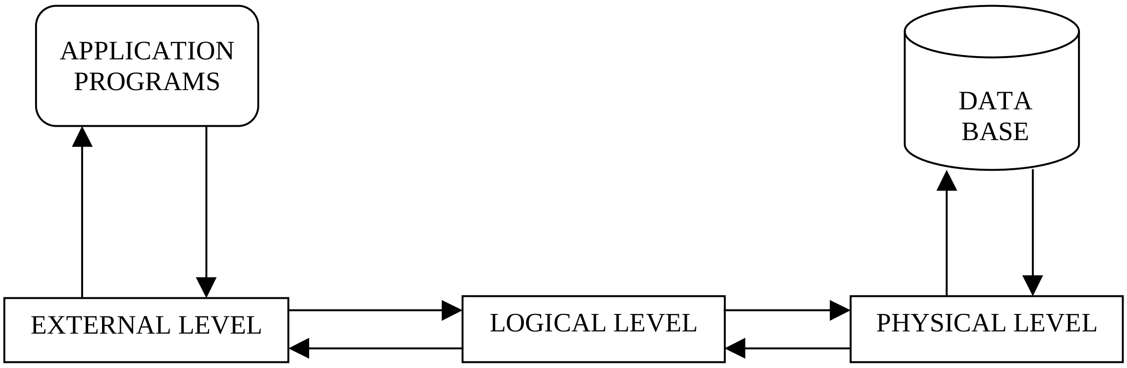


Figure 3. Access to a database through an application program.

2.4. The DBMS relational model.

As explained previously, there are more than one model for the implementation of a DBMS. The most often used are the relational and the object oriented models. For the IDOD project, the implementation of the database is realised using the relational model for the DBMS. This model defines the structure of the data and the operations on these data with the relational algebra. A relational database is constructed only with *relations* and *domains*. There is no pointer because this kind of object is not part of the relational model. The integrity constraints assure the respect of the model. The requests and operations on the data are made by use of predefined operators. The main operations are: searching, adding, updating and deletion.

2.5. The Entity/Relationship formalism for the conceptual datamodel.

2.5.1. Objectives.

The conceptual datamodel of the IDOD database is realised using the Entity/Relationship formalism (E/R formalism). As previously said, the conceptual datamodel is a static representation of the information system (IS) or all the data of the domain. It is designed with abstraction of the technical aspects of the database implementation and access techniques. It refers only to abstract objects and associations between these objects. The conceptual datamodel describes these objects and associations. There is no mention of what will be done with the data; only the semantic of the system is considered.

There are two ways to elaborate a conceptual datamodel:

1. A deductive approach that supposes the existence of a list of existing data to organise;
2. An inductive approach that aims to highlight all the concepts and ideas.

2.5.2. *Presentation of the E/R formalism.*

The Entity/Relationship formalism is a high-level semantic design tool that is set up on three fundamental constructive concepts: the *attributes* or *properties*, their gathering into *entities* and the links or *associations* between these entities.

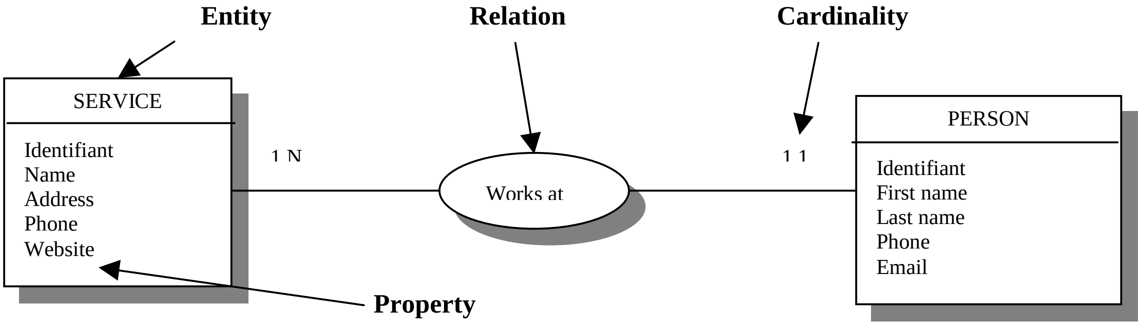


Figure 4. E/R formalism design.

2.5.3. *Attribute or property.*

The *attribute* or *property* is the representation of an elementary or atmomic information. It is the smallest part of information used by the system and having a meaning. The attribute can be single (first name, phone number, identifiant code, etc.) or compouned (address, date, etc.).

The attribute describes the entity or the relation and exists only if it is part of an entity or a relation. An attribute is unique in a conceptual datamodel and is part of only one entity or relation.

2.5.3. *Entity.*

The entity modelises a set of objetcs of the same type that are abstract or concrete. The entity is a group of attributes and represents only one semantic concept. The entity determines a type, a class or a group where all the elements are called the entity occurrences.

To refer directly to one occurrence of an entity, the entity requires a special attribute called an *identifiant* or *primary key*. One value of this identifiant corresponds to one and only one occurrence of the entity and the value of this identifiant must be invariant and can never be changed or modified until the deletation of the occurrence of the entity.

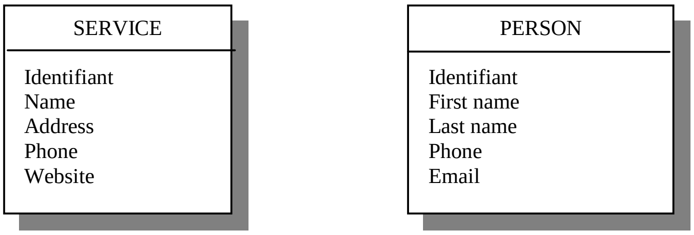


Figure 5. Examples of *Entities*.

2.5.4. Relation.

The relation is a group of associations of same type between two or more occurrences of entities of same or different types. The relation translates the words of the natural language. For example, in the figure 4, the relation "*works at*" between the entities SERVICE and PERSON is the translation of the sentence *"A person X works at service Y"*.

In opposition to the entity, a relation has no real existence. The relation is simply expressed by the implied entities. The minimum number of entities implied into a relation is at least two. The dimension of a relation is equal to the number of implied entities.

A relation can include some attributes like entity.

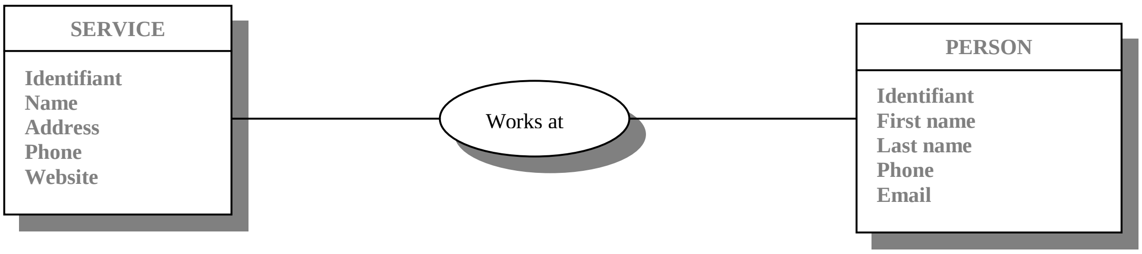


Figure 6. Exemple of Relation.

2.5.5. Cardinality.

For each pair of entity-relation, the cardinalities are the minimum and maximum numbers of occurrences of the relation that can exist for one occurrence of the entity.

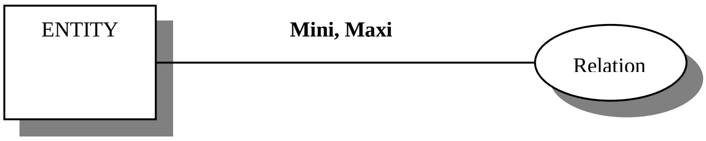


Figure 7. Example of Cardinalities.

The cardinality values are indicated upper or on the link between the entity and the relation. The most common values for cardinality are:

- **0,1**: one occurrence of the entity can exist without participating to the relation (0), and if it participates it is only on time (1);
- **0,N**: one occurrence of the entity can exist without participating to the relation (0), or can participate without limitation (N);
- **1,1**: one occurrence of the entity participates one and only one time to the relation;
- **1,N**: one occurrence of the entity participates at least one time to the relation.

2.6. Presentation and examples of the *DB-Main* case-tool.

The drawing of the conceptual datamodel of the IDOD database is made using the software *DB-Main 4.0*, downloaded in its demo version at the Web site address "*http://www.info.fundpc.ac.be/~dbm/*". It is a generic case tool for the database conception and was developed by the *Institut d'Informatique* of the *Facultés Universitaires Notre-Dame de la Paix* in Namur. It offers the possibilities to elaborate a conceptual datamodel with entities, relations, cardinalities, etc.

2.6.1. Entity.

In *DB-Main*, like in the E/R formalism, the entities are represented as concrete or abstract objects of the real world. Graphically, the entities have the same representation as in the E/R formalism.

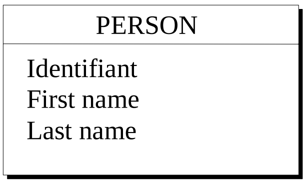


Figure 8. Representation of an entity in *DB-Main*.

2.6.2. Attribute.

The attributes are represented as illustrated in figure 8. Types can be affected to attributes but, in this case, *DB-Main* is just used as a drawing tool, and consequently this information is not showed.

2.6.3. Relation.

In *DB-Main*, the relation is represented by an irregular hexagon.

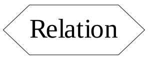


Figure 9. Representation of a relation in DB-Main.

A relation can also have attributes as illustrated in figure 10.

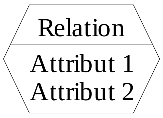


Figure 10. Representation of a relation with attributes in DB-Main.

2.6.4. Cardinalities.

As in the E/R formalism, the cardinalities are indicated on the links between the relation and associated entities. *DB-Main* allows to select between the set of cardinality values: (0,1), (0,N), (1,1) and (1,N); or to define specific values with respect of the rule *Mini<Maxi*.

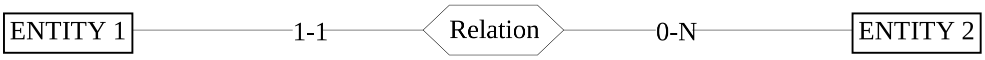


Figure 11. Representation of the cardinalities in *DB-Main*.

2.7. Normalisation

Normalisation is the process of gathering data items or properties into relations. Object diagrams makes the normalisation process easy because they represent groups of related properties. The goal of logical database design is to represent objects in the database using relations that provide the data needed to construct user objects and that are robust enough to allow items to be inserted, deleted, and modified without resulting in inconsistencies or errors in the stored data. The theory of normalisation has been introduced by E.F. Codd who defined a variety of *normal forms*. To understand the normalisation process, it is fundamental to know what are functional *dependency* and *key*. A functional dependency is a relationship between attributes. A key is a group of one or more attributes that uniquely identifies an occurrence of an object or entity. More details about this theory are out of the scope of this report and interested readers are invited to consult the specific literature mentioned in the bibliography.

The are five normal forms:

- the first normal form;
- the second normal form;
- the third normal form;
- the Boyce-Codd normal form or fourth normal form;
- the fifth normal form.

Generally, in practice, for a relational database, only the three first normal forms are applied.

The normal forms are structured in the order listed above. A relation in the second normal form is automatically in the first normal form and so on as illustrated in the figure 12.

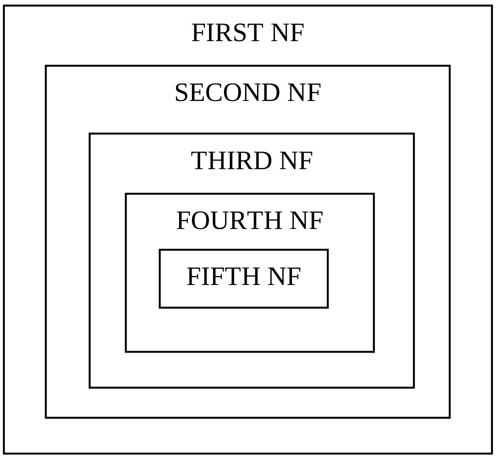


Figure 12. The normal form hierarchy.

- ***The first normal form***

A relation satisfies the first normal form if all properties are atomic value.

- ***The second normal form***

A relation satisfies the second normal form if and only if:
1. It is in first normal form;
2. Each attribute non-participating to a primary key is not in dependency of only a part of the primary key.

- ***The third normal form***

A relation satisfies the third normal form if and only if:
1. It is in second normal form;
2. Each attribute not included in a key is not depending of a non-key attribute.

3. The IDOD database.

The conception of the prototype of the IDOD database is the result of a long process of meta-data or inventory of data procedures. Several meetings with the different MUMM's teams led to the sketch of a first design that was modified many times before to obtain the conceptual design used for the prototype realisation with *Access97*. The main steps of the conception can be resume as:

1. Inventory of data ;
2. Definition of a structured form for every data with associated attributes ;
3. Research of general structures for extracting a global structure ;
4. Presentation of the model to the data providers for remarks and suggestions via the *Newsletter* ;
5. Validation of the conceptual datamodel ;
6. Normalisation ;
7. Conversion of the conceptual datamodel to the relational or logical datamodel ;
8. Implementation of the physical database with *Access97* ;
9. Introduction of data into the database by the use of specific realised requests ;
10. Realisation of forms for manual data input ;
11. Realisation of interrogation requests ;
12. Presentation layout for data output.

The design of the conceptual and logical models realised with the graphical tool of DB-Main are presented in annexes. There are two designs for each model: one for the continuous data values and a second one for the non-continuous data values. Both are very similar, except on some entities like SAMPLE, SUBSAMPLE, etc. They are presented in two designs only for clearer presentation and legibility.

At the present time, the prototype of the database is only devoted to the parameter concerning the quality of seawater. Two categories are considered: continuous and non-continuous data values. The continuous values are obtained by result of automated measurements at regular time intervals; while non-continuous values are not part of temporal series.

3.1. The conceptual design

After the normalisation process according to the three first normal forms, the IDOD conceptual design contains **23 entities** and **34 relations**. The conceptual design is strongly linked to the data dictionary that describes with more details all the attributes of the entities. This data dictionary is presented as an annex to this report. All the attributes mentioned as *Required* are mandatory and need to insert an item of this data. For example, when introducing a new service, it is required to know to which institute, city, country and at least one person this service is linked. These constraints are integrity constraints of the database; they are fundamental for the coherence of the information system.

More details about the required adaptations are mentioned in the Annex A.

Further in this report, some relation between entities will be described with associated transformations to the logical and physical designs.

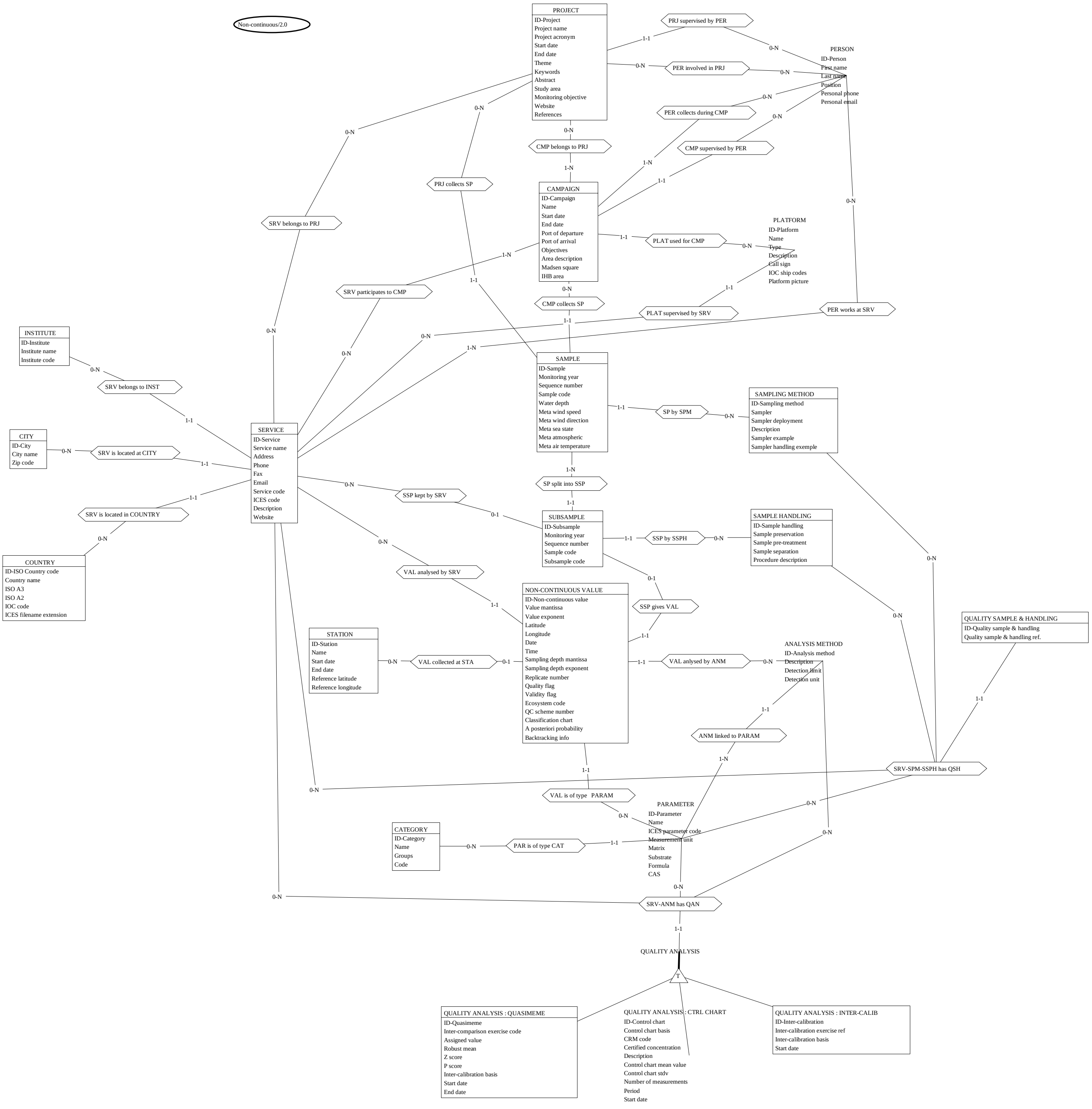


Figure 13. The conceptual datamodel for non-continuous values.

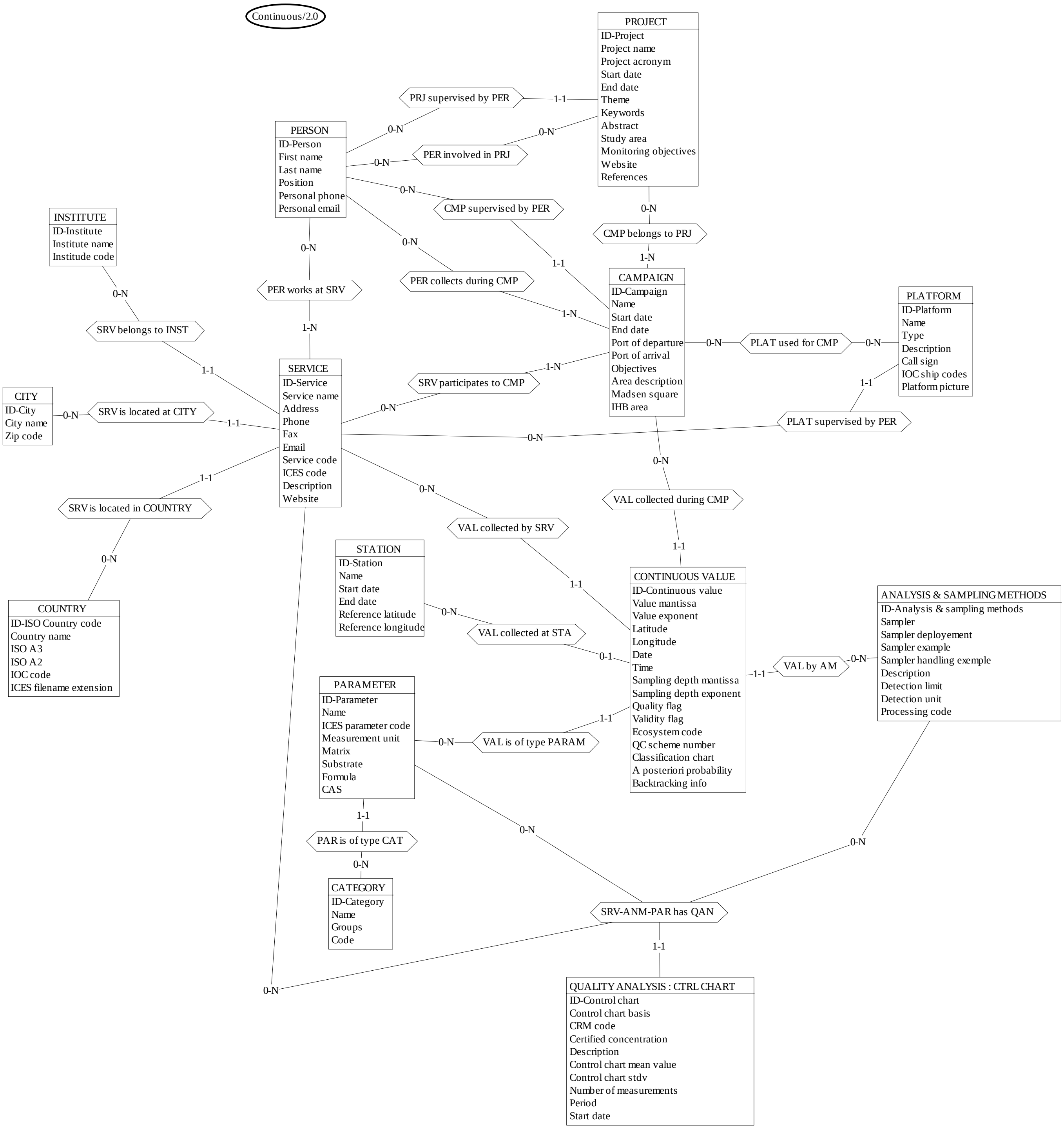


Figure 14. The conceptual datamodel for continuous values.

The conceptual datamodels illustrated in figures 13 and 14 are annexed to this report in a legibility form.

3.2. The logical design

In the logical design some attributes have to be added to materialise the links between the different tables. The use of such foreign keys depends on the type of relations between tables.

There are three types of relations to be considered:

1. One to one;
2. One to many;
3. Many to many.

The first and second relations are converted into the logical model by the use of new foreign keys into the tables. The third one is more complex to implement and requires the creation of an additional table for translation or relationship. Some examples will be given further in this report.

The figure 15 shown the relational model obtained with *DB-Main* for the continuous values. It is quite similar to the logical datamodel produced manually for the implementation of the prototype of the database. These schemas with the similar one for the non-continuous values are annexed to this report in a legibility form.

In the model of the IDOD database, there are *One to Many* and *Many to Many* relationships. Both will be explained for the transformation onto logical datamodel.

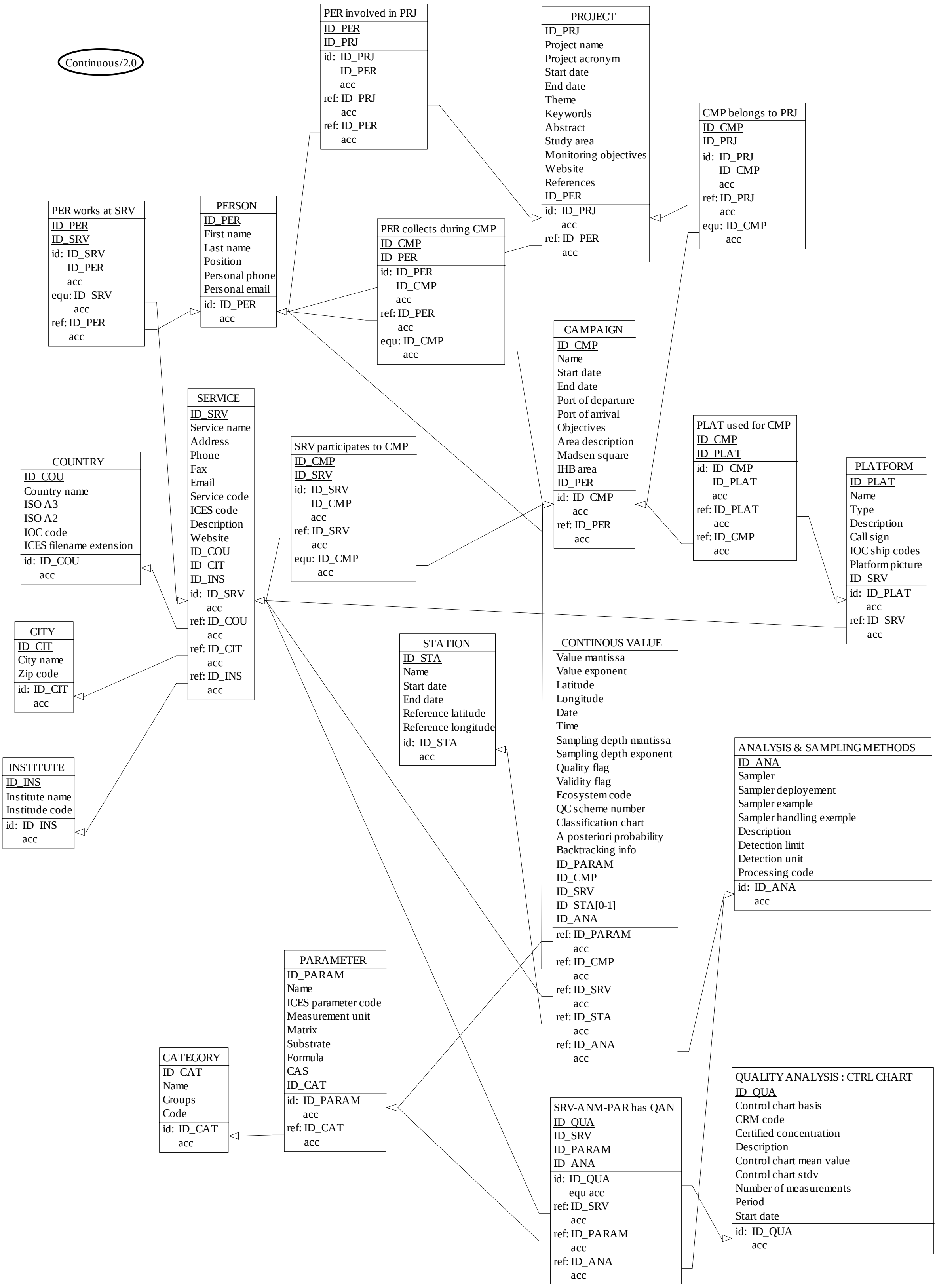


Figure 15. The relational datamodel for continuous values.

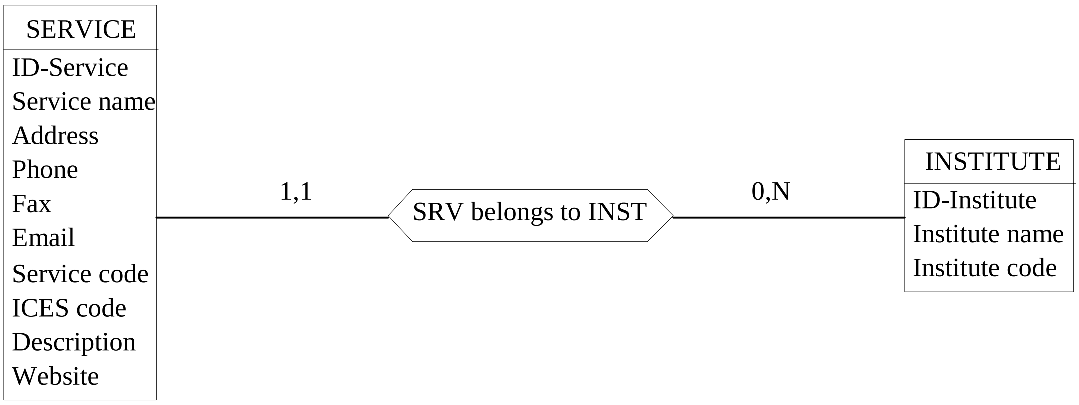


Figure 16. Conceptual datamodel.

The relation shown in the above figure 16 describes the belonging to a service to an institute. A service belongs to one and only one institute, but an institute groups 0 to N services.

The primary key of the service entity is *ID-Service* and for the institute entity it is *ID-Institute*.

The splitting of the service entity into different entities such as institute, city and country is a result of the normalisation procedure applied to the conceptual design. It is mainly to avoid incoherence into the database.

The transformation to the logical schema does not require an additional table (table is synonym of entity) because the relation is One-Many but it needs the adding of one foreign key to the Service table: *#Institute* that is the primary key of the Institute table.

A similar technique is applied to the relations with the city and country tables.

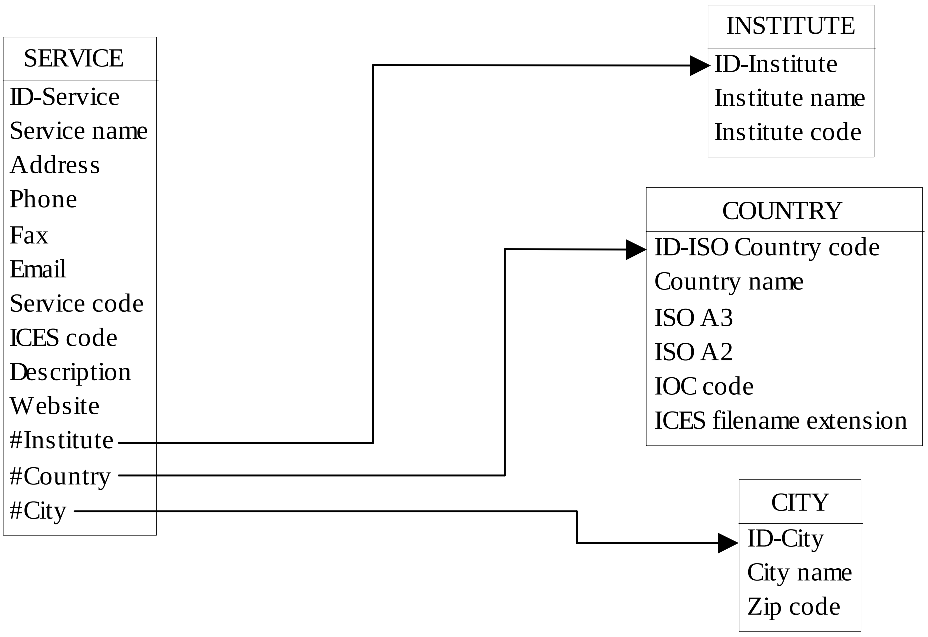


Figure 17. Logical datamodel.

All the three relations above are of type One-Many. For every records of the Service table, there are one link to the associated institute record, one link to the associated city record and one link to the associated country record. The arrow means that the link is mandatory: for each Service occurrence, the *Null* value is forbidden for the #Institute, #City and #Country pointers.

Another type of relation is the Many-Many as in the relation between the entities Project and Campaign. One Project participates to zero or N Campaigns and one Campaign is organised for 1 to N participating Projects.

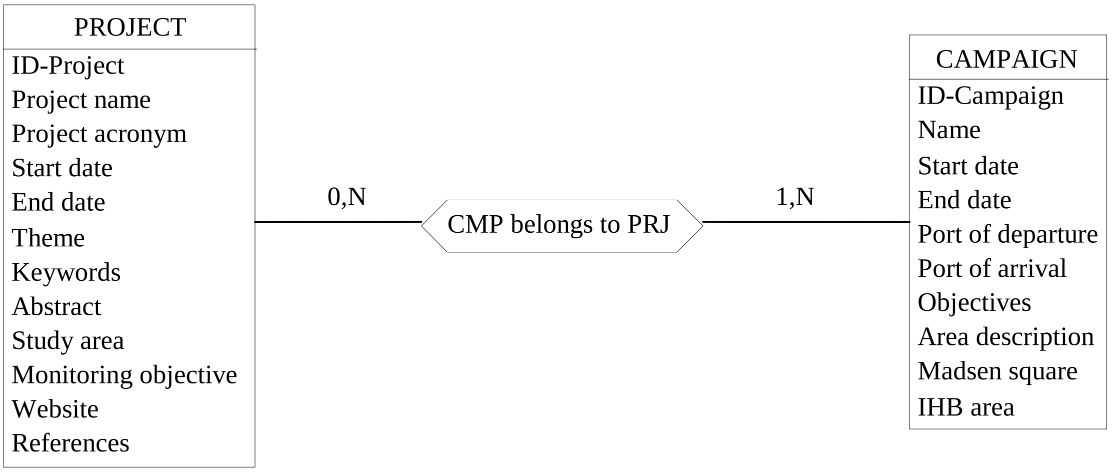


Figure 18. Conceptual datamodel.

For a relation Many-Many, it is necessary to incorporate a new table to establish the links between the two original tables. In this case, the new table will be called "Project – Campaign Relation". No foreign key is required within the two original tables.

The table "Project – Campaign Relation" is the logical representation of the relation "*CMP belongs to PRJ*" in the conceptual datamodel.

The new table contains two foreign keys that are the primary keys of the two linked tables. In this example, the cardinalities of the relation are different: (0,N) and (1,N). This is expressed in the logical datamodel as in the previous example by use of arrows onto the links. So, in the table "Project – Campaign Relation", all the records of the Campaign table have to be present and mentioned by their #ID-Camapaign. This means that every campaign includes almost one project. In other case, it is possible for a project to not participate to a campaign. In this second, case, the link is represented without an arrow and all the records of the Project table are not necessary present in the "Project – Campaign Relation" table.

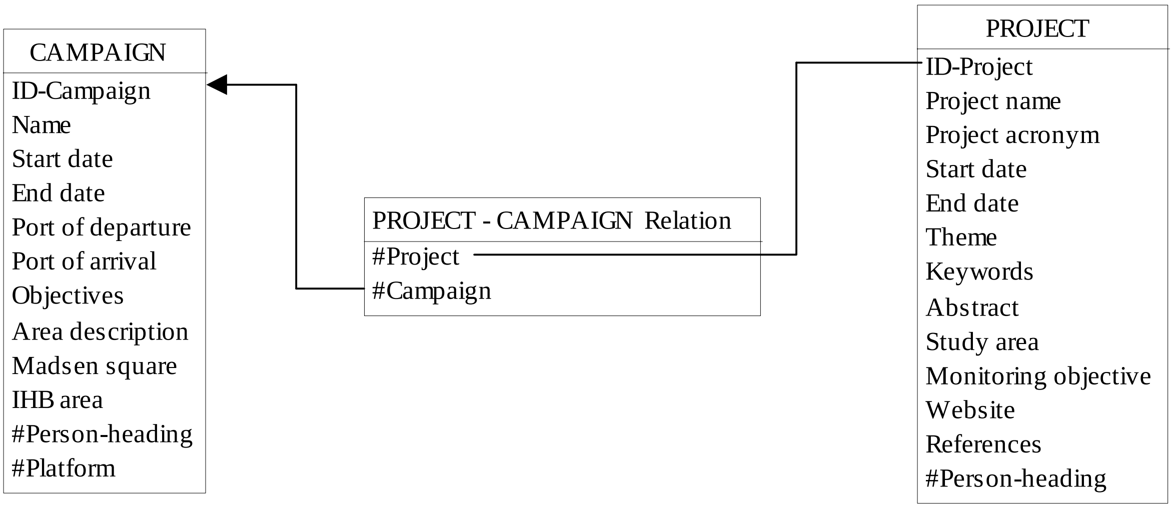


Figure 19. Logical datamodel.

Some pair entities of the conceptual datamodel are linked by more than one relation. For example, two relations "*PRJ supervised by PER*" and "*PER involved in PRJ*" link the entities Project and Person as in figure 20.

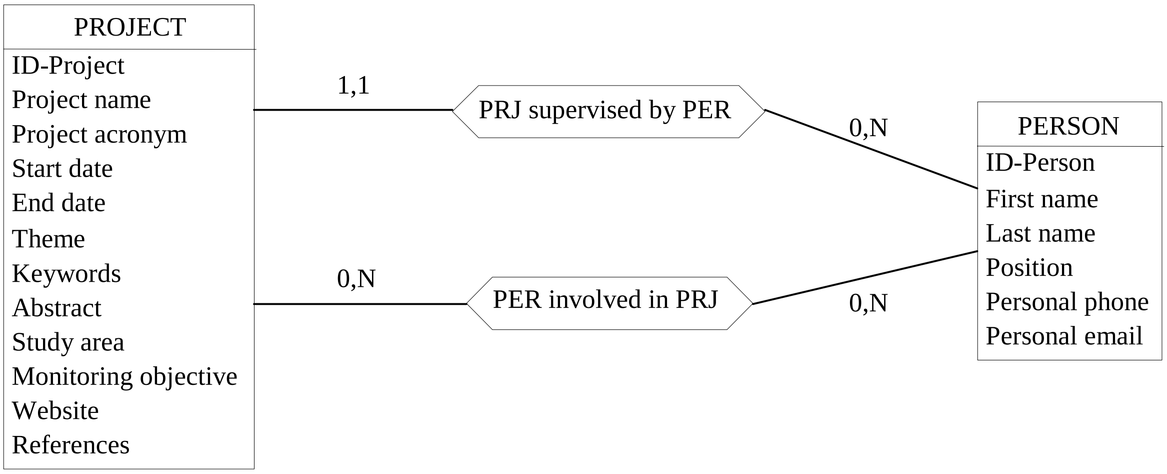


Figure 20. Conceptual datamodel.

This situation is converted into the logical design as illustrated in figure 21.

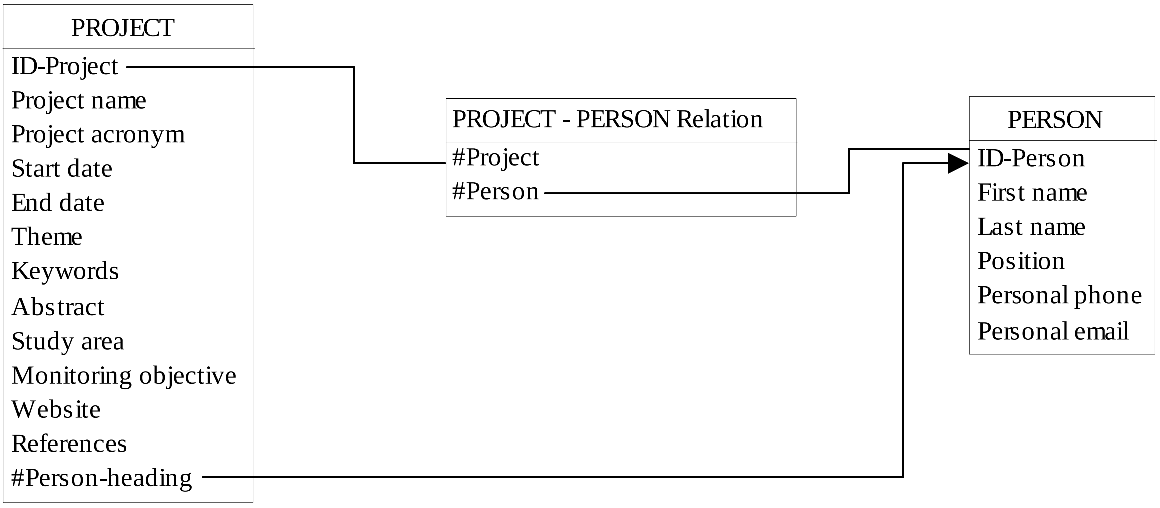


Figure 21. Logical datamodel.

In this last example, there is no arrow on the links with the transition table because these relations are not required. In opposite, any project always has a person-heading and then the arrow mentioned that the link is mandatory.

All the other relations from the conceptual datamodel are converted with the same rules of transformation. For these reasons, they will not be explained in details into this report. The interested readers can find the complete conceptual and logical datamodels annexed to this report. These models were designed using the *DB-Main* software and some concepts are sometimes differently represented. It is like this for the expression of the cardinalities after transformation to the relational model where the use of arrow is not use. In the previous presented examples, the requirement of a link is clearly expressed by the use of the arrow. It was all the same decided to use *DB-Main* for the drawing of the relational model to avoid the boring work of redrawing all the models after any modification added to the conceptual datamodel.

3.3. The logical datamodel tables

As mentioned in the previous section, all entities and relations of the conceptual datamodel have to be converted, after normalisation, to a relational or logical datamodel. The tables of the IDOD database logical datamodel are listed below. Some new tables were created: "Campaign – Person Relation", "Campaign – Project Relation", "Campaign – Service Relation", "Project – Person Relation", "Project – Service Relation" and "Service – Person Relation".

1. Service table:

SERVICE
ID-Service
Service name
Address
Phone
Fax
Email
Service code
ICES code
Description
Website
#Institute
#City
#Country

The #Institute, #City and #Country foreign keys are required because for each record of the Service table there is an associated institute, an associated city and an associated country that are all three mandatory.

2. City table:

CITY
ID-City
City name
Zip code

For the City table, no foreign key is added as consequence of the cardinalities of the relation associated with this entity.

3. Country table:

COUNTRY
ID-ISO Country code
Country name
ISO A3
ISO A2
IOC code
ICES filename extension

No foreign key to be added.

4. Institute table:

INSTITUTE
ID-Institute
Institute name
Institute code

No foreign key to be added.

5. Person table:

PERSON
ID-Person
First name
Last name
Position
Personal phone
Personal email

No foreign key to be added.

6. Project table:

PROJECT
ID-Project
Project name
Project acronym
Start date
End date
Theme
Keywords
Abstract
Study area
Monitoring objective
Website
References
#Person-heading

A new foreign key was added for the link with the person who is the chief supervisor of the project.

7. Campaign:

CAMPAIGN
ID-Campaign
Name
Start date
End date
Port of departure
Port of arrival
Objectives
Area description
Madsen square
IHB area
#Person-heading
#Platform

Two new foreign keys were added to establish the link with the chief supervisor of the campaign and with the platform used for the campaign.

8. Platform:

PLATFORM
ID-Platform
Name
Type
Description
Call sign
IOC ship codes
#Service

The foreign key #Service is added.

9. Sample:

SAMPLE
ID-Sample
Monitoring year
Sequence number
Sample code
Water depth
Meta wind speed
Meta wind direction
Meta sea state
Meta atmospheric
Meta air temperature
#Campaign
#Sampling method
#Project

Three foreign keys were added: #Campaign, #Sampling method, and #Project.

10. Subsample:

SUBSAMPLE
ID-Subsample Monitoring year Sequence number Sample code Subsample code #Sample #Sampling handling #Keeping service #Handling service

Four foreign keys were added: #Sample, #Sample handling, #Keeping service, and #Handling service.

11. Sampling method:

SAMPLING METHOD
ID-Sampling method Sampler Sampler deployment Description Sampler example Sampler handling exemple

No foreign key to be added.

12. Sample handling:

SAMPLE HANDLING
ID-Sample handling Sample preservation Sample pre-treatment Sample separation Procedure description

No foreign key to be added.

13. Analysis method:

ANALYSIS METHOD
ID-Analysis method
Description
Detection limit
Detection unit
#Parameter

One foreign key was added.

14. Analysis & sampling methods:

ANALYSIS & SAMPLING METHODS
ID-Analysis & sampling methods
Sampler
Sampler deployement
Sampler example
Sampler handling exemple
Description
Detection limit
Detection unit
Processing code

No foreign key to be added.

15. Category:

CATEGORY
ID-Category
Name
Groups
Code

No foreign key to be added.

16. Parameter:

PARAMETER
ID-Parameter
Name
ICES parameter code
Measurement unit
Matrix
Substrate
Formula
CAS
#Category

One foreign key was added.

17. Station:

STATION
ID-Station
Name
Start date
End date
Reference latitude
Reference longitude

No foreign key to be added.

18. Non-continuous value:

NON-CONTINUOUS VALUE
ID-Non-continuous value
Value mantissa
Value exponent
Latitude
Longitude
Date
Time
Sampling depth mantissa
Sampling depth exponent
Replicate number
Quality flag
Validity flag
Ecosystem code
QC scheme number
Classification chart
A posteriori probability
Backtracking info
#Station
#Analysis method
#Parameter
#Subsample
#Service

Five foreign keys were added: #Station, #Analysis method, #Parameter, #Subsample, and #Service.

19. Continuous value:

CONTINUOUS VALUE
ID-Continuous value
Value mantissa
Value exponent
Latitude
Longitude
Date
Time
Sampling depth mantissa
Sampling depth exponent
Quality flag
Validity flag
Ecosystem code
QC scheme number
Classification chart
A posteriori probability
Backtracking info
#Station
#Analysis & Sampling methods
#Parameter
#Campaign
#Service

Five foreign keys were added: #Station, #Analysis & Sampling methods, #Parameter, #Campaign, and #Service.

20. Quality sample and handling:

Not introduced into the prototype because there were no data available and some links are required. The adjunction of this table can be done quickly using the same transformation process.

21. Quality analysis – Quasimeme:

Not introduced into the prototype because there were no data available and some links are required. The adjunction of this table can be done quickly using the same transformation process.

22. Quality analysis – Control chart:

Not introduced into the prototype because there were no data available and some links are required. The adjunction of this table can be done quickly using the same transformation process.

23. Quality analysis – Inter-calibration:

Not introduced into the prototype because there were no data available and some links are required. The adjunction of this table can be done quickly using the same transformation process.

24. Campaign – Person Relation:

CAMPAIGN - PERSON Relation
#Campaign #Person

The two foreign keys #Campaign and #Person make the correspondence Many-Many between the tables Campaign and Person.

25. Campaign – Project Relation:

CAMPAIGN - PROJECT Relation
#Campaign #Project

The two foreign keys #Campaign and #Project make the correspondence Many-Many between the tables Campaign and Project.

26. Campaign – Service Relation:

CAMPAIGN - SERVICE Relation
#Campaign #Service

The two foreign keys #Campaign and #Service make the correspondence Many-Many between the tables Campaign and Service.

27. Project – Person Relation:

PROJECT - PERSON Relation
#Project #Person

The two foreign keys #Project and #Person make the correspondence Many-Many between the tables Project and Person.

28. Project – Service Relation:

PROJECT - SERVICE Relation
#Project #Service

The two foreign keys #Project and #Service make the correspondence Many-Many between the tables Project and Service.

29. Service – Person Relation:

SERVICE - PERSON Relation
#Service #Person

The two foreign keys #Service and #Person make the correspondence Many-Many between the tables Service and Person.

3.4. The DDL code to create the database

The DDL code (Data Definition Language) permit to define the structure of the database. The code was generated with the *DB-Main* case tool is presented in Annex B.

This code shows what is possible to realise automatically with a case tool. Because of some limitations of DB-Main, this case tool was not used for the generation of the prototype of the IDOD database and all the processes were realised manually.

3.5. The prototype with *Access97*

The choice of a Database Management System (DBMS) for the implementation of the IDOD database prototype was influenced by the opportunity of having the same DBMS in all the laboratories participating to the IDOD project. A common DBMS is *Microsoft Access97* available at Laboratory SURFACES, MUMM and UCS.

Microsoft Access97 is a relational DBMS offering a friendly interface to generate tables, forms, states, requests and macros. With its ODBC (Open Database Connectivity), it permits the connection and sharing of data with a lot of software. The ODBC connection will be used for the access to the database via the World Wide Web.

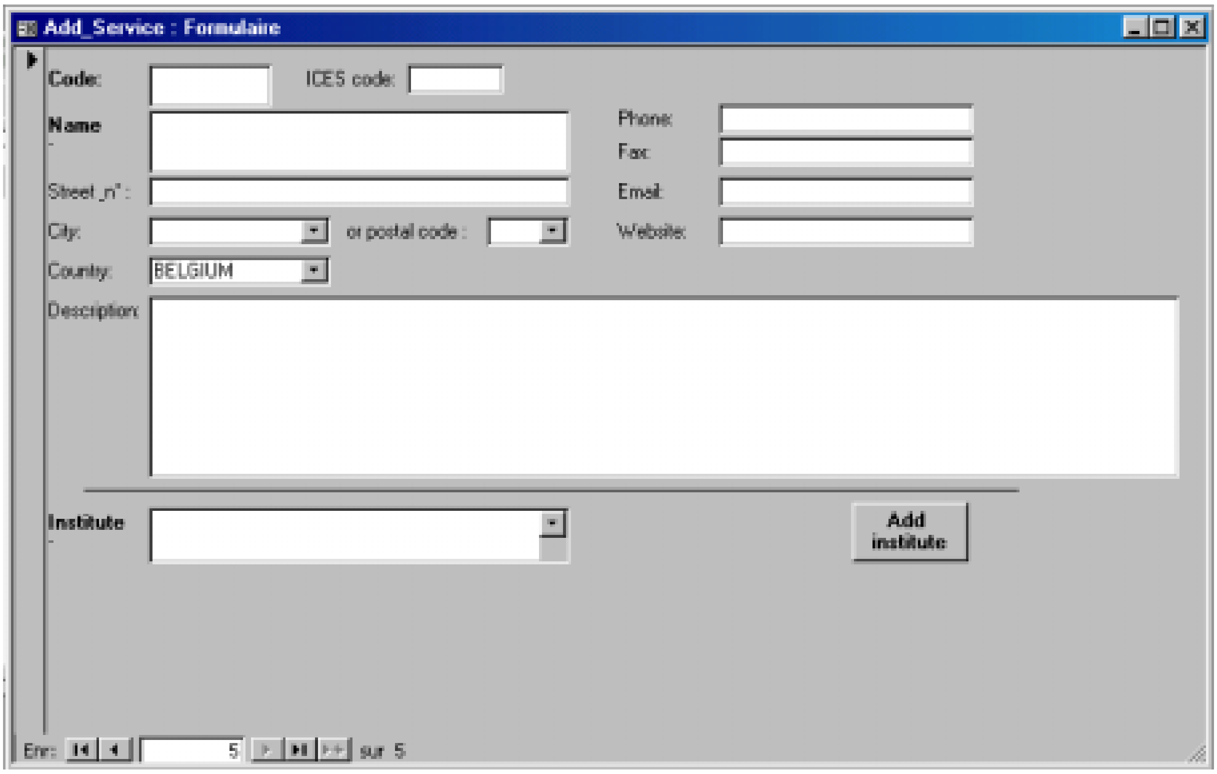
The fulfilling of the database prototype was made using temporary tables and specific adding requests. The data were provided in the "xls" format of *Microsoft Excel97* sheets. The data were firstly imported into temporary tables. The links between the tables were established step by step from the temporary tables by specific adding requests.

It is out of the scope of this report to present the specificity of Access97 and the interested readers are invited to report to the literature mentioned in the bibliography.

The prototype is now available in *Microsoft Access97* format as a MDB file.

3.5.1. Input of data using forms

Access97 allows defining forms for an easy input or browsing of data. Some examples of screen hardcopy show the forms used to add a new service or institute.



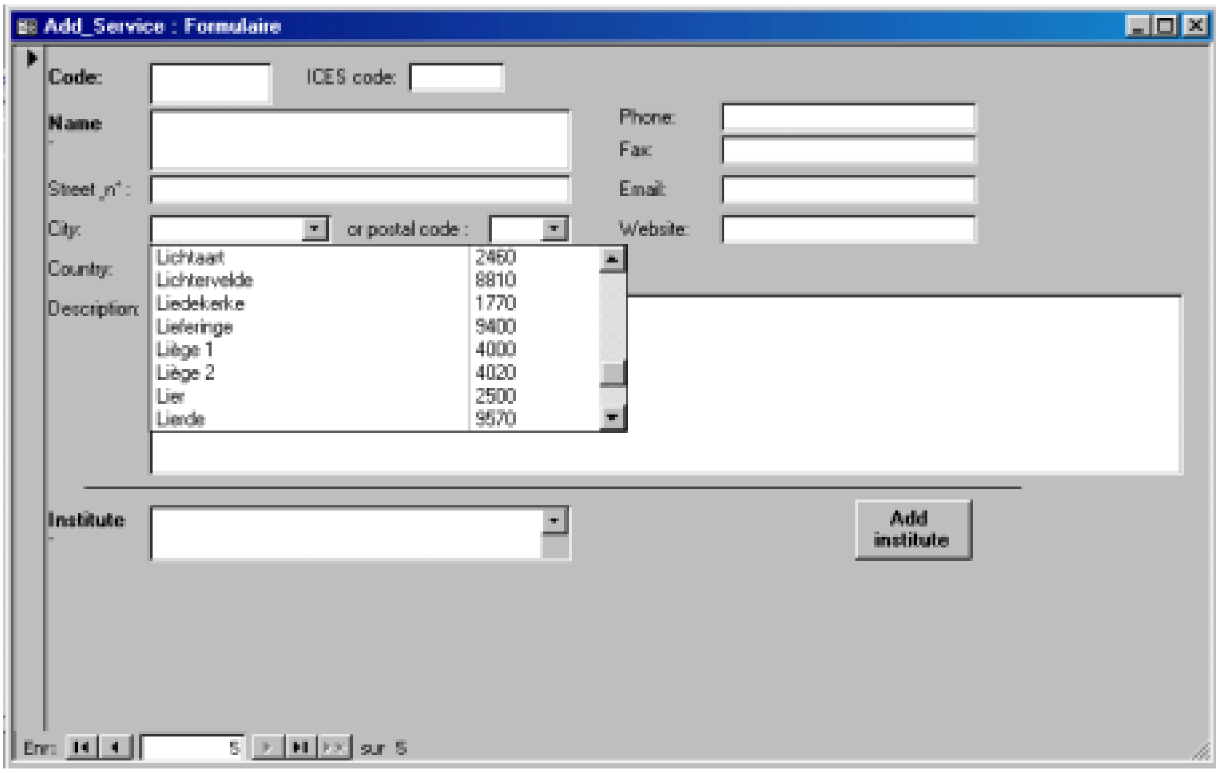
Country: BELGIUM

Enr: 5 sur 5

The service has to be attached to an institute that can be found in the scrollable list in the input form frame for institute. If the required institute does not exist in this list, the user can add the institute by clicking on the "Add institute" button.

The country, city and postal code are also available in scrollable list. In this case, it is only required to fill the city or postal code input form frame because these two attribute are part of the same entity.

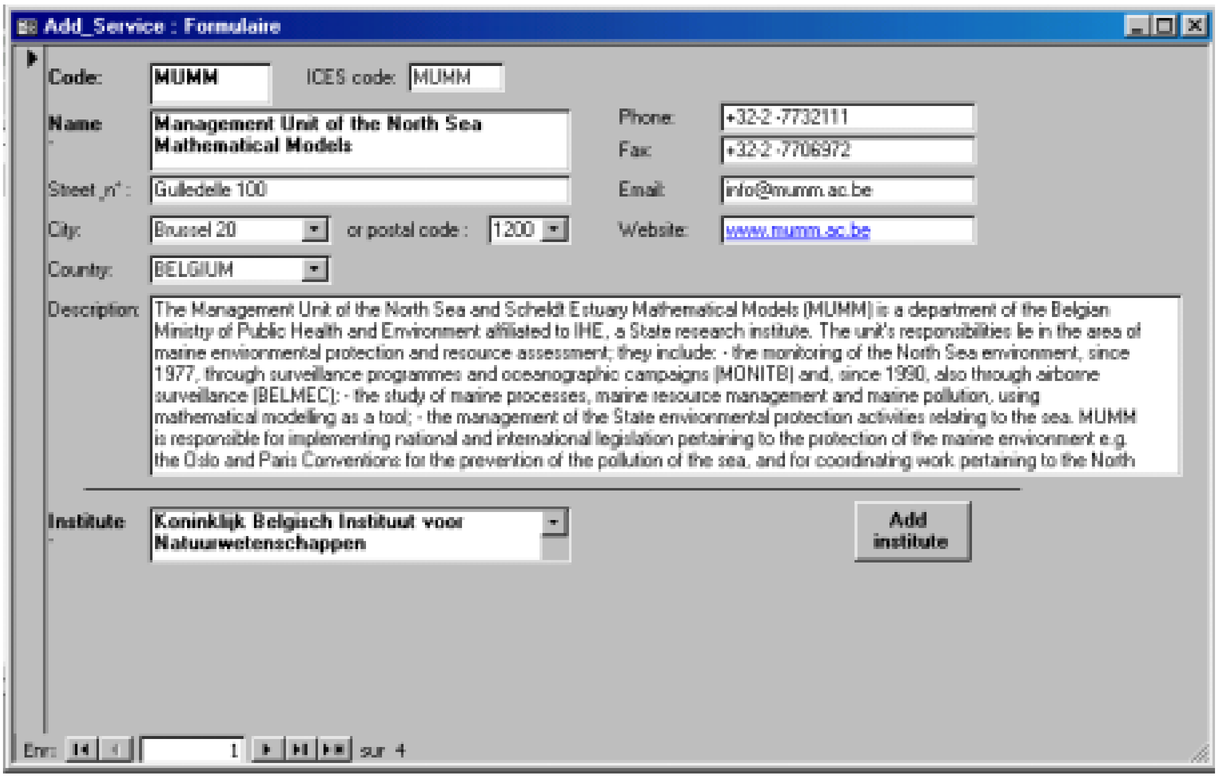
The following figure shows an example of the scrollable list for the cities with associated postal code.



Lichtaart	2460
Lichtervelde	8810
Liedekerke	1770
Lieferinge	9400
Liège 1	4000
Liège 2	4020
Lier	2500
Lierde	9570

As mentioned, the user has to scroll through the list to chose the city where the service is located. The corresponding postal code is automatically added.

The form can also be used to browse through the existing data. For example, it is possible to display the information relative to a particular service.



Code: MUMM ICES code: MUMM

Name: Management Unit of the North Sea Mathematical Models

Phone: +32-2-7732111

Fax: +32-2-7706972

Street ,n° : Gulledelle 100

Email: info@mumm.ac.be

City: Brussel 20 or postal code : 1200

Website: www.mumm.ac.be

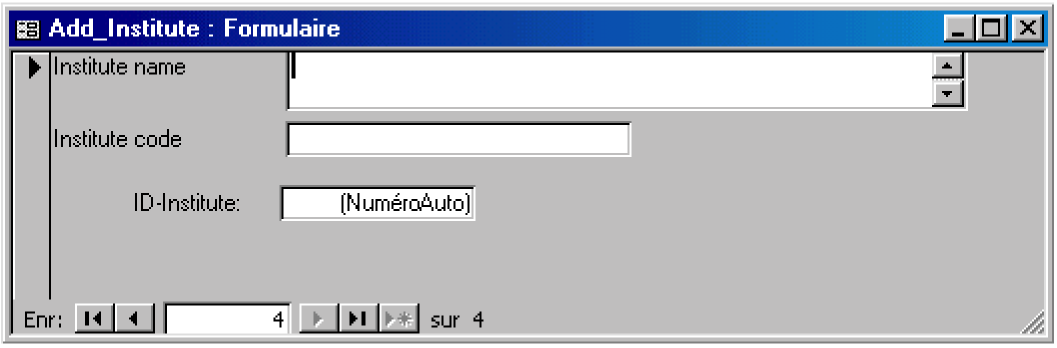
Country: BELGIUM

Description: The Management Unit of the North Sea and Scheldt Estuary Mathematical Models (MUMM) is a department of the Belgian Ministry of Public Health and Environment affiliated to IHE, a State research institute. The unit's responsibilities lie in the area of marine environmental protection and resource assessment; they include: - the monitoring of the North Sea environment, since 1977, through surveillance programmes and oceanographic campaigns (MONITB) and, since 1990, also through airborne surveillance (BELMEC); - the study of marine processes, marine resource management and marine pollution, using mathematical modelling as a tool; - the management of the State environmental protection activities relating to the sea. MUMM is responsible for implementing national and international legislation pertaining to the protection of the marine environment e.g. the Oslo and Paris Conventions for the prevention of the pollution of the sea, and for coordinating work pertaining to the North

Institute: Koninklijk Belgisch Instituut voor Natuurwetenschappen

In the previous example figure, the information concerning the MUMM service is listed using the form. In the lower left corner of the form, there are buttons to scroll through the records in the table. In this example, there are four services registered into the database. The adding of a new record can be done by moving up to the fifth position.

When the adding of a new institute is required, the action of pushing on the "Add institute" button opens a new form for institute input as illustrated in the following figure.



Add_Institute : Formulaire

Institute name

Institute code

ID-Institute: (NuméroAuto)

Enr: 4 sur 4

Similar input forms are used to introduce or update manually the content of the database.

The data can also be displayed using sheet mode for consultation, updating or adding of data. The following example shows this mode for the content of the Station table.

ID-Station	Name	Reference latitude	Reference longitude	Start date	End date
31	105	51,1833333333333	2,475	1/01/96	31/12/96
32	115	51,155	2,60333333333333	1/01/96	31/12/96
33	120	51,185	2,70116666666667	1/01/96	31/12/96
34	130	51,2708333333333	2,905	1/01/96	31/12/96
35	150	51,4161666666667	3,40216666666667	1/01/96	31/12/96
36	215	51,2766666666667	2,61333333333333	1/01/96	31/12/96
37	230	51,3083333333333	2,85	1/01/96	31/12/96
38	250	51,5166666666667	3,31666666666667	1/01/96	31/12/96
39	315	51,3228333333333	2,464	1/01/96	31/12/96
40	330	51,4333333333333	2,80833333333333	1/01/96	31/12/96
41	421	51,4805	2,45	1/01/96	31/12/96
42	435	51,5806666666667	2,79033333333333	1/01/96	31/12/96
43	545	51,7266666666667	3,05	1/01/96	31/12/96
44	700	51,3766666666667	3,22	1/01/96	31/12/96
45	710	51,4408333333333	3,13866666666667	1/01/96	31/12/96
46	800	51,8471666666667	2,86666666666667	1/01/96	31/12/96
47	S01	51,4166666666667	3,57	1/01/96	31/12/96
48	S04	51,345	3,825	1/01/96	31/12/96
49	S07	51,4366666666667	4	1/01/96	31/12/96
50	S09	51,37	4,07833333333333	1/01/96	31/12/96
51	S12	51,365	4,225	1/01/96	31/12/96

52	S15	51,3133333333333	4,27333333333333	1/01/96	31/12/96
53	S15b	51,2891666666667	4,32233333333333	1/01/96	31/12/96
54	S18	51,2666666666667	4,3	1/01/96	31/12/96
55	S18b	51,2548333333333	4,3175	1/01/96	31/12/96
56	S20	51,2408333333333	4,35	1/01/96	31/12/96
57	S22	51,2188333333333	4,39166666666667	1/01/96	31/12/96
58	ZG01	51,3333333333333	2,7	1/01/96	31/12/96
59	ZG02	51,3333333333333	2,5	1/01/96	31/12/96
60	ZG03	51,2616666666667	2,66666666666667	1/01/96	31/12/96

3.5.2. Queries on data

The interrogation of the database implies the use of specific requests. These requests use the relational algebra to submit the queries to the relational DBMS. The SQL code is a way to perform such queries.

Many queries are implemented in the prototype of the database and the results are presented as states that are text outputs.

Example 1:

Selection of all campaigns with associated services and platform.

This request can be performed using the following SQL code:

```
SELECT Campaign.Name, Campaign.[Start date], Campaign.[End date],
Campaign.Objectives, Service.[Service code], Platform.Name
FROM Service RIGHT JOIN (Platform RIGHT JOIN (Campaign RIGHT JOIN
[Relation Campaign-Service] ON Campaign.[ID-Campaign] = [Relation Campaign-
Service].[#Campaing]) ON Platform.[ID-Platform] = Campaign.[#Platform])
ON Service.[ID-Service] = [Relation Campaign-Service].[#Service];
```

The result of the request is shown in the Annex C.

Example 2:

List all stations with the total number of corresponding visits classified by campaigns.

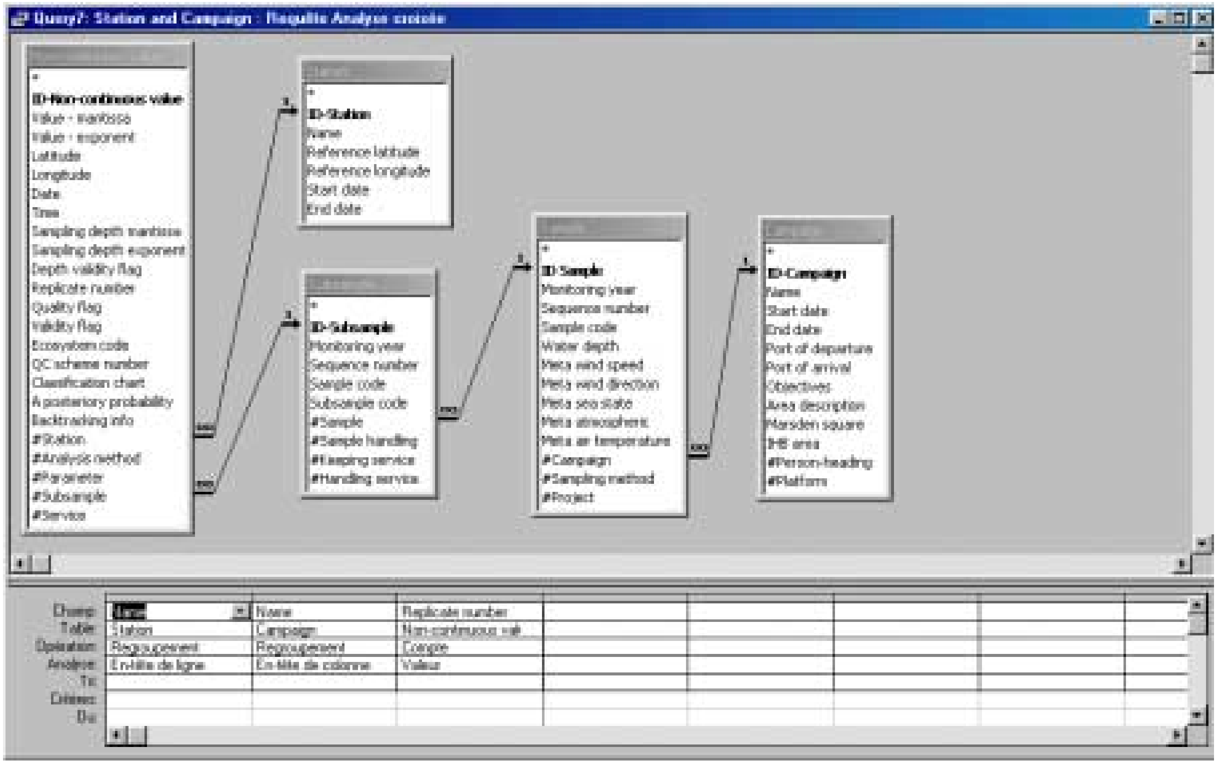
This request can be performed using the following SQL code:

```
TRANSFORM Count([Non-continuous value].[Replicate number]) AS
[CompteDeReplicate number]
SELECT Station.Name
```

FROM ((Campaign RIGHT JOIN Sample ON Campaign.[ID-Campaign] = Sample.[#Campaign]) RIGHT JOIN Subsample ON Sample.[ID Sample] = Subsample.[#Sample]) RIGHT JOIN (Station RIGHT JOIN [Non-continuous value] ON Station.[ID-Station] = [Non-continuous value].[#Station]) ON Subsample.[ID-Subsample] = [Non-continuous value].[#Subsample]
GROUP BY Station.Name
PIVOT Campaign.Name;

The result of the request is shown in Annex D.

Another way to perform the same query with Access97 is to define the following request in graphical mode.



Champ:	Name	Name	Replicate number
Table:	Station	Campaign	Non-continuous val
Opération:	Regroupement	Regroupement	Compte
Analyse:	En-tête de ligne	En-tête de colonne	Valeur
Tri:			
Critères:			
Ou:			

Example 3:

List all values from ICES parameter code "NTRA", "NTRI", "PSAL" or "CPHL*" and collected between 1996/03/10 and 1996/10/31, with associated station, campaign and project.*

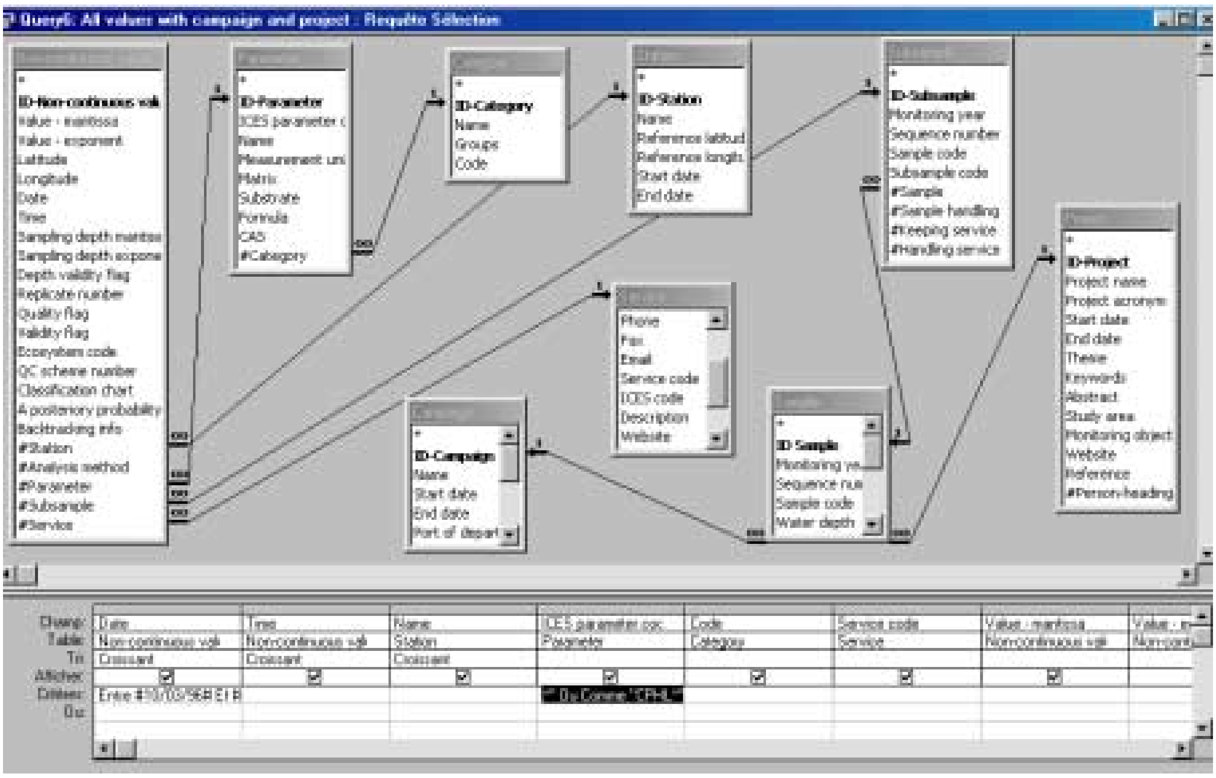
This request can be performed using the following SQL code:

SELECT [Non-continuous value].Date, [Non-continuous value].Time, Station.Name, Parameter.[ICES parameter code], Category.Code, Service.[Service code], [Non-continuous value].[Value - mantissa], [Non-continuous value].[Value - exponent], [Non-continuous value].Latitude, [Non-continuous value].Longitude, Campaign.Name, Project.[Project name]

FROM ((Project RIGHT JOIN (Campaign RIGHT JOIN Sample ON Campaign.[ID-Campaign] = Sample.[#Campaign]) ON Project.[ID-Project] = Sample.[#Project]) RIGHT JOIN Subsample ON Sample.[ID Sample] = Subsample.[#Sample]) RIGHT JOIN (Station RIGHT JOIN (Service RIGHT JOIN ((Category RIGHT JOIN Parameter ON Category.[ID-Category] = Parameter.[#Category]) RIGHT JOIN [Non-continuous value] ON Parameter.[ID-Parameter] = [Non-continuous value].[#Parameter]) ON Service.[ID-Service] = [Non-continuous value].[#Service]) ON Station.[ID-Station] = [Non-continuous value].[#Station]) ON Subsample.[ID-Subsample] = [Non-continuous value].[#Subsample]
WHERE ((([Non-continuous value].Date) Between #3/10/96# And #10/31/96#) AND ((Parameter.[ICES parameter code])="NTRA" Or (Parameter.[ICES parameter code])="NTRI" Or (Parameter.[ICES parameter code]) Like "PSAL*" Or (Parameter.[ICES parameter code]) Like "CPHL*"))
ORDER BY [Non-continuous value].Date, [Non-continuous value].Time, Station.Name;

The result of the request is shown in Annex E.

Another way to perform the same query with Access97 is to define the following request in graphical mode.



Champ:	Date	Time	Name	ICES parameter coc	Code	Service code	Value - mantissa	Value - e
Table:	Non-continuous valt	Non-continuous val	Station	Parameter	Category	Service	Non-continuous val	Non-conti
Tri:	Croissant	Croissant	Croissant					
Afficher:	☑	☑	☑	☑	☑	☑	☑	
Critères:	Entre #10/03/96# Et #			" Ou Comme "CPHL*"				
Ou:								

3.6. Perspectives

In a near future, the database will be extended to the other data types that are sediments, biota, data from mathematical models, etc. The data dictionary will be extended to incorporate all these new information.

The purchase of a case tool is envisaged in the next months after a benchmark for evaluation of the potentialities of all the different products available on the market. With such a tool, the conception of the final database will be easier. Indeed, some case tools are able to translate directly from the conceptual datamodel to the physical implementation into a DBMS. The auxiliary data such as the documentation or data dictionary of the database can also be generated and updated automatically. All modifications to the conceptual datamodel can be replied to the DBMS without having to make again all the complex steps necessary for the normalisation, conversion to intermediate datamodel, and physical implementation.

The final choice for a DBMS will be fixed when all the requirements for the IDOD database will be exactly known. The choice will be oriented to robust systems such as *Microsoft SQL Server, Sybase, Oracle,* etc.

Glossary

DBMS *Database Management System*. A program or collection of programs used to establish, maintain, and process a database.

Data dictionary A user-accessible catalogue of data about the database. An *active data dictionary* is one whose entries are updated automatically whenever changes are made to the database structure. A *passive data dictionary* is one whose entries need to be manually updated whenever database structure changes are made.

Database A self-describing collections of integrated records.

DDL *Data Definition Language*. The portion of a data model pertaining to the definition of database structure.

Entity A class of things that exist in the user's business environment.

Entity/Relationship diagram A diagram used in database design that illustrates business entities and associations among them. Also called *E/R diagram*.

Foreign key An attribute that is a key of a different relation.

GIS *Geographical Information System*.

IS *Information System*.

Meta-data Data about the structure of a database that is stored in the data dictionary.

Normal form A rule or set of rules governing the structure of a relation in order to eliminate certain modification anomalies. The normal forms have been defined: first, second, third, Boyce-Codd, fifth, and domain-key.

Normalisation The process by which attributes are grouped together to form a well-structured relation.

ODBC *Open Database Connectivity*.

Primary key A group of one or more attributes that uniquely identifies a record.

Property A characteristic of an object.

Relational data model A data model in which (1) data is stored in tables and (2) associations between tables are represented within the table data, not in overhead data structures.

Relationship An association between two or more entities.

SQL *Structured Query Language*. A relational database access language developed by IBM. SQL became the ANSI database standard in 1986.

Bibliography

DELMAL Pierre. *SQL 2. Application à Oracle, Access et RDB.* De Boeck Université,1998, 2ème éd., in-8°, 580 p.

DIONISI Dominique. *L'essentiel sur MERISE.* Eyrolles, 1997, in-8°, 257 p.

EMERI William J. & THOMSON Richard E. *Data analysis methods in physical oceanography*. Pergamon, 1997, in-8°, 634 p.

HOMER Alex, GILL Darren, JAKAB Stephen. *Interface entre Web et bases de données sous Windows NT*. Eyrolles, 1998, in-8°, 434 p.

KROENKE David M. & DOLAN Kathleen A. *Database Processing: fundamentals, design, implementation*. Science Research Associates, 1988, 3rd ed., in-8°, xxii+687 p.

NANCI Dominique & ESPINASSE Bernard. *Ingénierie des systèmes MERISE.* Sybex, 1996, 3ème éd., in-8°, 891 p.

PANTAZIS Dimos & DONNAY Jean-Paul. *La conception de SIG, méthode et formalisme*. Hermès, 1996, in-8°, 343 p.

SIMPSON Alan. *Access 97 pour Windows 95*. Sybex, 1997, in-4°, 1060 p.

VIESCAS John. *Microsoft Access 97 au quotidien*. Microsoft Press, 1997, in-4°, 969 p.

Annex A

This annex lists the main transformations and processes required passing from the first conceptual design of the IDOD database to its physical implement into *Access97*. They include the normalisation and the conversion to the logical model.

Conception and elaboration of a prototype for the IDOD database

Realisation: Lucien Schwind (MUMM)
Fabrice Muller (SURFACES – ULg)

Conceptual datamodel validation

Both conceptual models (continuous data and non-continuous data) have to be validated before their conversion to logical models. This step requires:

1. The checking of the entities and of the coherence of the properties of these ones;
2. The checking of the relations and the associated cardinalities;
3. The validation of the CDM (conceptual datamodel) following the first normal form;
4. The validation of the CDM following the second normal form;
5. The validation of the CDM following the third normal form.

The normalisation of the CDM aims to the following objectives:

- To represent all the relations minimising the redundancies;
- To make easier the data manipulations (insertion, updating and suppression) of relations avoiding the data storing anomalies;
- To reduce the necessity of restructuring of the relations after the adding of new data types.

The three first normal forms are required to avoid the functional dependencies inside the entities.

During the last months, the elaboration of the CDM was made with some respect of the design logic to end up at a model that does not required fundamental transformations before its conversion to a relational model. However, a few adaptations were needed as listed below.

Adaptation of the conceptual datamodel

To satisfy the first three normal forms, the following changes were applied to the E/R design model:

1. The SERVICE entity: the multilingual properties such as *Address*, *Name* and *Institute name* were removed and replaced by single properties.

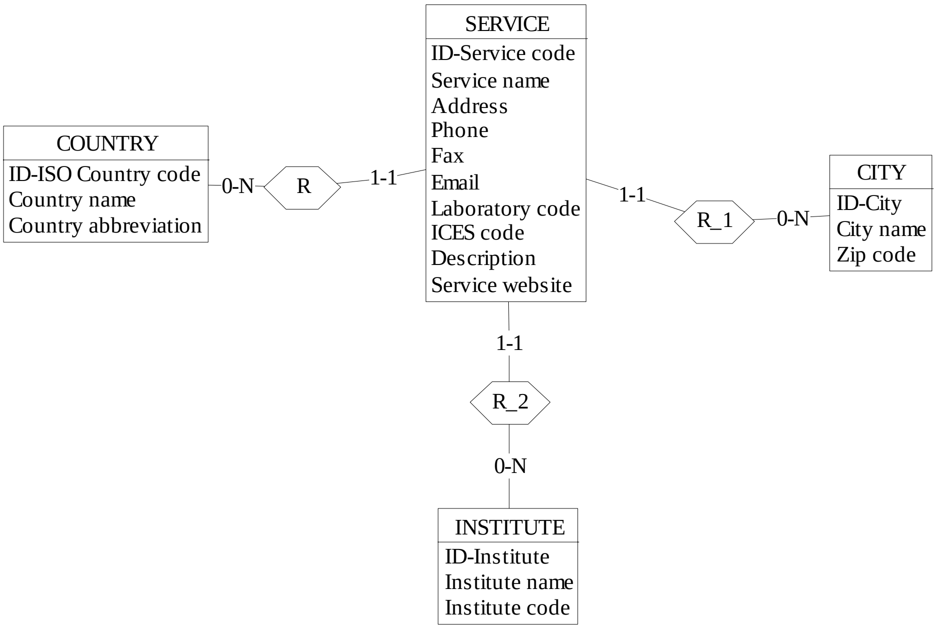
2. The PROJECT entity: the *Sponsor* and *Research Programme* properties must be extracted and converted into two new entities. *Sponsor* could also be considered as an occurrence of the SERVICE entity.
3. The QUALITY SAMPLE & HANDLING entity: this entity will not be included in the prototype of the database because it contains no property and is nonsense. This entity will be added when properties will be defined.
4. The CONTINUOUS VALUE and NON-CONTINUOUS VALUE entities: the *Comment* property must be extracted and converted into a new entity.
5. The *SRV supervised by PER* relation between the SERVICE and PERSON entities was removed.
6. The *SRV leads the CMP* relation between the SERVICE and CAMPAIGN entities was removed.
7. Several cardinalities were adapted to satisfy the coherence of the relations:
   7.1. Inside the *ANM linked to PARAM* relation between the PARAMETER and ANALYSIS METHOD entities, the 0-N cardinality changed to 1-N.
   7.2. Inside the *PER works at SRV* relation between the PERSON and SERVICE entities, the 0-N cardinality changed to 1-N.
   7.3. Inside the *PRJ collects SP* relation between the PROJECT and SAMPLE entities, the 0-1 cardinality changed to 1-1.
   7.4. Inside the *PER collects during CMP* relation between the PERSON and CAMPAIGN entities, the 0-N cardinality changed to 1-N.

8. Some properties of the SERVICE entity were extracted to constitute new entities in respect of the three first normal forms. For the 1NF (first normal form) it is required to split the *Address* property into one field for the street name and the office number, while the city, zip code and country are sent in other tables. Three new entities are created: INSTITUTE, CITY and COUNTRY. For each occurrence of SERVICE there are links to corresponding occurrences of the new entities.

   INSTITUTE:
   ID-Institute: the primary key of the entity.
   Institute name: the name of the institute.
   Institute code: the code of the institute.

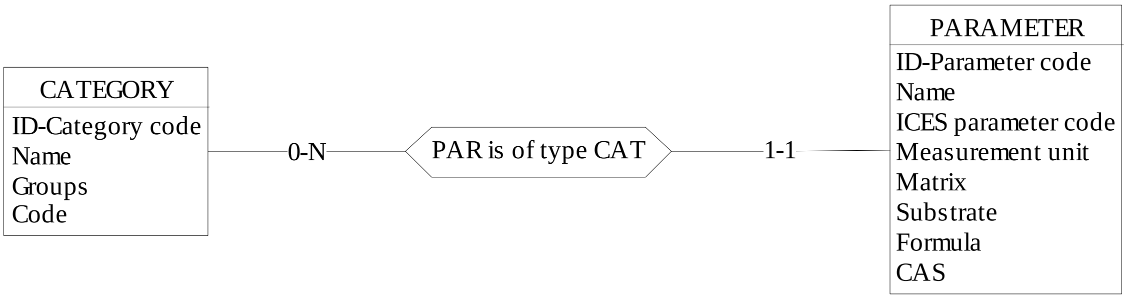
   CITY:
   ID-City: the primary key of the entity.
   City name: the name of the city.
   Zip code: the postal zip code of the city.

   COUNTRY:
   ID-ISO country code: the primary key country code from ISO 3166.
   Country name: the name of the country.
   Country abbreviation: the abbreviated name of the country.



9. In the conceptual datamodel, the QUALITY ANALYSIS entity is specialised into three sub-entities: QUASIMEME, CONTROL CHART and INTER-CALIBRATION. The only common properties are the *ID co*de and the *Start date*. So, in the relational design, there will be three distinct entities, all with their own properties.
10. According to the available information data, some properties were added to the COUNTRY entity: *Country name*, *ISO A3*, *ISO A2*, *IOC code*, *ICES filename extension*.
11. Data from the "Service.xls" file were added to the INSTITUTE table.
12. A complete list of the Belgian zip codes was downloaded from the World Wide Web http://www.charline.be/info/codepost/cpost.htm. This list mentions the zip codes and the associated locality names. This information was added to the CITY table.
13. Data from the "Country codes.xls" file were added to the COUNTRY table.
14. Data from the "Services.xls" file were added to the SERVICE table. The required links with related tables were created automatically during this input procedure by the use of specific adding requests in SQL script. The SERVICE table required links with the PLATFORM and PERSON tables and the *Null* value is forbidden for these links.
15. Data from the "Campaign.xls" file were added to the CAMPAIGN table. The required links with related tables were created automatically during this input procedure by the use of specific adding requests in SQL script. The CAMPAIGN table required links with the PROJECT, PLATFORM, SERVICE and PERSON tables and the *Null* value is forbidden for the links with the PLATFORM, SERVICE and PERSON tables.
16. The CAMPAIGN-SERVICE relation table was fulfilled to establish the links between the CAMPAIGN and SERVICE tables for the relation *SRV participates to CMP*.
17. The CAMPAIGN-PERSON relation table was fulfilled to establish the links between the CAMPAIGN and PERSON tables for the relation *PER collects during CMP*.
18. The PROJECT-PERSON relation table was fulfilled to establish the links between the PROJECT and PERSON tables for the relation *PER involved in PRJ*.
19. The PROJECT-SERVICE relation table was fulfilled to establish the links between the PROJECT and SERVICE tables for the relation *SRV belongs to PRJ*.

20. The CAMPAIGN-PROJECT relation table was fulfilled to establish the links between the CAMPAIGN and PROJECT tables for the relation *CMP belongs to PRJ*.
21. The SERVICE-PERSON relation table was fulfilled to establish the links between the SERVICE and PERSON tables for the relation *PER works at SRV*.
22. Data were introduced into the SAMPLE table. The required links with the CAMPAIGN, SAMPLING METHOD and PROJECT tables were created using SQL scripts requests. The required link with SUBSAMPLE will be established when the data for the subsamples will be available.
23. Data were introduced into the ANALYSIS METHOD table.
24. Data were introduced into the CATEGORY table. This table results from the splitting of the PARAMETER table according to the third normal form.



Data were introduced into the PARAMETER table. The required links with the ANALYSIS METHOD and CATEGORY tables were created using SQL scripts.

25. In the SAMPLE table, the following properties were added:
    - Monitoring year: integer type;
    - Sequence number: integer type;
    - Sample code: integer type.

    These three new attributes are necessary for the input of data into the NON-CONTINUOUS VALUE table.
26. Data were introduced into the SUBSAMPLE table.
27. In the NON-CONTINUOUS VALUE table, the *Latitude* and *Longitude* attributes are required only if the value for the link to STATION is *Null*.
28. In the NON-CONTINUOUS VALUE table, the *Sampling depth* attribute is split into two new attributes: *Sampling depth mantissa* and *Sampling depth exponent*, both of integer type.
29. In the NON-CONTINUOUS VALUE table, one new attribute was added: *Depth validity code* of integer type.

Annex B

In this annex is presented the DDL code that can be used to generate the structure of the database for non-continuous value. A very similar DDL code can be obtained for the continuous value. The code was generated using the *DB-Main* case tool.

The types of some attributes are not always in concordance with the actual ones defined within the IDOD database dictionary. This implies to the restriction inherent to *DB-Main*.

```
-- *********************************************
-- * Standard SQL generation                   *
-- *-------------------------------------------*
-- * Generator date:  Nov  2 1998              *
-- * Generation date: Wed Dec 23 15:44:23 1998 *
-- *********************************************


-- Database Section
-- ________________

create database Non-continuous;


-- DBSpace Section
-- _______________


-- Table Section
-- _____________

create table ANALYSIS METHOD (
     ID_ANA char(10) not null,
     ID-Analysis method numeric(10) not null,
     Description char(1000) not null,
     Detection limit float(1) not null,
     Detection unit char(15) not null,
     ID_PARAM char(10) not null,
     primary key (ID_ANA));

create table CAMPAIGN (
     ID_CMP char(10) not null,
     ID-Campaign numeric(10) not null,
     Name char(50) not null,
     Start date date not null,
     End date date not null,
     Port of departure char(50) not null,
     Port of arrival char(50) not null,
     Objectives char(255) not null,
     Area description char(1000) not null,
     Madsen square char(3) not null,
     IHB area char(50) not null,
     ID_PLAT char(10) not null,
     ID_PER char(10) not null,
     primary key (ID_CMP));

create table CATEGORY (
     ID_CAT char(10) not null,
     ID-Category numeric(10) not null,
     Name char(50) not null,
     Groups char(50) not null,
     Code char(10) not null,
     primary key (ID_CAT));

create table CITY (
     ID_CIT char(10) not null,
     ID-City numeric(10) not null,
     City name char(50) not null,
     Zip code char(10) not null,
     primary key (ID_CIT));

create table CMP belongs to PRJ (
     ID_CMP char(10) not null,
     ID_PRJ char(10) not null,
     primary key (ID_PRJ, ID_CMP));

create table COUNTRY (
     ID_COU char(10) not null,
     ID-ISO Country code numeric(10) not null,
     Country name char(50) not null,
     ISO A3 char(3) not null,
     ISO A2 char(2) not null,
     IOC code char(2) not null,
     ICES filename extension char(2) not null,
     primary key (ID_COU));

create table INSTITUTE (
     ID_INS char(10) not null,
     ID-Institute numeric(10) not null,
     Institute name char(100) not null,
     Institute code char(50) not null,
     primary key (ID_INS));

create table NON-CONTINUOUS VALUE (
     ID_SSP char(10) not null,
     ID-Non-continuous value numeric(10) not null,
     Value mantissa numeric(15) not null,
     Value exponent char(5) not null,
     Latitude float(1) not null,
     Longitude float(1) not null,
     Date date not null,
     Time date not null,
     Sampling depth mantissa numeric(15) not null,
     Sampling depth exponent numeric(5) not null,
     Replicate number numeric(3) not null,
     Quality flag char(25) not null,
     Validity flag char(25) not null,
     Ecosystem code char(25) not null,
     QC scheme number numeric(15) not null,
     Classification chart char(35) not null,
     A posteriori probability float(1) not null,
     Backtracking info char(30) not null,
     ID_PARAM char(10) not null,
     ID_STA char(10),
     ID_ANA char(10) not null,
     ID_SRV char(10) not null,
     primary key (ID_SSP));

create table PARAMETER (
     ID_PARAM char(10) not null,
     ID-Parameter numeric(10) not null,
     Name char(50) not null,
     ICES parameter code char(5) not null,
     Measurement unit char(50) not null,
     Matrix char(50) not null,
     Substrate char(50) not null,
     Formula char(50) not null,
     CAS char(50) not null,
     ID_CAT char(10) not null,
     primary key (ID_PARAM));

create table PER collects during CMP (
     ID_CMP char(10) not null,
     ID_PER char(10) not null,
     primary key (ID_PER, ID_CMP));

create table PER involved in PRJ (
     ID_PER char(10) not null,
     ID_PRJ char(10) not null,
     primary key (ID_PER, ID_PRJ));

create table PER works at SRV (
     ID_PER char(10) not null,
     ID_SRV char(10) not null,
     primary key (ID_PER, ID_SRV));

create table PERSON (
     ID_PER char(10) not null,
     ID-Person numeric(10) not null,
     First name char(15) not null,
     Last name char(15) not null,
     Position char(25) not null,
     Personal phone char(25) not null,
     Personal email char(80) not null,
     primary key (ID_PER));

create table PLATFORM (
     ID_PLAT char(10) not null,
     ID-Platform numeric(10) not null,
     Name char(50) not null,
     Type char(30) not null,
     Description char(1000) not null,
     Call sign char(50) not null,
     IOC ship codes char(50) not null,
     Platform picture char(0) not null,
     ID_SRV char(10) not null,
     primary key (ID_PLAT));

create table PROJECT (
     ID_PRJ char(10) not null,
     ID-Project numeric(10) not null,
     Project name char(255) not null,
     Project acronym char(15) not null,
     Start date date not null,
     End date date not null,
     Theme char(50) not null,
     Keywords char(50) not null,
     Abstract char(1000) not null,
     Study area char(1000) not null,
     Monitoring objective char(1000) not null,
     Website char(255) not null,
     References char(1000) not null,
     ID_PER char(10) not null,
     primary key (ID_PRJ));

create table QUALITY ANALYSIS (
     ID_QUA char(10) not null,
     QUALITY ANALYSIS : CTRL CHART char(10),
     QUALITY ANALYSIS : INTER-CALIB char(10),
     QUALITY ANALYSIS : QUASIMEME char(10),
     primary key (ID_QUA));

create table QUALITY ANALYSIS : CTRL CHART (
     ID_QUA char(10) not null,
     ID-Control chart numeric(10) not null,
     Control chart basis char(25) not null,
     CRM code char(15) not null,
     Certified concentration float(1) not null,
     Description char(1000) not null,
     Control chart mean value float(1) not null,
     Control chart stdv float(1) not null,
     Number of measurements numeric(3) not null,
     Period numeric(3) not null,
     Start date date not null,
     primary key (ID_QUA));

create table QUALITY ANALYSIS : INTER-CALIB (
     ID_QUA char(10) not null,
     ID-Inter-calibration numeric(10) not null,
     Inter-calibration exercise ref char(25) not null,
     Inter-calibration basis char(25) not null,
     Start date date not null,
     primary key (ID_QUA));

create table QUALITY ANALYSIS : QUASIMEME (
     ID_QUA char(10) not null,
     ID-Quasimeme numeric(10) not null,
     Inter-comparison exercise code char(25) not null,
     Assigned value float(1) not null,
     Robust mean float(1) not null,
     Z score float(1) not null,
     P score float(1) not null,
     Inter-calibration basis char(50) not null,
     Start date date not null,
     End date date not null,
     primary key (ID_QUA));

create table QUALITY SAMPLE & HANDLING (
     ID_QSH char(10) not null,
     ID-Quality sample & handling numeric(10) not null,
     Quality sample & handling ref. char(1000) not null,
     primary key (ID_QSH));

create table SAMPLE (
     ID_SP char(10) not null,
     ID-Sample numeric(10) not null,
     Monitoring year numeric(4) not null,
     Sequence number numeric(5) not null,
     Sample code numeric(6) not null,
     Water depth float(1) not null,
     Meta wind speed float(1) not null,
     Meta wind direction float(1) not null,
     Meta sea state numeric(2) not null,
     Meta atmospheric float(1) not null,
     Meta air temperature float(1) not null,
     ID_SPM char(10) not null,
     ID_PRJ char(10) not null,
     ID_CMP char(10) not null,
     primary key (ID_SP));

create table SAMPLE HANDLING (
     ID_SSPM char(10) not null,
     ID-Sample handling numeric(10) not null,
     Sample preservation char(50) not null,
     Sample pre-treatment char(50) not null,
     Sample separation char(50) not null,
     Procedure description char(1000) not null,
     primary key (ID_SSPM));

create table SAMPLING METHOD (
     ID_SPM char(10) not null,
     ID-Sampling method numeric(10) not null,
     Sampler char(100) not null,
     Sampler deployment char(100) not null,
     Description char(1000) not null,
     Sampler example char(0) not null,
     Sampler handling exemple char(0) not null,
     primary key (ID_SPM));

create table SERVICE (
     ID_SRV char(10) not null,
     ID-Service numeric(10) not null,
     Service name char(100) not null,
     Address char(80) not null,
     Phone char(25) not null,
     Fax char(25) not null,
     Email char(80) not null,
     Service code char(10) not null,
     ICES code char(10) not null,
     Description char(1000) not null,
     Website char(255) not null,
     ID_COU char(10) not null,
     ID_CIT char(10) not null,
     ID_INS char(10) not null,
     primary key (ID_SRV));

create table SRV belongs to PRJ (
     ID_PRJ char(10) not null,
     ID_SRV char(10) not null,
     primary key (ID_SRV, ID_PRJ));

create table SRV participates to CMP (
     ID_CMP char(10) not null,
     ID_SRV char(10) not null,
     primary key (ID_SRV, ID_CMP));

create table SRV-ANM has QAN (
     ID_QUA char(10) not null,
     ID_SRV char(10) not null,
     ID_PARAM char(10) not null,
     ID_ANA char(10) not null,
     primary key (ID_QUA));
```

```
create table SRV-SPM-SSPH has QSH (
     ID_QSH char(10) not null,
     ID_SRV char(10) not null,
     ID_SPM char(10) not null,
     ID_SSPM char(10) not null,
     ID_PARAM char(10) not null,
     primary key (ID_QSH));

create table STATION (
     ID_STA char(10) not null,
     ID-Station numeric(10) not null,
     Name char(10) not null,
     Start date date not null,
     End date date not null,
     Reference latitude float(1) not null,
     Reference longitude float(1) not null,
     primary key (ID_STA));

create table SUBSAMPLE (
     ID_SSP char(10) not null,
     ID-Subsample numeric(10) not null,
     Monitoring year numeric(4) not null,
     Sequence number numeric(5) not null,
     Sample code numeric(5) not null,
     Subsample code numeric(5) not null,
     ID_SRV char(10),
     ID_SSPM char(10) not null,
     ID_SP char(10) not null,
     primary key (ID_SSP));


-- Constraints Section
-- ___________________

alter table ANALYSIS METHOD add constraint FKANM linked to PARAM
     foreign key (ID_PARAM)
     references PARAMETER;

--alter table CAMPAIGN add constraint
--     check(exists(select * from CMP belongs to PRJ
--                  where CMP belongs to PRJ.ID_CMP = ID_CMP));

--alter table CAMPAIGN add constraint
--     check(exists(select * from PER collects during CMP
--                  where PER collects during CMP.ID_CMP = ID_CMP));

--alter table CAMPAIGN add constraint
--     check(exists(select * from SRV participates to CMP
--                  where SRV participates to CMP.ID_CMP = ID_CMP));

alter table CAMPAIGN add constraint FKPLAT used for CMP
     foreign key (ID_PLAT)
     references PLATFORM;

alter table CAMPAIGN add constraint FKCMP supervised by PER
     foreign key (ID_PER)
     references PERSON;

alter table CMP belongs to PRJ add constraint FKCMP_PRO
     foreign key (ID_PRJ)
     references PROJECT;
```

```
alter table CMP belongs to PRJ add constraint FKCMP_CAM
     foreign key (ID_CMP)
     references CAMPAIGN;

alter table NON-CONTINUOUS VALUE add constraint FKVAL is of type PARAM
     foreign key (ID_PARAM)
     references PARAMETER;

alter table NON-CONTINUOUS VALUE add constraint FKVAL collected at STA
     foreign key (ID_STA)
     references STATION;

alter table NON-CONTINUOUS VALUE add constraint FKVAL anlysed by ANM
     foreign key (ID_ANA)
     references ANALYSIS METHOD;

alter table NON-CONTINUOUS VALUE add constraint FKVAL analysed by SRV
     foreign key (ID_SRV)
     references SERVICE;

alter table NON-CONTINUOUS VALUE add constraint FKSSP gives VAL
     foreign key (ID_SSP)
     references SUBSAMPLE;

--alter table PARAMETER add constraint
--     check(exists(select * from ANALYSIS METHOD
--                  where ANALYSIS METHOD.ID_PARAM = ID_PARAM));

alter table PARAMETER add constraint FKPAR is of type CAT
     foreign key (ID_CAT)
     references CATEGORY;

alter table PER collects during CMP add constraint FKPER_PER_2
     foreign key (ID_PER)
     references PERSON;

alter table PER collects during CMP add constraint FKPER_CAM
     foreign key (ID_CMP)
     references CAMPAIGN;

alter table PER involved in PRJ add constraint FKPER_PRO
     foreign key (ID_PRJ)
     references PROJECT;

alter table PER involved in PRJ add constraint FKPER_PER_1
     foreign key (ID_PER)
     references PERSON;

alter table PER works at SRV add constraint FKPER_SER
     foreign key (ID_SRV)
     references SERVICE;

alter table PER works at SRV add constraint FKPER_PER
     foreign key (ID_PER)
     references PERSON;

alter table PLATFORM add constraint FKPLAT supervised by SRV
     foreign key (ID_SRV)
     references SERVICE;
```

```
alter table PROJECT add constraint FKPRJ supervised by PER
     foreign key (ID_PER)
     references PERSON;

--alter table QUALITY ANALYSIS add constraint
--     check(exists(select * from SRV-ANM has QAN
--                  where SRV-ANM has QAN.ID_QUA = ID_QUA));

--alter table QUALITY ANALYSIS add constraint ISAQUALITY ANALYSIS
--     check(QUALITY ANALYSIS : QUASIMEME is not null or QUALITY ANALYSIS :
INTER-CALIB is not null or QUALITY ANALYSIS : CTRL CHART is not null);

alter table QUALITY ANALYSIS : CTRL CHART add constraint FKQUA_QUA_2
     foreign key (ID_QUA)
     references QUALITY ANALYSIS;

alter table QUALITY ANALYSIS : INTER-CALIB add constraint FKQUA_QUA_1
     foreign key (ID_QUA)
     references QUALITY ANALYSIS;

alter table QUALITY ANALYSIS : QUASIMEME add constraint FKQUA_QUA
     foreign key (ID_QUA)
     references QUALITY ANALYSIS;

--alter table QUALITY SAMPLE & HANDLING add constraint
--     check(exists(select * from SRV-SPM-SSPH has QSH
--                  where SRV-SPM-SSPH has QSH.ID_QSH = ID_QSH));

--alter table SAMPLE add constraint
--     check(exists(select * from SUBSAMPLE
--                  where SUBSAMPLE.ID_SP = ID_SP));

alter table SAMPLE add constraint FKSP by SPM
     foreign key (ID_SPM)
     references SAMPLING METHOD;

alter table SAMPLE add constraint FKPRJ collects SP
     foreign key (ID_PRJ)
     references PROJECT;

alter table SAMPLE add constraint FKCMP collects SP
     foreign key (ID_CMP)
     references CAMPAIGN;

--alter table SERVICE add constraint
--     check(exists(select * from PER works at SRV
--                  where PER works at SRV.ID_SRV = ID_SRV));

alter table SERVICE add constraint FKSRV is located in COUNTRY
     foreign key (ID_COU)
     references COUNTRY;

alter table SERVICE add constraint FKSRV is located at CITY
     foreign key (ID_CIT)
     references CITY;

alter table SERVICE add constraint FKSRV belongs to INST
     foreign key (ID_INS)
     references INSTITUTE;

alter table SRV belongs to PRJ add constraint FKSRV_SER_3
```

```
     foreign key (ID_SRV)
     references SERVICE;

alter table SRV belongs to PRJ add constraint FKSRV_PRO
     foreign key (ID_PRJ)
     references PROJECT;

alter table SRV participates to CMP add constraint FKSRV_SER_2
     foreign key (ID_SRV)
     references SERVICE;

alter table SRV participates to CMP add constraint FKSRV_CAM
     foreign key (ID_CMP)
     references CAMPAIGN;

alter table SRV-ANM has QAN add constraint FKSRV_SER_1
     foreign key (ID_SRV)
     references SERVICE;

alter table SRV-ANM has QAN add constraint FKSRV_QUA_1
     foreign key (ID_QUA)
     references QUALITY ANALYSIS;

alter table SRV-ANM has QAN add constraint FKSRV_PAR_1
     foreign key (ID_PARAM)
     references PARAMETER;

alter table SRV-ANM has QAN add constraint FKSRV_ANA
     foreign key (ID_ANA)
     references ANALYSIS METHOD;

alter table SRV-SPM-SSPH has QSH add constraint FKSRV_SER
     foreign key (ID_SRV)
     references SERVICE;

alter table SRV-SPM-SSPH has QSH add constraint FKSRV_SAM_1
     foreign key (ID_SPM)
     references SAMPLING METHOD;

alter table SRV-SPM-SSPH has QSH add constraint FKSRV_SAM
     foreign key (ID_SSPM)
     references SAMPLE HANDLING;

alter table SRV-SPM-SSPH has QSH add constraint FKSRV_QUA
     foreign key (ID_QSH)
     references QUALITY SAMPLE & HANDLING;

alter table SRV-SPM-SSPH has QSH add constraint FKSRV_PAR
     foreign key (ID_PARAM)
     references PARAMETER;

alter table SUBSAMPLE add constraint FKSSP kept by SRV
     foreign key (ID_SRV)
     references SERVICE;

alter table SUBSAMPLE add constraint FKSSP by SSPH
     foreign key (ID_SSPM)
     references SAMPLE HANDLING;

alter table SUBSAMPLE add constraint FKSP split into SSP
     foreign key (ID_SP)
     references SAMPLE;


-- Index Section
-- _____________

create unique index ID
     on ANALYSIS METHOD (ID_ANA);

create index FKANM linked to PARAM
     on ANALYSIS METHOD (ID_PARAM);

create unique index ID
     on CAMPAIGN (ID_CMP);

create index FKPLAT used for CMP
     on CAMPAIGN (ID_PLAT);

create index FKCMP supervised by PER
     on CAMPAIGN (ID_PER);

create unique index ID
     on CATEGORY (ID_CAT);

create unique index ID
     on CITY (ID_CIT);

create unique index IDCMP belongs to PRJ
     on CMP belongs to PRJ (ID_PRJ, ID_CMP);

create index FKCMP_PRO
     on CMP belongs to PRJ (ID_PRJ);

create index FKCMP_CAM
     on CMP belongs to PRJ (ID_CMP);

create unique index ID
     on COUNTRY (ID_COU);

create unique index ID
     on INSTITUTE (ID_INS);

create index FKVAL is of type PARAM
     on NON-CONTINUOUS VALUE (ID_PARAM);

create index FKVAL collected at STA
     on NON-CONTINUOUS VALUE (ID_STA);

create index FKVAL anlysed by ANM
     on NON-CONTINUOUS VALUE (ID_ANA);

create index FKVAL analysed by SRV
     on NON-CONTINUOUS VALUE (ID_SRV);

create unique index FKSSP gives VAL
     on NON-CONTINUOUS VALUE (ID_SSP);

create unique index ID
     on PARAMETER (ID_PARAM);

create index FKPAR is of type CAT
```

```
     on PARAMETER (ID_CAT);

create unique index IDPER collects during CMP
     on PER collects during CMP (ID_PER, ID_CMP);

create index FKPER_PER_2
     on PER collects during CMP (ID_PER);

create index FKPER_CAM
     on PER collects during CMP (ID_CMP);

create unique index IDPER involved in PRJ
     on PER involved in PRJ (ID_PER, ID_PRJ);

create index FKPER_PRO
     on PER involved in PRJ (ID_PRJ);

create index FKPER_PER_1
     on PER involved in PRJ (ID_PER);

create unique index IDPER works at SRV
     on PER works at SRV (ID_PER, ID_SRV);

create index FKPER_SER
     on PER works at SRV (ID_SRV);

create index FKPER_PER
     on PER works at SRV (ID_PER);

create unique index ID
     on PERSON (ID_PER);

create unique index ID
     on PLATFORM (ID_PLAT);

create index FKPLAT supervised by SRV
     on PLATFORM (ID_SRV);

create unique index ID
     on PROJECT (ID_PRJ);

create index FKPRJ supervised by PER
     on PROJECT (ID_PER);

create unique index ID
     on QUALITY ANALYSIS (ID_QUA);

create unique index FKQUA_QUA_2
     on QUALITY ANALYSIS : CTRL CHART (ID_QUA);

create unique index FKQUA_QUA_1
     on QUALITY ANALYSIS : INTER-CALIB (ID_QUA);

create unique index FKQUA_QUA
     on QUALITY ANALYSIS : QUASIMEME (ID_QUA);

create unique index ID
     on QUALITY SAMPLE & HANDLING (ID_QSH);

create unique index ID
     on SAMPLE (ID_SP);
```

```
create index FKSP by SPM
     on SAMPLE (ID_SPM);

create index FKPRJ collects SP
     on SAMPLE (ID_PRJ);

create index FKCMP collects SP
     on SAMPLE (ID_CMP);

create unique index ID
     on SAMPLE HANDLING (ID_SSPM);

create unique index ID
     on SAMPLING METHOD (ID_SPM);

create unique index ID
     on SERVICE (ID_SRV);

create index FKSRV is located in COUNTRY
     on SERVICE (ID_COU);

create index FKSRV is located at CITY
     on SERVICE (ID_CIT);

create index FKSRV belongs to INST
     on SERVICE (ID_INS);

create unique index IDSRV belongs to PRJ
     on SRV belongs to PRJ (ID_SRV, ID_PRJ);

create index FKSRV_SER_3
     on SRV belongs to PRJ (ID_SRV);

create index FKSRV_PRO
     on SRV belongs to PRJ (ID_PRJ);

create unique index IDSRV participates to CMP
     on SRV participates to CMP (ID_SRV, ID_CMP);

create index FKSRV_SER_2
     on SRV participates to CMP (ID_SRV);

create index FKSRV_CAM
     on SRV participates to CMP (ID_CMP);

create index FKSRV_SER_1
     on SRV-ANM has QAN (ID_SRV);

create unique index FKSRV_QUA_1
     on SRV-ANM has QAN (ID_QUA);

create index FKSRV_PAR_1
     on SRV-ANM has QAN (ID_PARAM);

create index FKSRV_ANA
     on SRV-ANM has QAN (ID_ANA);

create index FKSRV_SER
     on SRV-SPM-SSPH has QSH (ID_SRV);
```

```
create index FKSRV_SAM_1
     on SRV-SPM-SSPH has QSH (ID_SPM);

create index FKSRV_SAM
     on SRV-SPM-SSPH has QSH (ID_SSPM);

create unique index FKSRV_QUA
     on SRV-SPM-SSPH has QSH (ID_QSH);

create index FKSRV_PAR
     on SRV-SPM-SSPH has QSH (ID_PARAM);

create unique index ID
     on STATION (ID_STA);

create unique index ID
     on SUBSAMPLE (ID_SSP);

create index FKSSP kept by SRV
     on SUBSAMPLE (ID_SRV);

create index FKSSP by SSPH
     on SUBSAMPLE (ID_SSPM);

create index FKSP split into SSP
     on SUBSAMPLE (ID_SP);
```

Annex C

All campaigns with associated services and platforms.

Campaign

Campaign.Name *BE96/1*

Service code	Start date	End date	Objectives	Platform.Name
BMMOST	30/01/96	2/02/96	Monitoring	Belgica
GMMA	30/01/96	2/02/96	Monitoring	Belgica

Campaign.Name *BE96/10*

Service code	Start date	End date	Objectives	Platform.Name
ANCH	2/05/96	10/05/96	Geology	Belgica
GMMA	2/05/96	10/05/96	Geology	Belgica

Campaign.Name *BE96/11*

Service code	Start date	End date	Objectives	Platform.Name
GMMA	13/05/96	15/05/96	Monitoring	Belgica

Campaign.Name *BE96/12*

Service code	Start date	End date	Objectives	Platform.Name
GMMA	20/05/96	24/05/96	Monitoring	Belgica

Campaign.Name *BE96/13*

Service code	Start date	End date	Objectives	Platform.Name
GMMA	28/05/96	31/05/96	Geology	Belgica

Campaign.Name *BE96/14*

Service code	***Start date***	***End date***	***Objectives***	***Platform.Name***
ANCH	3/06/96	14/06/96	Fishery	Belgica
BMMOST	3/06/96	14/06/96	Fishery	Belgica

Campaign.Name *BE96/15*

Service code	***Start date***	***End date***	***Objectives***	***Platform.Name***
MUMM	18/06/96	1/07/96	Geology	Belgica

Campaign.Name *BE96/16*

Service code	***Start date***	***End date***	***Objectives***	***Platform.Name***
MUMM	3/07/96	5/07/96	Environment	Belgica

Campaign.Name *BE96/17*

Service code	***Start date***	***End date***	***Objectives***	***Platform.Name***
GMMA	7/07/96	12/07/96	Ecosystem	Belgica

Campaign.Name *BE96/18*

Service code	***Start date***	***End date***	***Objectives***	***Platform.Name***
ANCH	15/07/96	17/07/96	Ecosystem	Belgica

Campaign.Name *BE96/19*

Service code	***Start date***	***End date***	***Objectives***	***Platform.Name***
ANCH	21/08/96	29/08/96	Fishery	Belgica
BMMOST	21/08/96	29/08/96	Fishery	Belgica

GMMA	21/08/96	29/08/96	Fishery	Belgica

Campaign.Name *BE96/2*

Service code	***Start date***	***End date***	***Objectives***	***Platform.Name***
MUMM	6/02/96	9/02/96	Geology	Belgica

Campaign.Name *BE96/20*

Service code	***Start date***	***End date***	***Objectives***	***Platform.Name***
GMMA	2/09/96	20/09/96	Ecosystem	Belgica
MUMM	2/09/96	20/09/96	Ecosystem	Belgica

Campaign.Name *BE96/21*

Service code	***Start date***	***End date***	***Objectives***	***Platform.Name***
ANCH	23/09/96	26/09/96	Geology	Belgica

Campaign.Name *BE96/22a*

Service code	***Start date***	***End date***	***Objectives***	***Platform.Name***
ANCH	30/09/96	4/10/96	Monitoring	Belgica

Campaign.Name *BE96/22b*

Service code	***Start date***	***End date***	***Objectives***	***Platform.Name***
MUMM	7/10/96	10/10/96	Monitoring	Belgica

Campaign.Name *BE96/23*

Service code	***Start date***	***End date***	***Objectives***	***Platform.Name***
ANCH	14/10/96	24/10/96	Ecosystem	Belgica

mardi 15 décembre 1998 ***Page 3 sur 6***

BMMOST	14/10/96	24/10/96	Ecosystem	Belgica
GMMA	14/10/96	24/10/96	Ecosystem	Belgica

Campaign.Name *BE96/24*

Service code	***Start date***	***End date***	***Objectives***	***Platform.Name***
GMMA	28/10/96	31/10/96	Geology	Belgica

Campaign.Name *BE96/25*

Service code	***Start date***	***End date***	***Objectives***	***Platform.Name***
GMMA	4/11/96	8/11/96	Fishery	Belgica

Campaign.Name *BE96/26*

Service code	***Start date***	***End date***	***Objectives***	***Platform.Name***
ANCH	12/11/96	14/11/96	Ecosystem	Belgica
BMMOST	12/11/96	14/11/96	Ecosystem	Belgica

Campaign.Name *BE96/27*

Service code	***Start date***	***End date***	***Objectives***	***Platform.Name***
MUMM	18/11/96	22/11/96	Monitoring	Belgica

Campaign.Name *BE96/28*

Service code	***Start date***	***End date***	***Objectives***	***Platform.Name***
BMMOST	25/11/96	28/11/96	Geology	Belgica

Campaign.Name *BE96/29*

Service code	***Start date***	***End date***	***Objectives***	***Platform.Name***

GMMA	3/12/96	6/12/96	Monitoring	Belgica

Campaign.Name *BE96/3*

Service code	***Start date***	***End date***	***Objectives***	***Platform.Name***
ANCH	13/02/96	16/02/96	Monitoring	Belgica
MUMM	13/02/96	16/02/96	Monitoring	Belgica

Campaign.Name *BE96/30*

Service code	***Start date***	***End date***	***Objectives***	***Platform.Name***
ANCH	9/12/96	13/12/96	Ecosystem	Belgica

Campaign.Name *BE96/31*

Service code	***Start date***	***End date***	***Objectives***	***Platform.Name***
ANCH	16/12/96	19/12/96	Geology	Belgica
MUMM	16/12/96	19/12/96	Geology	Belgica

Campaign.Name *BE96/4*

Service code	***Start date***	***End date***	***Objectives***	***Platform.Name***
MUMM	19/02/96	29/02/96	Ecosystem	Belgica

Campaign.Name *BE96/5*

Service code	***Start date***	***End date***	***Objectives***	***Platform.Name***
ANCH	4/03/96	8/03/96	Fishery	Belgica
BMMOST	4/03/96	8/03/96	Fishery	Belgica

Campaign.Name *BE96/6*

Service code	***Start date***	***End date***	***Objectives***	***Platform.Name***
MUMM	11/03/96	15/03/96	Geology	Belgica

Campaign.Name *BE96/7*

Service code	***Start date***	***End date***	***Objectives***	***Platform.Name***
BMMOST	18/03/96	29/03/96	Monitoring	Belgica

Campaign.Name *BE96/8*

Service code	***Start date***	***End date***	***Objectives***	***Platform.Name***
BMMOST	15/04/96	26/04/96	Ecosystem	Belgica

Campaign.Name *BE96/9*

Service code	***Start date***	***End date***	***Objectives***	***Platform.Name***
ANCH	29/04/96	30/04/96	Monitoring	Belgica
BMMOST	29/04/96	30/04/96	Monitoring	Belgica
GMMA	29/04/96	30/04/96	Monitoring	Belgica
MUMM	29/04/96	30/04/96	Monitoring	Belgica

Annex D

List all stations with the total number of corresponding visits classified by campaigns.

Name	*BE96/1*	*BE96/12*	*BE96/29*	*BE96/3*
			52	
105	13	13	13	
115	13	13	13	
120	13	13	13	
130	13	13	13	
150	13	13	13	
215	13	13	13	
230	13	13	13	
250	13	13	13	
315	13	13	13	
330	13	13	13	
421	13	13		
435	13	13		
545	13	13		
700	13	13	13	
710	13	13	13	
800	13	13		
S01	13	13	13	13
S04	13	13	13	13
S07	13	13	13	13
S09	13	13	13	13
S12	13	13		13
S15	13	13		13
S15b	13			13
S18	13	13		13
S18b				13
S20				13
S22	13	13	13	13
ZG01	13	13	13	
ZG02	13	13	13	
ZG03	13	13	13	

Annex E

List all values with associated station, campaign and project.

Date	Time	Station.Name	Value mantissa	Value exponent -	Campaign.Name	Project name
20/05/96	9:13	700	3	1	BE96/12	Monitoring of the quality of the marine environment
20/05/96	9:13	700	5	-6	BE96/12	Monitoring of the quality of the marine environment
20/05/96	9:13	700	8	-7	BE96/12	Monitoring of the quality of the marine environment
20/05/96	9:13	700	3	1	BE96/12	Monitoring of the quality of the marine environment
20/05/96	9:13	700	3	-5	BE96/12	Monitoring of the quality of the marine environment
20/05/96	14:23	250	3	1	BE96/12	Monitoring of the quality of the marine environment
20/05/96	14:23	250	3	1	BE96/12	Monitoring of the quality of the marine environment
20/05/96	14:23	250	4	-7	BE96/12	Monitoring of the quality of the marine environment
20/05/96	14:23	250	2	-5	BE96/12	Monitoring of the quality of the marine environment
20/05/96	14:23	250	4	-6	BE96/12	Monitoring of the quality of the marine environment
20/05/96	17:27	545	3	1	BE96/12	Monitoring of the quality of the marine environment
20/05/96	17:27	545	3	1	BE96/12	Monitoring of the quality of the marine environment
20/05/96	17:27	545	2	-7	BE96/12	Monitoring of the quality of the marine environment
20/05/96	17:27	545	7	-6	BE96/12	Monitoring of the quality of the marine environment
20/05/96	17:27	545	2	-6	BE96/12	Monitoring of the quality of the marine environment
20/05/96	18:46	800	1	-7	BE96/12	Monitoring of the quality of the marine environment
20/05/96	18:46	800	5	-8	BE96/12	Monitoring of the quality of the marine environment
20/05/96	18:46	800	3	1	BE96/12	Monitoring of the quality of the marine environment

Date	Time	Station.Name	Value mantissa	Value exponent -	Campaign.Name	Project name
20/05/96	18:46	800	3	1	BE96/12	Monitoring of the quality of the marine environment
20/05/96	18:46	800	4	-7	BE96/12	Monitoring of the quality of the marine environment
21/05/96	4:48	435	2	-6	BE96/12	Monitoring of the quality of the marine environment
21/05/96	4:48	435	3	1	BE96/12	Monitoring of the quality of the marine environment
21/05/96	4:48	435	3	1	BE96/12	Monitoring of the quality of the marine environment
21/05/96	4:48	435	2	-7	BE96/12	Monitoring of the quality of the marine environment
21/05/96	4:48	435	4	-6	BE96/12	Monitoring of the quality of the marine environment
21/05/96	8:16	421	3	1	BE96/12	Monitoring of the quality of the marine environment
21/05/96	8:16	421	3	1	BE96/12	Monitoring of the quality of the marine environment
21/05/96	8:16	421	1	-7	BE96/12	Monitoring of the quality of the marine environment
21/05/96	8:16	421	4	-7	BE96/12	Monitoring of the quality of the marine environment
21/05/96	8:16	421	4	-7	BE96/12	Monitoring of the quality of the marine environment
21/05/96	9:32	315	5	-6	BE96/12	Monitoring of the quality of the marine environment
21/05/96	9:32	315	2	-6	BE96/12	Monitoring of the quality of the marine environment
21/05/96	9:32	315	2	-7	BE96/12	Monitoring of the quality of the marine environment
21/05/96	9:32	315	3	1	BE96/12	Monitoring of the quality of the marine environment
21/05/96	9:32	315	3	1	BE96/12	Monitoring of the quality of the marine environment
21/05/96	10:59	ZG02	3	1	BE96/12	Monitoring of the quality of the marine environment
21/05/96	10:59	ZG02	1	-6	BE96/12	Monitoring of the quality of the marine environment
21/05/96	10:59	ZG02	9	-6	BE96/12	Monitoring of the quality of the marine environment
21/05/96	10:59	ZG02	3	-7	BE96/12	Monitoring of the quality of the marine environment

Date	Time	Station.Name	Value mantissa	Value exponent -	Campaign.Name	Project name
21/05/96	10:59	ZG02	3	1	BE96/12	Monitoring of the quality of the marine environment
21/05/96	12:14	105	3	1	BE96/12	Monitoring of the quality of the marine environment
21/05/96	12:14	105	3	1	BE96/12	Monitoring of the quality of the marine environment
21/05/96	12:14	105	4	-7	BE96/12	Monitoring of the quality of the marine environment
21/05/96	12:14	105	9	-6	BE96/12	Monitoring of the quality of the marine environment
21/05/96	12:14	105	1	-6	BE96/12	Monitoring of the quality of the marine environment
21/05/96	13:21	215	8	-7	BE96/12	Monitoring of the quality of the marine environment
21/05/96	13:21	215	3	1	BE96/12	Monitoring of the quality of the marine environment
21/05/96	13:21	215	3	1	BE96/12	Monitoring of the quality of the marine environment
21/05/96	13:21	215	1	-5	BE96/12	Monitoring of the quality of the marine environment
21/05/96	13:21	215	5	-7	BE96/12	Monitoring of the quality of the marine environment
21/05/96	14:56	ZG01	3	1	BE96/12	Monitoring of the quality of the marine environment
21/05/96	14:56	ZG01	3	1	BE96/12	Monitoring of the quality of the marine environment
21/05/96	14:56	ZG01	4	-7	BE96/12	Monitoring of the quality of the marine environment
21/05/96	14:56	ZG01	1	-5	BE96/12	Monitoring of the quality of the marine environment
21/05/96	14:56	ZG01	2	-6	BE96/12	Monitoring of the quality of the marine environment
21/05/96	15:54	ZG03	4	-7	BE96/12	Monitoring of the quality of the marine environment
21/05/96	15:54	ZG03	2	-6	BE96/12	Monitoring of the quality of the marine environment
21/05/96	15:54	ZG03	1	-5	BE96/12	Monitoring of the quality of the marine environment
21/05/96	15:54	ZG03	3	1	BE96/12	Monitoring of the quality of the marine environment
21/05/96	15:54	ZG03	3	1	BE96/12	Monitoring of the quality of the marine environment

Date	Time	Station.Name	Value mantissa	Value exponent -	Campaign.Name	Project name
21/05/96	16:44	230	3	-6	BE96/12	Monitoring of the quality of the marine environment
21/05/96	16:44	230	3	1	BE96/12	Monitoring of the quality of the marine environment
21/05/96	16:44	230	3	1	BE96/12	Monitoring of the quality of the marine environment
21/05/96	16:44	230	9	-7	BE96/12	Monitoring of the quality of the marine environment
21/05/96	16:44	230	2	-5	BE96/12	Monitoring of the quality of the marine environment
21/05/96	18:27	130	3	1	BE96/12	Monitoring of the quality of the marine environment
21/05/96	18:27	130	2	-6	BE96/12	Monitoring of the quality of the marine environment
21/05/96	18:27	130	3	-5	BE96/12	Monitoring of the quality of the marine environment
21/05/96	18:27	130	2	-6	BE96/12	Monitoring of the quality of the marine environment
21/05/96	18:27	130	3	1	BE96/12	Monitoring of the quality of the marine environment
21/05/96	19:47	120	3	1	BE96/12	Monitoring of the quality of the marine environment
21/05/96	19:47	120	2	-5	BE96/12	Monitoring of the quality of the marine environment
21/05/96	19:47	120	1	-6	BE96/12	Monitoring of the quality of the marine environment
21/05/96	19:47	120	3	1	BE96/12	Monitoring of the quality of the marine environment
21/05/96	19:47	120	2	-6	BE96/12	Monitoring of the quality of the marine environment
21/05/96	21:12	115	2	-5	BE96/12	Monitoring of the quality of the marine environment
21/05/96	21:12	115	9	-7	BE96/12	Monitoring of the quality of the marine environment
21/05/96	21:12	115	3	1	BE96/12	Monitoring of the quality of the marine environment
21/05/96	21:12	115			BE96/12	Monitoring of the quality of the marine environment
21/05/96	21:12	115	3	1	BE96/12	Monitoring of the quality of the marine environment
22/05/96	9:29	710	6	-7	BE96/12	Monitoring of the quality of the marine environment

Date	Time	Station.Name	Value mantissa	Value exponent -	Campaign.Name	Project name
22/05/96	9:29	710	4	-6	BE96/12	Monitoring of the quality of the marine environment
22/05/96	9:29	710	2	-5	BE96/12	Monitoring of the quality of the marine environment
22/05/96	9:29	710	3	1	BE96/12	Monitoring of the quality of the marine environment
22/05/96	9:29	710	3	1	BE96/12	Monitoring of the quality of the marine environment
22/05/96	14:28	330	8	-7	BE96/12	Monitoring of the quality of the marine environment
22/05/96	14:28	330	3	1	BE96/12	Monitoring of the quality of the marine environment
22/05/96	14:28	330	3	1	BE96/12	Monitoring of the quality of the marine environment
22/05/96	14:28	330	3	-7	BE96/12	Monitoring of the quality of the marine environment
22/05/96	14:28	330	1	-5	BE96/12	Monitoring of the quality of the marine environment
23/05/96	4:16	150	5	-6	BE96/12	Monitoring of the quality of the marine environment
23/05/96	4:16	150	7	-7	BE96/12	Monitoring of the quality of the marine environment
23/05/96	4:16	150	3	-5	BE96/12	Monitoring of the quality of the marine environment
23/05/96	4:16	150	3	1	BE96/12	Monitoring of the quality of the marine environment
23/05/96	4:16	150	3	1	BE96/12	Monitoring of the quality of the marine environment
23/05/96	5:50	S01	3	-6	BE96/12	Monitoring of the quality of the marine environment
23/05/96	5:50	S01	4	-5	BE96/12	Monitoring of the quality of the marine environment
23/05/96	5:50	S01	10	-7	BE96/12	Monitoring of the quality of the marine environment
23/05/96	5:50	S01	3	1	BE96/12	Monitoring of the quality of the marine environment
23/05/96	5:50	S01	3	1	BE96/12	Monitoring of the quality of the marine environment
23/05/96	8:15	S04	3	1	BE96/12	Monitoring of the quality of the marine environment
23/05/96	8:15	S04	3	1	BE96/12	Monitoring of the quality of the marine environment

Date	Time	Station.Name	Value mantissa	Value exponent -	Campaign.Name	Project name
23/05/96	8:15	S04	3	-6	BE96/12	Monitoring of the quality of the marine environment
23/05/96	8:15	S04	1	-4	BE96/12	Monitoring of the quality of the marine environment
23/05/96	8:15	S04	3	-6	BE96/12	Monitoring of the quality of the marine environment
23/05/96	9:28	S07	2	1	BE96/12	Monitoring of the quality of the marine environment
23/05/96	9:28	S07	3	-6	BE96/12	Monitoring of the quality of the marine environment
23/05/96	9:28	S07	2	-4	BE96/12	Monitoring of the quality of the marine environment
23/05/96	9:28	S07	6	-6	BE96/12	Monitoring of the quality of the marine environment
23/05/96	9:28	S07	2	1	BE96/12	Monitoring of the quality of the marine environment
23/05/96	10:15	S09	1	-5	BE96/12	Monitoring of the quality of the marine environment
23/05/96	10:15	S09	2	1	BE96/12	Monitoring of the quality of the marine environment
23/05/96	10:15	S09	2	1	BE96/12	Monitoring of the quality of the marine environment
23/05/96	10:15	S09	2	-4	BE96/12	Monitoring of the quality of the marine environment
23/05/96	10:15	S09	4	-6	BE96/12	Monitoring of the quality of the marine environment
23/05/96	11:20	S12	2	-4	BE96/12	Monitoring of the quality of the marine environment
23/05/96	11:20	S12	1	-5	BE96/12	Monitoring of the quality of the marine environment
23/05/96	11:20	S12	5	-6	BE96/12	Monitoring of the quality of the marine environment
23/05/96	11:20	S12	1	1	BE96/12	Monitoring of the quality of the marine environment
23/05/96	11:20	S12	1	1	BE96/12	Monitoring of the quality of the marine environment
23/05/96	11:59	S15	4	-6	BE96/12	Monitoring of the quality of the marine environment
23/05/96	11:59	S15	3	-4	BE96/12	Monitoring of the quality of the marine environment
23/05/96	11:59	S15	2	-5	BE96/12	Monitoring of the quality of the marine environment

Date	Time	Station.Name	Value mantissa	Value exponent -	Campaign.Name	Project name
23/05/96	11:59	S15	1	1	BE96/12	Monitoring of the quality of the marine environment
23/05/96	11:59	S15	1	1	BE96/12	Monitoring of the quality of the marine environment
23/05/96	12:48	S18	9	-6	BE96/12	Monitoring of the quality of the marine environment
23/05/96	12:48	S18	3	-4	BE96/12	Monitoring of the quality of the marine environment
23/05/96	12:48	S18	2	-5	BE96/12	Monitoring of the quality of the marine environment
23/05/96	12:48	S18	6	0	BE96/12	Monitoring of the quality of the marine environment
23/05/96	12:48	S18	6	0	BE96/12	Monitoring of the quality of the marine environment
23/05/96	13:33	S22	1	0	BE96/12	Monitoring of the quality of the marine environment
23/05/96	13:33	S22	6	-6	BE96/12	Monitoring of the quality of the marine environment
23/05/96	13:33	S22	1	0	BE96/12	Monitoring of the quality of the marine environment
23/05/96	13:33	S22	1	-4	BE96/12	Monitoring of the quality of the marine environment
23/05/96	13:33	S22	2	-5	BE96/12	Monitoring of the quality of the marine environment

PRIME MINISTER'S SERVICES
Federal Office for Scientific, Technical and Cultural Affairs

SCIENTIFIC SUPPORT PLAN
FOR A SUSTAINABLE DEVELOPMENT POLICY

«SUSTAINABLE MANAGEMENT OF THE NORTH SEA»

RESEARCH CONTRACTS MN/DD/60, 61 & 62

INTEGRATED AND DYNAMICAL OCEANOGRAPHIC DATA MANAGEMENT

JOINT SCIENTIFIC REPORTS
for the year 1999

January 2000

The present document gathers the yearly scientific reports of the three partners to the *Integrated and Dynamical Oceanographic Data Management* project, performed on behalf of the Federal Office for Scientific, Technical and Cultural Affairs.

It covers the activities performed during the year 1999. It contains, in sequence :

- the scientific report of the Management Unit of the Mathematical Model of the North Sea (MUMM), and its annexes;
- the scientific report of the SURFACES laboratory (ULg);
- the scientific report of the *Universitair Centrum voor Statistiek* (UCS, KUL), and its appendices;
- three general annexes (IDOD Newsletter # 2, IDOD Newsletter # 3, paper submitted for publication).

January 2000

MANAGEMENT UNIT OF THE
NORTH SEA MATHEMATICAL MODELS

DATA MANAGEMENT GROUP

Integrated and Dynamical Oceanographic Data Management - IDOD

Karien De Cauwer, Mia Devolder, Siegrid Jans

DMG/0/KDC, MD, SJ/199901/EN/IR

Joint Scientific Report n°3 (January-December 1999)
Prepared for OSTC, contract MN/DD/60

MUMM
100 Gulledelle
B–1200 Brussels
Belgium

LIST OF ANNEXES

LIST OF FIGURES

1. Introduction

In 1999, the update of the inventory of data sets and the screening procedure used for the acceptation of data and data documentation that is performed before data can be entered in the database, were considered as important and ongoing tasks. After the screening of the data sets, the analysis of their structure was performed. These tasks concentrated on the data sets from the Programme “Sustainable Management of the North Sea”. A detailed description of the work is given in chapter 2.

The analysis of the structure of the data sets gave rise to an update of the conceptual scheme for the seawater database. Furthermore, they formed the basis for a first draft of the conceptual schemes for the plankton and sediment databases. A detailed overview of the schemes is available in chapter 3.

Chapter 4 illustrates the implementation of the scheme for the seawater database. The result is a prototype of database in which a query was developed to retrieve seawater data.

The task concerning spatial visualisation was also considered and resulted in the purchase of Arcview. Actually, the exploratory phase of the geographic information system is ongoing. Further details are included in chapter 5.

In order to clarify the conditions of data transfer and data access to the database, a convention between the data providers, the OSTC and the IDOD data centre was prepared (see chapter 6).

The products, such as the Newsletters and the IDOD web page, prepared in the frame of the IDOD project are summarised in chapter 7.

2. Data sets

In order to construct a database, an inventory of the data sets that should be stored has to be made. As the types of data are diverse, the data characteristics have to be studied to be able to construct and/or adapt a conceptual scheme. The analysis of the structure of data sets was an important and time-consuming work in 1999.

In 1999, we mainly focused on the data sets from the Programme ‘Sustainable Development of the North Sea’. Those data sets cover a wide range of disciplines, and form as such a good basis for the development of an integrated oceanographic database. As the projects are ongoing, it is easy to contact the data providers for more explanation on the characteristics and missing meta-information.

2.1. Inventory

Concerning the existing data sets, a list of data sets or data sources available at MUMM had been established in 1998. This list has not been changed, but new contacts are taken in order to obtain specific geographical data. Besides this, the European Directory of Marine Environmental Data (EDMED) is still being updated and extended.

2.1.1. European Directory of Marine Environmental Data (EDMED)

EDMED consists of two inventory forms : data sets and data holding centres. The inventory of Belgian data sets was last updated in 1995. In the frame of EURONODIM, an EC Concerted Action (http://www.sea-search.net), the inventory is being reviewed. The objective is to review and upgrade the content, the format and the internet functionality for entry and retrieval.

A number of revisions with respect to format were agreed on. Web links, forms for the entry of multiple data contacts per centre, extra query fields (e.g. geographic NSEW co-ordinates) and other fields are included.

The update of the content occurs in different phases. The first phase consisted in the review of names and addresses of data holding organisations and has been done by MUMM. The second phase consists in the update of data set descriptions of data centres already present, and the last phase will be the inventory of data sets of new data centres. The last two phases have not yet been started as the MSAccess input programme was not yet received.

2.1.2. Data collected by the programme « Sustainable Development of the North Sea »

A first basic inventory of the parameters measured by the laboratories in the frame of the Programme « Sustainable Development of the North Sea » was made in 1998 (IDOD, 1999). However, further contacts with the data providers have put in evidence changes in the availability of results for these parameters. This resulted in a new list of expected parameters and meta-data. The most recent inventory is joined in Annex 1.

The technical specifications of the research projects mention that the data sets have to be sent to the IDOD centre « at the end of March following the year when the corresponding samples were taken ». As decided at a coordination meeting, the data sets pertaining to the years 1997 and 1998 had to be submitted to the IDOD data centre at the end of March 99. The “flow rate” evaluated by mid-December is represented in Figure 1.

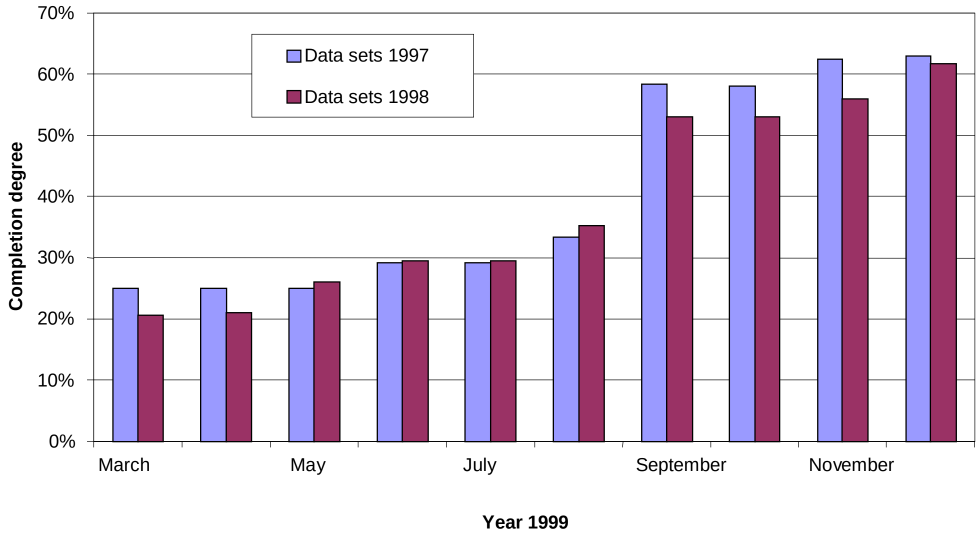


Figure 1. Flow rate of data sets to the IDOD centre during the year 1999 (12 data sets were expected for 1997 and 17 for 1998)

By the end of March 99, only about 25 % of the data sets of 1997 and 20 % of the data sets of 1998 were received by MUMM. At the end of 1999 less than 70 % of the data from 1997 and 1998 were received.

As shown in Figure 1, data sets were received with a delay of often several months and some data sets are still incomplete or missing. This delay in data transfer of course led to a delay in the set-up of the conceptual database scheme, but also induced an increasing difficulty for the reflection about the database structure itself.

As data sets were received, a more elaborated inventory could be made, including information on the results of the screening.

2.2. *Quality Control*

Quality control is essential for the credibility of a database. The control process takes place at different levels.

The first level is the <u>acceptation of data and data documentation</u>.
This step implies the verification of the data sets using the screening procedure:

1.a. Availability of the data set
- yes (date of delivery) / no

1.b. Availability of the meta-information:
- general (date, time, position, depth)
- methods (sampling, sample handling, analysis method, sta-

tistical analysis)

- quality control (Quasimeme score, certified and internal reference, inter-calibration)
- meteorology tracked by ODAS (PAR, wind information, solar radiation, atmospheric pressure, temperature)

1.c. Availability of the expected parameters:

- yes / no

1.d. Quality of the data set:

- values expressed in significant figures
- clear description of parameter and measurement unit
- detection limit and expected accuracy

The second step is the control of integrity, which will be performed at the entrance of the data in the database: syntax check, external value check, combination of fields check, check of replicate data, position check, internal value check (IDOD, 1999).
The last level is the statistical quality control, which is being developed by KUL-UCS.

MUMM has screened the different data sets sent by the Project Data Managers (PDM). The results of the screening procedure are summarised in a table for every data set (see Annex 2). For each network, the expected and received parameters are listed. Additional questions and comments are listed in order to clarify some characteristics on data sets during our visit to the laboratory. A good comprehension of the data is a prerequisite for a correct introduction in the database.

2.3. Analysis of the data structure

Since March 1999, data sets collected in the frame of the Programme Sustainable Management of the North Sea were received at MUMM. Every data set had its own characteristics and not all of these could be deduced from the file received. In order to understand the specificity of the data sets and of the four considered substrates (seawater, air, sediment and biota), several meetings took place between the data providers and IDOD. During a visit to the laboratory, more information was received on the applied methodology, the importance of certain items (e.g. exact positioning), extra required information (e.g. volume sampled) etc. The reports of each meeting are presented in Annex 3.

Meta-information and precise description of each data sets are essential for the set-up of the database structure. Based on the files and the reports, adaptations were done on the conceptual scheme.

3. Database structure

3.1. Conceptual scheme for seawater and air

The conceptual scheme for seawater was subject to different changes. First of all as a result of problems encountered with the IDOD prototype during data entry and retrieval. Secondly, adaptations were necessary to enable the storage of data sets other than the seawater monitoring data available at MUMM.

Also data resulting from air samples were received. The structure of this data set was similar to that of concentrations measured in seawater, so the same conceptual scheme can be used. The only difference is that sampling depth will be a negative value.

The adapted conceptual scheme (version2.0) for non-continuous values related to seawater and air is presented in Figure 3.

Parameter and methodology

One of the major changes is the way information on parameters and methodologies is organised. During one sampling event, the same parameter is often measured by different services and/or analysed with different methods. A clear display and correct statistical analysis of these results appeared to be very complicated. More particularly, it was difficult to distinguish values obtained with different methods. As certain services measure a parameter for monitoring purposes with a precise, quality-controlled analytical method, and others measure that parameter for relative comparison with other data, the users should only mix these data in further analysis with care.

This problem was solved by adding a reference of the analytical method and responsible service to the parameter name instead of to the value. The data display resulting from the search for the parameter 'Nitrate+Nitrite' is shown as example in Figure 2. This involves that the user should always take into consideration the applied methodology before proceeding with the data analysis.

Such a combination of parameter code, method, substrate and matrix is encountered in other existing oceanographic databases like OMEX-1 (Lowry, R.K, 1998) and NOWESP (Radach *et al.*, 1996). In these databases character codes are used to indicate the analysis method. In the IDOD database, the method identifier (an AutoNumber) will be used.

Another advantage is that the applied methods can be listed directly for every laboratory (in the previous scheme this had to be done by passing through all values).

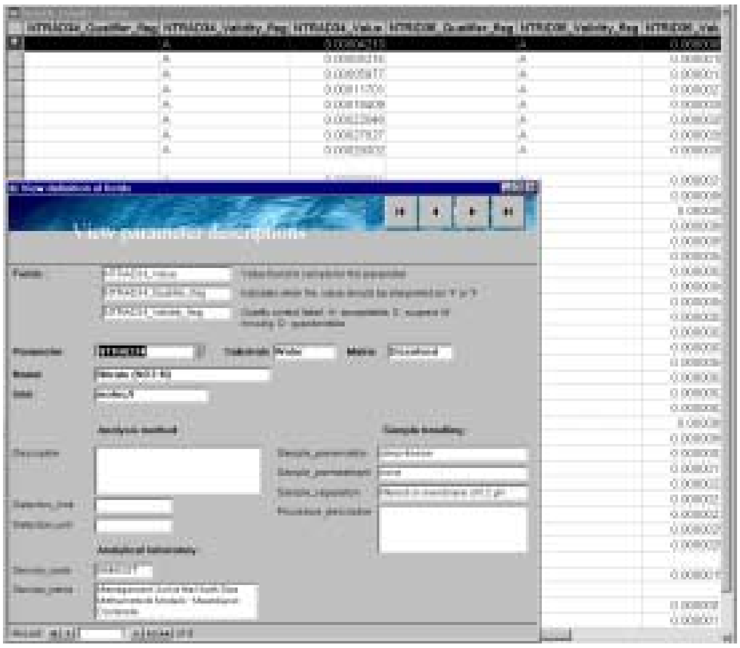


Figure 2. Example of data display

The sampling event

- Sampling position/time has changed into start and end sampling position/time, to enable the storage of data resulting from sampling tracks in the same structure (e.g. collection of air particles, continuous collection of suspended particles by use of the centrifuge on the seawater circuit). Start and end sampling position will also be necessary for other data types, more specifically those obtained by hauls.
- The attribute ‘exceptional circumstances’ is added as a memo-field to store the notes of the laboratory on the sampling conditions.
- Besides the actual sampling depth, a reference sampling depth is included to enable retrieval and analysis based on depth levels *e.g.* “surface”, “bottom”, “middle”, “-3 m”.
- A field is added to indicate the time reference system. If the reference system is known, it will be converted to UTC, if not, it will be noted as unknown. Similarly, another field states the position reference system. Both flags are especially necessary for historical data.
- The volume sampled is added. This is necessary when the volume sampled is variable, e.g. samples of seawater and air particles collected with a pump. Also for other data types, e.g. plankton data obtained with nets, this field will be required. If a sample is taken with a recipient of known volume, this volume is added to the gear type in the entity ‘Sampling Method’.

Analysis method

With the description of the analysis method, a field ‘method codes’ is added. Character codes will be appended in one string, as is done to report methodology information to Quasimeme (Quality Assurance of Information from Marine Environmental Monitoring in Europe). This way the different characteristics of importance for a certain method can be reflected in a flexible way. More information can be given in the description.

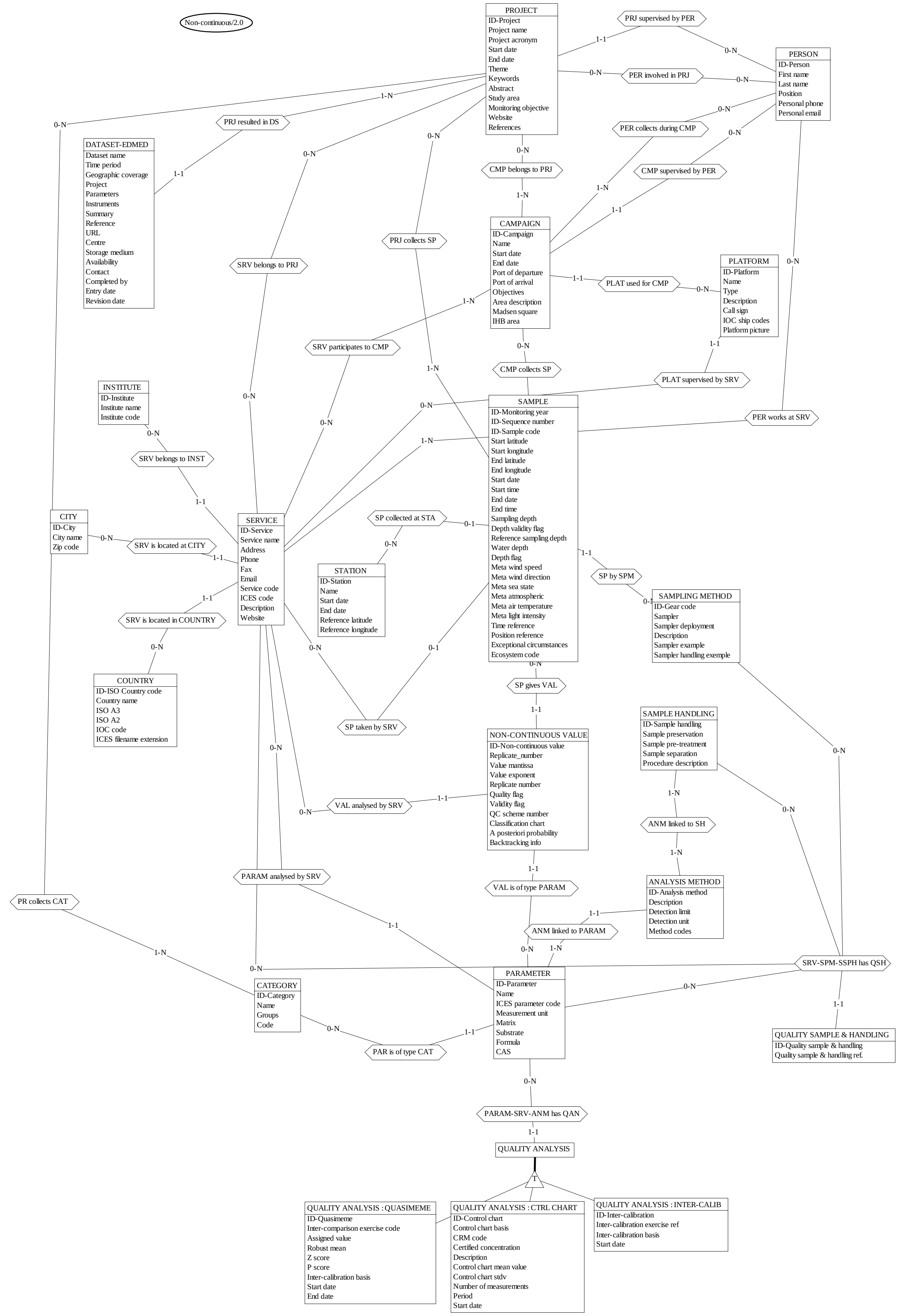


Figure 3. Conceptual scheme (version 2.0) for non-continuous values measured in seawater and air

3.2. Conceptual scheme for plankton

To develop a conceptual scheme for this data type, different information sources were used. The structure of the data collected in the frame of the Programme was analysed. The 'reporting format for biological community data' of ICES (ICES, 1999) was discussed with the data providers and was used as a basis for the construction of the conceptual scheme. Besides this, other oceanographic databases, e.g. OMEX-I (BODC, 1997) and Mariner (The Irish Marine Data Centre, 1999), were consulted.

Data on the diversity and abundance of plankton species are closely related to concentrations measured in seawater. The parameters, e.g. number of individuals of different species, can be stored as water column parameters. Therefore, it was decided to extend the scheme for seawater to store plankton data. The resulting draft scheme is shown in Figure 4. The additions and adaptations of the seawater scheme are indicated in bold.

The seawater scheme was extended in the following way :

- Plankton abundance can be obtained by the analysis of water samples (e.g. bottle, bucket) or through the deployment of 'nets'. Nets can be used in vertical, horizontal or oblique hauls. Therefore additional information will be stored in a separate entity ('Haul'), like the minimum and maximum sampling depth, the approximate horizontal and vertical speed of the tow. The type of net is specified in the entity 'Sampling method'. Extra fields are included : the mesh size of the net and the diameter of the net opening ('sample area opening'). The sample area opening should be inserted to evaluate whether the sample is representative depending on the site characteristics (current and wave properties). The volume of seawater that passed the net, calculated based on flowmeter readings, is given for every sample.
- For phytoplankton, the volume used for sedimentation should be stored for every sample to evaluate if the sample is representative. This volume is variable, depending on the density of specimens encountered. According to the analytical laboratory, the sedimentation time for phytoplankton is constant, so this field is foreseen in the entity 'Sample handling', and not in 'Sample' as in the ICES format. The magnification used to identify the phytoplankton species, will be described with the analysis method.
- For zooplankton, when the density is too high, the sample is divided and only a part is counted. The technique used for splitting should be reported in 'Sample handling'. To be able to report information on the sex code of species, the number of males and females can be inserted together with abundance in the table 'Value'. The development stage will be included in the parameter name, e.g.
  - Centropages hamatus adults – abundance
  - Centropages hamatus nauplii – abundance
  - Copepod other copepodites – abundance

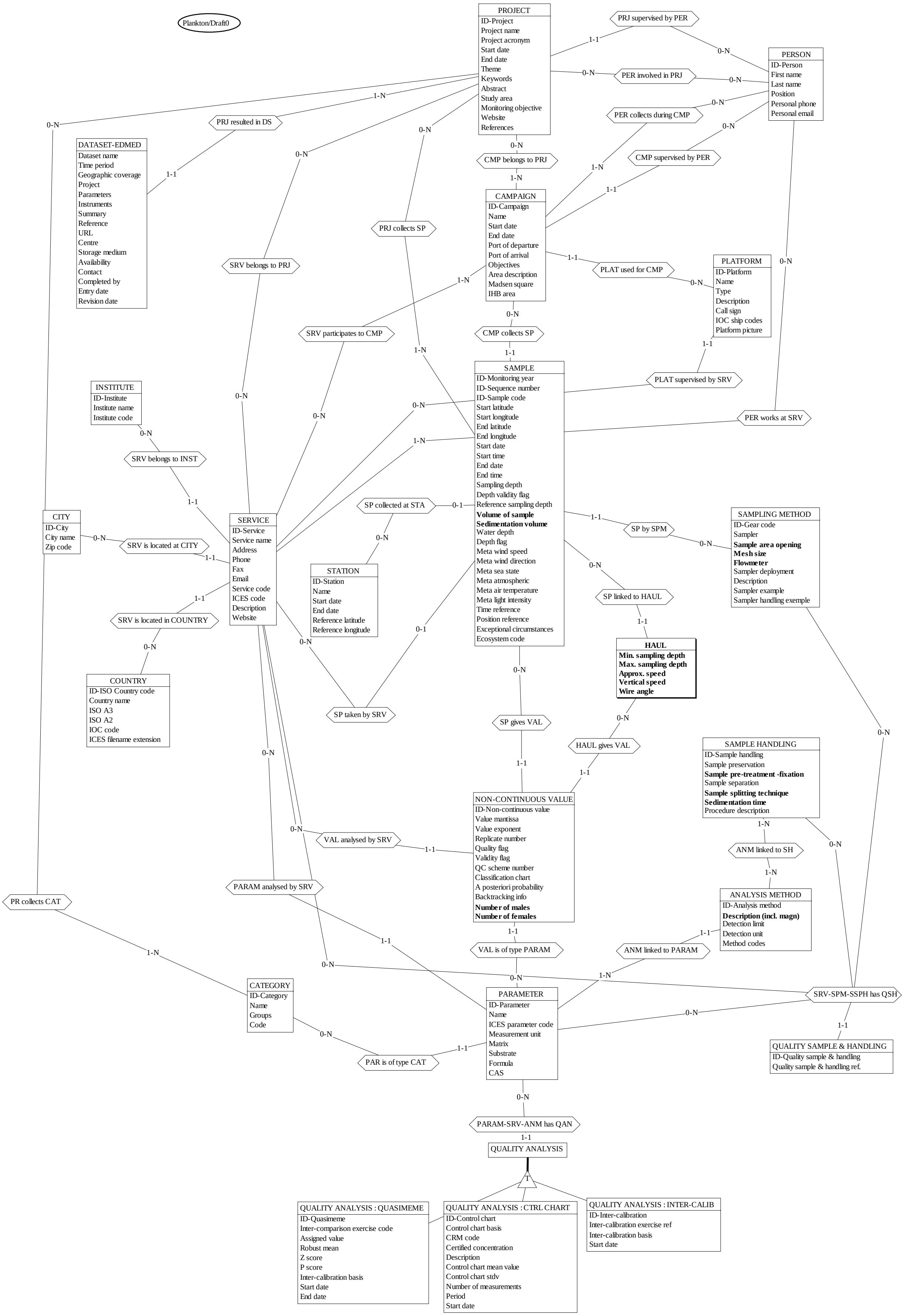


Figure 4. Conceptual scheme for plankton (Version Draft0)

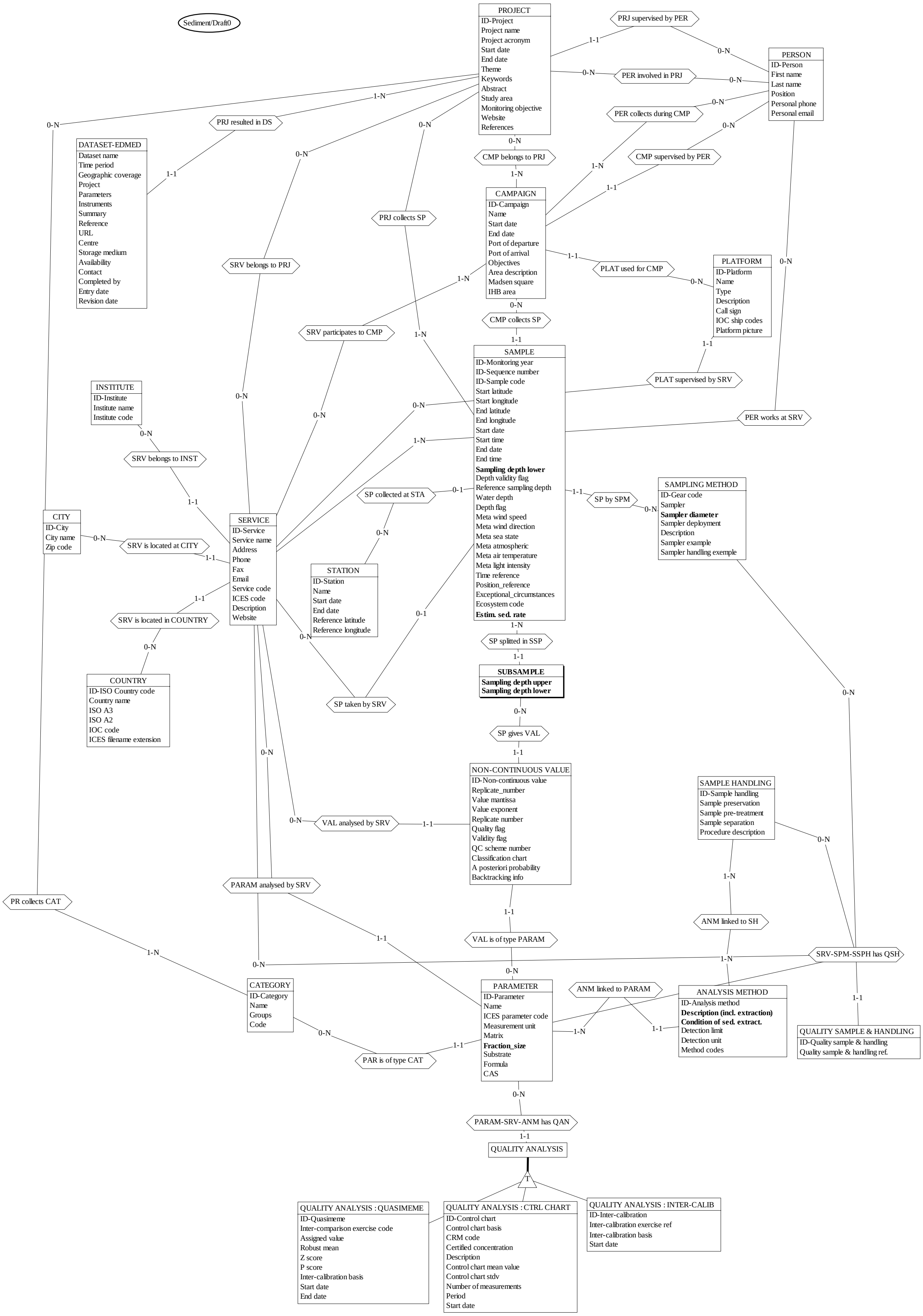
3.3. Conceptual scheme for sediment

The sediment data actually considered are physico-chemical parameters (e.g. grain size distribution) and contaminant concentrations. Compared to the conceptual scheme of seawater, some extra information should be stored for a good comprehension of these data. The main characteristics of sediment data are the sediment fractions and the division of sediment samples in slices along a depth profile.

Figure 5 shows the draft scheme for data on sediment. The additions and adaptations of the seawater scheme are indicated in bold. The scheme is based on the sediment monitoring data actually stored in the ICES format (ICES, 1994) and several discussions with data providers.

Following adaptations were done :

- Each sediment core is a sample and each slice of the sediment core is a sub-sample. The entity ‘Sub-sample’ was introduced with the fields, upper and lower sampling depth. According to the ICES format, when a grid is sampled with grab samplers or box-corers to study the local variability or the reproducibility of the grab samples, the different physical samples are considered as sub-samples and all together they form one sample. However, different sub-samples can be taken from one grab sampler. It was found appropriate by IDOD to indicate the difference between different physical samples and different sub-samples. So every physical sample would be stored separately. The number of replicate samples can still be retrieved by making a search on the monitoring year and sequence number, identifying a sampling occasion. If only average results are available for the set of samples, the start and end positions define the boundaries of the area sampled and the sampler type will contain the number of individual samples taken.
- The estimated sedimentation rate is foreseen by ICES. This parameter is only important in areas with a high sedimentation rate (e.g. Baltic Sea). This field will be included in the entity ‘Sample’ of the IDOD database.
- The description of the methodology should contain the condition of sediment when extracted (e.g. oven dried, freeze dried). The condition of the sediment determines the set of parameters that can be measured on the sample. Also the extraction method should be specified.
- The matrix specifies whether the parameter was measured on pore water, fractionated or unfractionated sediment. If a fraction was analysed, the information on the fraction size will be added in the table ‘Parameter’.



- Figure 5. Conceptual scheme for sediment (Version Draft0)

4. Database prototype for seawater

The prototype for seawater data is implemented in MSAccess. To retrieve data based on a combination of criteria, a general query has been constructed within an Access form. *Microsoft Visual Basic* was used to construct the SQL-code (Viescas, 1997). The code is provided in Annex 4. Figure 6 gives an example of a query for nutrients, dissolved oxygen, temperature and salinity. Figure 7 shows a part of the resulting table.

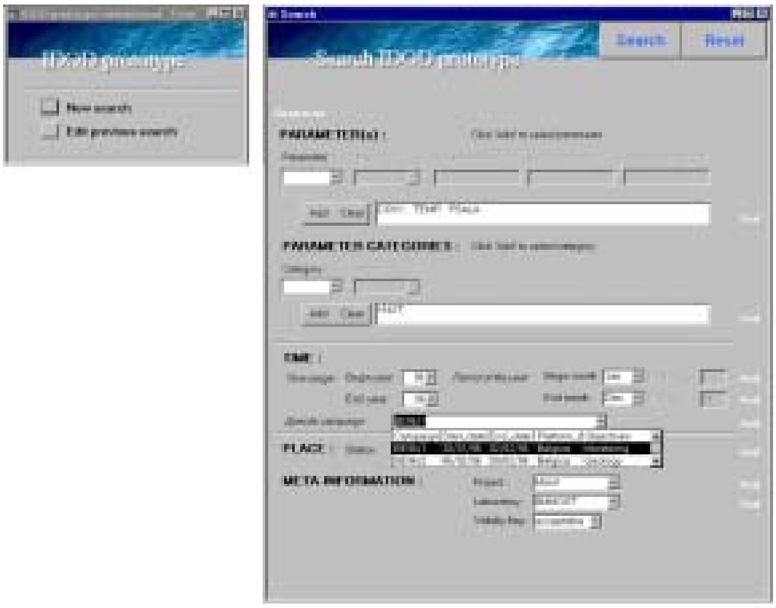
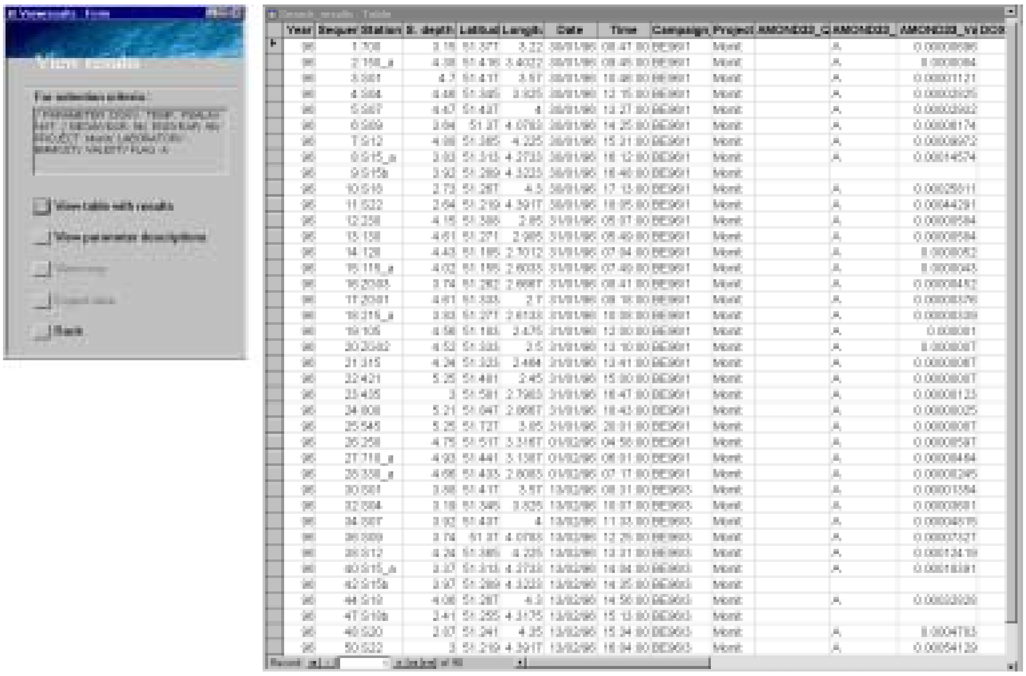


Figure 6. Search form in the IDOD-prototype



Year	Sequen	Station	S. depth	Latitud	Length	Date	Time	Campaign	Project	AMONDI32_Q	AMONDI32_	AMONDI32_Va	DCS
96	1	708	3.18	51.377	3.22	30/01/96	08:47:00	BE961	Mont		A	0.00008595	
96	2	158_a	4.38	51.418	3.4022	30/01/96	09:45:00	BE961	Mont		A	0.0000084	
96	3	S01	4.7	51.417	3.57	30/01/96	10:48:00	BE961	Mont		A	0.00001121	
96	4	S04	4.48	51.385	3.825	30/01/96	12:15:00	BE961	Mont		A	0.00002925	
96	5	S07	4.47	51.437	4	30/01/96	13:27:00	BE961	Mont		A	0.00002932	
96	6	S09	3.84	51.37	4.0783	30/01/96	14:25:00	BE961	Mont		A	0.00008174	
96	7	S12	4.08	51.385	4.225	30/01/96	15:21:00	BE961	Mont		A	0.00008872	
96	8	S15_a	3.93	51.318	4.2733	30/01/96	16:12:00	BE961	Mont		A	0.00014574	
96	9	S15b	3.92	51.289	4.3223	30/01/96	16:40:00	BE961	Mont				
96	10	S18	2.73	51.287	4.5	30/01/96	17:13:00	BE961	Mont		A	0.00025511	
96	11	S22	2.84	51.219	4.3917	30/01/96	18:05:00	BE961	Mont		A	0.00044291	
96	12	230	4.15	51.588	2.85	31/01/96	05:37:00	BE961	Mont		A	0.00008584	
96	13	150	4.61	51.271	2.985	31/01/96	05:49:00	BE961	Mont		A	0.00008584	
96	14	120	4.43	51.185	2.7012	31/01/96	07:04:00	BE961	Mont		A	0.0000762	
96	15	115_a	4.02	51.155	2.6035	31/01/96	07:49:00	BE961	Mont		A	0.0000043	
96	16	2303	3.74	51.282	2.6687	31/01/96	08:41:00	BE961	Mont		A	0.00008452	
96	17	2301	4.61	51.333	2.7	31/01/96	08:18:00	BE961	Mont		A	0.00008376	
96	18	215_a	3.93	51.271	2.6133	31/01/96	10:08:00	BE961	Mont		A	0.00008339	
96	19	105	4.58	51.183	2.475	31/01/96	12:00:00	BE961	Mont		A	0.000081	
96	20	2302	4.53	51.333	2.5	31/01/96	13:10:00	BE961	Mont		A	0.0000087	
96	21	315	4.34	51.333	3.484	31/01/96	13:41:00	BE961	Mont		A	0.00000087	
96	22	421	5.25	51.481	2.45	31/01/96	15:00:00	BE961	Mont		A	0.00000087	
96	23	435	3	51.581	2.7983	31/01/96	16:47:00	BE961	Mont		A	0.00008123	
96	24	808	5.21	51.047	2.0687	31/01/96	18:43:00	BE961	Mont		A	0.00008025	
96	25	545	5.25	51.727	3.05	31/01/96	20:01:00	BE961	Mont		A	0.00000087	
96	26	258	4.75	51.517	3.3167	01/02/96	04:58:00	BE961	Mont		A	0.00008597	
96	27	718_a	4.93	51.441	3.1587	01/02/96	08:01:00	BE961	Mont		A	0.00008464	
96	28	330_a	4.68	51.435	2.6083	01/02/96	07:17:00	BE961	Mont		A	0.00008295	
96	30	S01	3.88	51.417	3.57	13/02/96	08:31:00	BE963	Mont		A	0.00001584	
96	32	S04	3.18	51.385	3.825	13/02/96	10:07:00	BE963	Mont		A	0.00003581	
96	34	S07	3.92	51.437	4	13/02/96	11:33:00	BE963	Mont		A	0.00004575	
96	36	S09	3.74	51.37	4.0783	13/02/96	12:25:00	BE963	Mont		A	0.00007327	
96	38	S12	4.34	51.385	4.225	13/02/96	13:31:00	BE963	Mont		A	0.00012418	
96	40	S15_a	3.37	51.318	4.2733	13/02/96	14:04:00	BE963	Mont		A	0.00018381	
96	42	S15b	3.97	51.289	4.3223	13/02/96	14:25:00	BE963	Mont				
96	44	S18	4.06	51.287	4.5	13/02/96	14:56:00	BE963	Mont		A	0.00032928	
96	47	S18b	2.41	51.255	4.3175	13/02/96	15:13:00	BE963	Mont				
96	48	S20	3.07	51.241	4.35	13/02/96	15:34:00	BE963	Mont		A	0.0004783	
96	50	S22	3	51.219	4.3917	13/02/96	16:04:00	BE963	Mont		A	0.00054129	

Figure 7. Result of the search

Display and analysis of replicate samples and replicate values

A problem for display and statistical analysis occurs with data for one parameter obtained on several samples at one sampling station. Such 'simultaneous' values provide information on sampling reproducibility or the homogeneity of the substrate sampled.

Replicate analyses of a given parameter in a particular sample provide information on the analytical reproducibility.

For both problems it was decided to only display and use the average value in the standard procedures of the statistical analysis. Of course for specific applications, the individual values can still be retrieved.

Software

Actually contacts take place with software providers to evaluate different Relational Database Management Systems (RDBMS). The systems considered are *Sybase*, *Oracle* and *SQL-server*. To improve the quality and rapidity of the database design and implementation, also a case tool will be purchased offering a complete workflow from the conceptual model to the physical implementation into a RDBMS.

5. GIS and visualisation

The task concerning spatial visualisation has been considered by MUMM during the year 1999. A comparative study of the existing geographic information systems was made, ArcView 3.1. and the extension Spatial Analyst, best fulfilled our needs and was purchased. Actually, the exploratory phase of the tool with different data sets is ongoing. ArcView 3.1 is already successfully linked with the database prototype, using an ODBC connection. Data sets are envisaged to produce different types of maps.

The first task considered is the evaluation of the interpolation methods proposed by the software. The aim is to elaborate, from punctual measurements, a spatially continuous distribution that best represents the real situation. However, different results are obtained with the different interpolation methods and it is difficult to define the most adequate method since this depends on the parameter considered. Therefore, a scientific evaluation of the results is necessary. The four interpolation methods actually studied with ArcView are: Inverse Distance Weighted, Spline, Kriging and Trend. From the ArcView literature (ArcView, 1996), some interpolators base assumptions are described.

The IDW interpolator assumes that each input point has a local influence that diminishes with distance. It weights the points closer to the processing cell greater than those farther away. A specified number of points, or all points within a specified radius, can be used to determine the output value for each location. Use of this method assumes that the variable being mapped decreases in influence with distance from its sampled location.

The Spline interpolator is a general purpose interpolation method that fits a minimum-curvature surface through the input points. It fits a mathematical function to a specified number of nearest input points, while passing through the sample points. This method is best for gently varying surfaces such as elevation, water table heights or pollution concentrations. It is not appropriate if there are large changes in the surface within a short horizontal distance because it can overshoot estimated values.

The Kriging interpolator is a specialised interpolation method that assumes the distance or direction between sample points shows spatial correlation that helps describe the surface. Kriging fits a mathematical functions to a specified number of points, or all points within a specified radius, to determine the output value for each location. The use of kriging involves several steps: exploratory statistical analysis of the data, variogram modelling, then creating the surface and analysing its optional variance surface. This function is most appropriate when you know about spatial correlated distance or directional bias in the data.

The Trend interpolator fits a mathematical function, polynomial of specified order, to all input points. When calculating the mathematical function to describe the resulting surface, Trend uses a least-squares regression fit. The resulting surface minimises the variance in relation to the input point values. That is, at all the know input points, the sum of the differences between actual and estimated values will be as small as possible.

An elementary analysis consists in a visual “evaluation” of the spatial distribution map produced. IDW seems to give good results in the estuary, while Kriging is more adapted for open sea interpolation. The Spline method gives also quite good results, but it does not allow to set a barrier line theme.

However, more elaborated evaluation techniques have to be considered to evaluate the most suitable interpolation method: comparison with continuous data sets, with satellite images, ... of course only possible for a limited set of parameters.

Besides interpolations and cartographic representation of data, Arc View will be used to answer specific spatial queries and facilitate decision making.

Till now, the GIS has been used to elaborate maps for the National Comments (1998) and for a slideshow presented during the workshop entitled “A la recherche d’un développement durable entre science et politique – Op zoek naar een duurzame dialoog tussen onderzoek en beleid” (Brussels, 24-25 November 1999).

6. Convention

The technical annexes of the research projects of the programme “Sustainable Management of the North Sea” specify that the data sets of the projects have to be submitted to the IDOD data centre (see § 2.1.2).

In order to clarify the conditions of data transfer and data access, a convention between the OSTC, the IDOD data centre and the project promoters was drafted by MUMM. This convention defines the rights and duties of the parties involved, such as the date of data-submission to the IDOD data centre and the access rules and embargo for the different user groups of the IDOD database. This draft convention was sent to all project promoters for comments and was discussed during the IDOD co-ordination meetings. At the end of 1999, a nearly agreed version, that has to be finalised by bilateral contacts between MUMM and the project promoters, was available. The most recent version of the convention is included as Annex 5.

7. Products

7.1. International reporting obligations

MUMM has fulfilled the Belgian international obligations in the frame of the “Joint Assessment and Monitoring Programme” (JAMP) and “Nutrient Monitoring Programme” of the Oslo and Paris Commission (OSPAR) by reporting the monitoring data for 1998 to ICES.

The submitted data are general hydrographic data such as salinity, temperature, dissolved oxygen; concentrations of the nutrients nitrate, nitrite, ammonium, phosphate and silicate and the pigments chlorophyll and phaeophytine in the Western Scheldt estuary and the North Sea. For each parameter, a method description for sampling and analysis is included. The validity of the data is checked by using the ICES screening programme. The results are presented in the first draft of the National Comments for the monitoring year 1998. Due to the delay of the submission of the monitoring data these draft National Comments could not yet be finalised.

7.2. *Newsletter 2 and 3*

In 1999, two Newsletters were produced to inform data providers about the developments within the IDOD project. Issues 2 and 3 of the Newsletter became respectively available in March 1999 and December 1999. They were distributed to the members of the co-ordination committee, the project promoters, the project data managers and other involved persons.

Issue 2 and 3 of the Newsletter are attached as Annexes A and B to the Joint Scientific Report.

7.3. *IDOD webpages*

The IDOD web pages give a description of the IDOD project. Also the Newsletters and the comments on them are made available on the website.

In 1999, the IDOD team prepared an extensive description of the project in the form of a slideshow for the workshop “A la recherche d’un dialogue durable entre science et politique – Op zoek naar een duurzame dialoog tussen onderzoek en beleid” (Brussels, 24-25 November 1999). This slideshow remains available on the IDOD website (http://idod.mumm.ac.be/slideshow/index.htm).

7.4. *Publication*

The design of the IDOD seawater database was presented during a CoastGIS symposium at Brest, France (9 tot 11 September 1999). The presented paper “Design of an oceanographic database” is selected for publication in the CoastGIS’99 book. The paper is included as Annex C to the Joint Scientific Report.

8. Conclusion

The update of the inventory of the data sets has continued. A new list of parameters measured in the frame of the Programme was established and EDMED is being updated. The first data sets of the Programme were received in March 1999. A procedure was set-up to screen the data and data documentation for completeness and comprehensibility. Although all the laboratories were asked to submit the missing meta-information, this information is still not available for some data sets.

Once the data were received, an evaluation of the characteristics was made during personal contacts with the data providers. These informations were essential for the set-up of the conceptual schemes for sediments and plankton and for adaptations and improvements of the conceptual scheme for seawater and air. For plankton and sediment data, also the ICES reporting formats for environmental data and existing oceanographic databases were used to construct the conceptual schemes. The scheme for seawater was already implemented in a prototype in which the changes to the conceptual scheme were incorporated. This prototype was tested with real data using a first user interface to enable data retrieval based on different search criteria. These tests also resulted in changes of the conceptual scheme.

A link is made between the prototype and a geographic information system. The GIS, ArcView 3.1 and its extension Spatial Analyst, purchased by MUMM is actually explored. It offers the possibility to elaborate spatially continuous distributions from punctual measurements. Actually, the different interpolation methods are considered to define which one will give the more realistic results. This tool will also permit to answer specific spatial query from the data sets.

The Convention between the IDOD data centre, the OSTC and the data provider was discussed efficiently in 1999 and is nearly finalised. It clarifies the conditions of data transfer and data access to the database for different user groups.

9. References

ArcView, 1996, Spatial Analysis User Guide.

BODC, 1997. OMEX-1 Project Data Set CD-ROM.

ICES, 1994. Reporting Format for Sediment Data. Version 2.2 – November 1994. ICES Environmental Data Reporting Formats.

ICES, 1999. Reporting Format for Biological Community Data. DRAFT Version BRF99-September 1999. ICES Environmental Data Reporting Formats.

IDOD, 1999. Joint Scientific Reports for the year 1998. Scientific Support Plan for a Sustainable Development Policy. “Sustainable Management of the North Sea”. Research Contracts MN/DD/60, 61 & 62.

JANS, S., 1998, Draft National Comments to the 1998 Belgian monitoring data for seawater. Submitted by Belgium for ICES and PARCOM, Ref. DMG/1/SJ/199908/EN/IR.

LOWRY, K.R., LONCAR, Z. and DOWNER, R., 1998. Data Management for the OMEX I project. A Case Study. In : BOHLE-CARBONELL, M. (Ed.). Experiences in project data management. Marine science and technology programme.

MULLER, F., K. DE CAUWER, L. SCHWIND, M. DEVOLDER and S. SCORY, 1999. Design of an oceanographic database. Submitted to conference COASTGIS '99, Brest, 9-11 September 1999.

OSLO AND PARIS COMMISSIONS, 1995. Nutrient Monitoring Programme. Adopted by OSPAR 1995, OSPAR 95/15/1, Annex 12.

OSLO AND PARIS COMMISSIONS, 1995. The Joint Assessment and Monitoring Programme. ISBN 0 946955 41 7.

RADACH, G., GEKELER, J., BECKER, G., BOT, P., CASTAING, P., COLIJN, F., DAMM, P., DANIELSSEN, D., FOYN, L., GAMBLE, J., LAANE, R., MOMMAERTS, J.P., NEHRING, D. , PEGLER, K., VAN RAAPHORST, W. and WILSON, J., 1996. The NOWESP Research Data Base. German Journal of Hydrography, Vol. 48 (241-259).

The Irish Marine Data Centre, 1999. The MARINER System User Guide.

VIESCAS, J., 1997. Microsoft Access au quotidien. Microsoft Press, 969 p.

COLOPHON

This report was issued by MUMM in January 2000.

Its reference code is MOD code.

Status ☐ draft
☐ final version
☐ revised version of document
☒ confidential

Available in ☒ English
☐ Dutch
☐ French

If you have any questions or wish to receive additional copies of this document, please send an e-mail to *idod@mumm.ac.be*, quoting the reference, or write to:

MUMM
100 Gulledelle
B–1200 Brussels
Belgium
Phone: +32 2 773 2111
Fax: +32 2 770 6972
http://www.mumm.ac.be/

MANAGEMENT UNIT OF THE
NORTH SEA MATHEMATICAL MODELS

MODELLING GROUP

The typefaces used in this document are Gudrun Zapf-von Hesse’s *Carmina Medium* at 10/14 for body text, and Frederic Goudy’s *Goudy Sans Medium* for headings and captions.

Annex 1 – Data sets inventory : Situation in January 2001

LABORATORY	Data 97	Data 98	Data 99	See comments
Pr. M. Vincx	March 99 (I)	March 99 (I)	🟦	X
Pr. E. Kuijken	Sept 99 (C)	Sept 99 (C)	🟦	X
Pr. F. Ollevier	august 99 (C)	October 00 (C)	March 00 (C)	
Pr. J-M. Bouquegneau	March 99 (I)	March 99 (I)	March 00 (I)	X
Pr. Fr. Coignoul	March 99 (I)	March 99 (I) / Sept 99 (C)	March 00 (I)	X
Pr. Cl. Joiris	🟦	🟦	🟦	X
Pr. P. Meire	Sept 99 (C)	Sept 99 (C)	Sept 99 (I)	X
Pr. Ch. Lancelot	June 99 (C)	June 99 (C)	🟦	X
Dr. M. Tackx	XXX	Dec 99 (C)	June 00 (I)	X
Dr. K. Ruddick	Sept 99 (C)	Sept 99 (C)	July 00 (C)	
Pr. R. van Grieken	XX	May 99 (C)	March 00 (C)	
Pr. W. Baeyens	XXX	March 00 (I) / Nov 00 (C)	March (I) / Nov 00 (C)	
Pr. H. Van Langenhove	March 99 (C)	March 00 (C) *	March 00 (C)	
Pr. W. Wollast	March 99 (I)/ March 00 (I)	March 99 (I)	🟦	X
Dr. Ph. Dubois	XXX	March 99 (I) / March 00(C)	March 00 (C)	
Pr. M. Jangoux	XXX	March 99 (I) / March 00(C)	March 00 (C)	
Pr. R. Flammang	XXX	March 99 (I) / March 00(C)	March 00 (I)	X

🟦 nothing received/example

C=complete

I=incomplete

XXX = no campaign

XX = only testing, no real valid data

* Delay due to problems with analytical techniques

Figure 1. Data sets inventory - Situation in January 2001

1. Data sets analysis (20/12/1999)

1.1. The structural and functional biodiversity of the North Sea ecosystems – Species and their habitats as indicators for the sustainable management of the Belgian coastal shelf

PDM: André Cattrijsse (RUG – Marine Biology Section)

1.1.1. Data set from Pr. M. Vincx (A. Cattrijsse)

1.a. Availability of the data set	YES
Date of delivery	26/03/1999
File name	Marbioug.wk4 Epi-do.wk4 (given as examples)

1.b. Availability of the meta-information	General	Incomplete
	Methods	To receive*
	Quality Control	Not available**
	Meteorology	See ODAS

1.c. Availability of the expected parameters	Pr. M. Vincx	Incomplete (1997) Incomplete (1998)

1.d. Quality of the data set	Significant figures	To check
	Description of parameter	YES
	Unit	YES
	Detection limit	Not available*
	Quantification limit	Not available*
	Expected accuracy	Not available*

* Discussed during the meeting at UG, the information will be sent by Andre Cattrijse for the data gathered by the *Sectie Mariene Biologie*

** Discussed during the meeting at UG for data gathered by the *Sectie Mariene Biologie*, it should be checked whether this information is available for nutrient and pigment analysis

<u>Expected / received* parameters:</u>

<u>For metadata:</u>

- Meteorology* (see ODAS):
  - wind speed
  - wind direction
  - solar radiation
  - atmospheric pressure
  - temperature

For water:

- Physical* (see CTD-profiles):
  - temperature
  - depth
  - PAR
- Major inorganic* (see CTD-profiles)::
  - salinity
- Nutrients:
  - nitrate*
  - nitrite*
  - phosphate*
  - ammonium*
- Pigments:
  - chlorophyl-a*
  - chlorophyl-c*
  - fuccoxanthine*
- Optical* (see CTD-profiles)::
  - optical backscatter

For sediment:

- Physical:
  - granulometry
- Interstitial water:
  - nutrients: nitrate, nitrite, ammonia, phosphate
  - pigments: chlorophyl-a and –c, fuccoxanthine

For biota:

- Benthos:
  - diversity index
  - #species*
  - density*
  - biomass
  - dominance index
  - length frequency distribution
  - weight frequency distribution

Questions / comments:

More information on the data sets of the Sectie Mariene Biologie was gathered during a visit to the laboratory (10th of June, 1999). Some final data sets (with meta-information : date, time, position) are expected in September : hyper- and epibenthos and stomach analyses. Also a description of the methodology will be sent. Meio- and macrobenthos and sediment analysis results are still missing.

1.1.2. Data set from F. Ollevier (E. Gysels)

1.a. Availability of the data set	YES
Date of delivery	26/08/1999
File name	Alpmin.wk4

1.b. Availability of the meta-information	General	Incomplete
	Methods	NO
	Quality Control	NO
	Meteorology	NO

1.c. Availability of the expected parameters	Pr. F. Ollevier	Incomplete (1997) Incomplete (1998)

1.d. Quality of the data set	Significant figures	To check
	Description of parameter	NO
	Unit	NO
	Detection limit	NO
	Quantification limit	NO
	Expected accuracy	NO

Expected / received* parameters:

For biota:

- Fish – varia:
  - genetic structure*
  - parasites: species and incidence
  - stomach analysis

Questions / comments :

- Species analysed
- Descriptive information for samples/individuals taken (e.g. size, weight, sex, …)
- Samples only taken to study genetic structure ? Or is same sample or subsample also analysed by another laboratory ?
- List of codes

1.1.3. Data set from Pr. E. Kuijken (J. Seys)

1.a. Availability of the data set	YES
Date of delivery	01/09/99
File name	BsasIDOD.doc BsasIDOD.mdb

1.b. Availability of the meta-information	General	YES
	Methods	NO
	Quality Control	NO
	Meteorology	NO

1.c. Availability of the expected parameters	Pr. P. Meire	YES (1992-1999)

1.d. Quality of the data set	Significant figures	To check
	Description	YES
	Unit	YES
	Detection limit	NO
	Quantification limit	NO
	Expected accuracy	NO

Questions/comments:

- Table ‘BELGTRIP’
  - Contains information on executed campaign days (308). Can attributes differ during a campaign (e.g. Count_type, use_of_binoculars ...) ?
  - How is *tripkey* composed ?
  - ‘Base_type’ ?
  - Transect width : depending on visibility ?
  - Sometimes conflict between number of Observer_names and Number_of_X
  - Species_observed : depending on time or experience ?

- BELGBASE2
  - How is ‘poskey’ defined ?
  - Wave, Wind_force and visibility : estimated ?

1.2. Birds and marine mammals of the North Sea: Pathology and Ecotoxicology

PDM: Virginie Debacker (Ulg – Laboratoire d'Océanologie)

1.2.1. Data set from Pr. J.-M. Bouquegneau (V. Debacker)

1.a. Availability of the data set	YES
Date of delivery	24/03/1999 (old version) 23/08/1999 (new version)
File name	IDOD.doc IDOD.wk4

1.b. Availability of the meta-information	General	YES
	Methods	YES
	Quality Control	YES
	Meteorology	Not available

1.c. Availability of the expected parameters	Pr. J.-M. Bouquegneau	Incomplete (1997) Incomplete (1998)

1.d. Quality of the data set	Significant figures	To check
	Description	YES
	Unit	YES
	Detection limit	YES
	Quantification limit	YES
	Expected accuracy	YES (Recovery)

*These data have been discussed during the meeting held at ULg and will be sent by Virginie Debacker.

<u>Expected / received* parameters:</u>

<u>For biota:</u>

- Seabirds – Marine mammals:
  - Metals: Cd*, Cr*, Cu*, Fe*, Ni*, Pb*, Zn*, Selenium*(for kidneys and livers of mammals)
  - Organic: total* lipids
  - Varia: metallothioneins

<u>Questions / comments:</u>

The visit held at ULg (22/06/99) has permitted us to receive answers on the different questions.

<u>Data set from Pr. J.-M. Bouquegneau:</u>
Biota:

- The dataset is considered as incomplete because metallothionein concentrations are not still available . The technique is in development (even so the analyses will be made on biota collected in years 97-98).
- Lipid concentration data: the values represent the total lipids.

- Selenium concentrations: the analyses are only made on kidneys and livers of mammals (no data for muscles). Selenium concentrations are not measured for birds.
- Lipid concentrations are only measured for livers and muscles (no data for kidneys).
- For birds: if the bird stays more than three days in an asylum, it’s not analysed because it has receive health care which can distort the results.
- Last year, there have been between 2500 and 3500 birds grounded, of which 60% in February.
- For mammals: some mammals accidentally catched by fishermen are analysed, but only 1/10 of the catches (estimation) is received for analysis.
- The number of mammals analysed (toxicology) each year is around 20. The number of autopsies can be greater as putrefied animals are autopsied but not sent for toxicological analyses.

1.2.2. Data set from Pr. F. Coignoul (T. Jaugniaux)

1.a. Availability of the data set	YES
Date of delivery	24/03/1999 (old version) 23/08/1999 (new version)
File name	IDOD.doc IDOD.wk4

1.b. Availability of the meta-information	General	YES
	Methods	To receive*
	Quality Control	To receive*
	Meteorology	Not available

1.c. Availability of the expected parameters	Pr. F. Coignoul	Incomplete (1997) Incomplete (1998)

1.d. Quality of the data set	Significant figures	To check
	Description	YES
	Unit	YES
	Detection limit	YES
	Quantification limit	YES
	Expected accuracy	To receive*

*These data have been discussed during the meeting held at ULg and will be sent by T. Jaugniaux.

Expected / received* parameters:

For biota:

- Seabirds – Marine mammals:
  - Organic: PAHs
  - Varia: pathology*, parasites* (*only for birds)

Questions / comments:

Data set from Pr. F. Coignoul:

Biota:

- The dataset is considered as incomplete because the detailed autopsy results have not been sent (+ an example of autopsy report).
- PAHs: as the collaboration project with another laboratory has not been successful, these analyses are not made actually.
- Pathology and parasites: these data are available in a report for every specimen. A summary of the results will be sent (not easy to put them into a database because they are presented in a text format; so a code for each type of pathology has to be evolved).

1.2.3. Data set from P. Meire (J. Seys)

1.a. Availability of the data set	YES
Date of delivery	01/09/99
File name	OlieIDOD.xls

1.b. Availability of the meta-information	General	YES
	Methods	NO
	Quality Control	NO
	Meteorology	NO

1.c. Availability of the expected parameters	Pr. P. Meire	1992- March 1999

1.d. Quality of the data set	Significant figures	YES
	Description	YES
	Unit	YES
	Detection limit	Not applicable
	Quantification limit	Not applicable
	Expected accuracy	Not applicable

Expected / received* parameters:

For biota:

- Seabirds – Marine mammals:
  - Ecology: diversity*, density*, #per species*, development*, plumage stage*

Questions / comments:

- Data set from January 1992 to March 1999
- Very well structured

- Transects : difference OONP and NPOO ? Km are constant, sometimes however 15 km or even 17 and 17.5 ?
- Only two transects in March and April, counting type however is noted as weekly ?
- Wind : how determined ?
- Method : walking, car or bike .. influence on results ? what if empty ?
- Oil on beach : what if empty ?
- Observer : codes are missing (e.g. WB, DB, PG, PJ, …; EA *et alli* ?)
- Euringcode : international standard ?
- Age : what if empty ?
- Plumage : how is the schedule of the year determined ? Are combinations possible ?
- What if the species can not be determined ? euring = 0 ?
- Counting of birds brought to recovery centre ?

1.2.4. Data set from Ludo Holsbeek (Pr. C. Joiris)

1.a. Availability of the data set	YES
Date of delivery	24/09/99
File name	VUB1, Hg en dwfw.xls VUB2, PCB and opDDE.xls

1.b. Availability of the meta-information	General	YES
	Methods	NO*
	Quality Control	NO*
	Meteorology	Not available

1.c. Availability of the expected parameters	Pr. C. Joiris	Incomplete (1997) Incomplete (1998)

1.d. Quality of the data set	Significant figures	To check
	Description	YES
	Unit	YES
	Detection limit	NO*
	Quantification limit	NO*
	Expected accuracy	NO*

*These data have been discussed and will be sent by Ludo Holsbeek.

Expected / received* parameters:

For biota (seabirds and marine mammals):

- Metals: total Hg*, methyl Hg*
- Organic: hydrocarbons, polar lipids*, PCBs*
- Organochlorines: DDE*, DDT, aldrin, lindane, heptachlor epoxide

Questions / comments:

The data were discussed during the visit held at VUB-ETOX (14/10/99). A report with more details on the dataset can be found in the annex.

PCB's are measured on lipids, total Hg on fresh weight and MMHg on dry weight. The necessary ratio's (dw/fw, lw/dw) will be sent as well to enable conversion.

1.3. *AMORE – Advanced MOdelling and Research on Eutrophication*

PDM: Véronique Rousseau (ULB – GMMA)

1.3.1. Data set from Véronique Rousseau (Pr. Ch. Lancelot)

1.a. Availability of the data set	YES
Date of delivery	22/06/98
File name	ESA1997.xls ESA1998.xls ESAdia97.xls ESAdia98.xls

1.b. Availability of the meta-information	General	YES
	Methods	YES
	Quality Control	YES
	Meteorology	ODAS

1.c. Availability of the expected parameters	Pr. Ch. Lancelot	Incomplete (1997) Incomplete (1998)

1.d. Quality of the data set	Significant figures	OK
	Description	YES
	Unit	YES
	Detection limit	YES
	Quantification limit	NO
	Expected accuracy	YES

Expected / received* parameters:

For metadata:

- Meteorology: wind speed, wind direction

For water:

- Physical: temperature*, suspended matter
- Major inorganic: salinity*
- Nutrients: nitrate+nitrite*, nitrate, nitrite, phosphate*, silicate*, ammonia*
- Pigments: chlorophyll-a*

For biota:

- Phytoplankton ecology: composition*, enumeration*
- Bacterio-, nanophytozooplankton, nanoprotozoo-, microprotozooplankton: composition, enumeration (data for 1998 to receive; no data for 1997)

Questions / comments:

Data set from Pr. Ch. Lancelot:

Water:

- Physical parameter: MES not received. No data available for 1997 due the use of an unsuitable filter. Data for 1998 to receive.
- Nutrients: separate data for nitrite and nitrate to receive
- Measurements concerning carbon in water will be made by ULB from 1999.

1.3.2. Data set from Dr. M. Tackx (E. Antajan)

1.a. Availability of the data set	YES
Date of delivery	09/12/99
File name	PigMethod.doc PigResults98.xls ZooMethod.doc ZooResults98.xls

1.b. Availability of the meta-information	General	YES
	Quality Control	YES
	Meteorology	ODAS

1.c. Availability of the expected parameters	Dr. M. Tackx	No data expected (1997) Complete (1998)

1.d. Quality of the data set	Significant figures	To check
	Description	YES
	Unit	YES
	Detection limit	YES
	Quantification limit	NO
	Expected accuracy	YES

Expected / received* parameters:

For water: (new parameters to include in the list)

- Pigments: chlorophyll-a*, chlorophyll-c, fuccoxanthine*, phaeopigment (peridinin, 19’hexanoxanthine, diadinoxanthine, alloxanthine)

For biota:

- Mesozooplankton: composition*, abundance*

Questions / comments:

New parameters have to be included in the parameter list: Pigments in water (chlorophyll-a, chlorophyll-c, Peridinin, Fucoxanthin, 19’hexanoxanthine, diadinoxanthine, alloxanthin) are measured by VUB.

1.3.3. Data set from Kevin Ruddick

1.a. Availability of the data set	YES
Date of delivery	24/09/99
File name	1997-1998-water.xls 1997-1998-water.doc 1997-16-PNF.xls 1997-16-PNF.doc 1998-08-RASrad.xls Coordinates.doc

1.b. Availability of the meta-information	General	YES
	Methods	YES
	Quality Control	YES
	Meteorology	ODAS

1.c. Availability of the expected parameters	K. Ruddick	Only CTD and yellow substance abs. missing

1.d. Quality of the data set	Significant figures	To check
	Description	YES
	Unit	YES
	Detection limit	YES
	Quantification limit	NO
	Expected accuracy	NO for PNF measurements

Expected / received* parameters:

For metadata:

- Meteorology: PAR, wind speed, wind direction

For water:

- Physical: temperature (v,c), suspended matter*, secchi depth*
- Major inorganic: salinity (v,c)
- Pigments: chlorophyll-a*, phaeopigment*
- Maps – satellite derived: suspended matter, chlorophyll-a
- Optical: artificial fluorescence, downwelling PAR irradiance, optical back-scatter, scalar PAR irradiance*, upwelling fluorescence radiance*, PAR attenuation coefficient*, upwelling radiance spectra*, downwelling irradiance spectra*, sub-surface irradiance spectra, phytoplankton absorption spectra, yellow substances absorption spectra

Questions / comments:

- Artificial fluorescence: the artificial fluorimeter has been lost at sea, so there will have no data in the immediate future (so this parameter has to be deleted from the list)
- IDOD agrees with the condition asked by Machteld Rijkeboer on the use of data : optical property data and PR-650 measurements of spectral reflec-

tance collected in the frame of the MULTICOLOR project.

- Yellow substance absorption data collected in the frame of MULTICOLOR: data gathered during cruises by HD and lab tests by IS with problems of quality as the methodology had to be tested. This should be mentioned clearly with the project description in the database.
- PNF measurements (vertical profiles temp, PAR, refPAR, LuChl and kPAR):
  - Correction of time for stations MC0, MC1 and MC2 on 9/7/1997.
  - Every record corresponds to a certain time interval.
- RAS floating radiometer
- CTD data
  - IDOD is interested in down-cast after pump stabilisation with suitable smoothing and despiking every 0.5 m. First, only profiles executed in the frame of AMORE and MUMM monitoring will be stored in a separate table 'CTD profile'. The marked files will be used to add the information with point samples.

1.4. The biochemistry of nutrients, metals, and organic micropollutants in the North Sea

1.4.1. Data set from Pr. R. van Grieken (K. Eyckmans, UIA – Centrum voor Micro- en Spore analyse)

1.a. Availability of the data set	YES
Date of delivery	03/05/1999
File name	IDOD data.doc

1.b. Availability of the meta-information	General	YES
	Methods	Not detailed*
	Quality Control	NO
	Meteorology	ODAS

1.c. Availability of the expected parameters	Pr. R. van Grieken (UIA)	No data expected (1997) Incomplete (1998)

1.d. Quality of the data set	Significant figures	To check
	Description	Incomplete*
	Unit	YES
	Detection limit	NO*
	Quantification limit	NO*
	Expected accuracy	NO*

*These data have been discussed during and will be sent by Kurt Eyckmans.

Expected / received* parameters:

For air:

- Nutrients * : NO_2^-, NO_3^-, SO_4^{2-}, PO_4^{3-}
- Major inorganic *(qualitative): F^-, Cl^-, sea salt, gypsum, alumino-silicate, Si
- Metals * (qualitative): Ag, Cu, Zn, Fe

Questions / comments:

The data were discussed during the visit held at UIA – Centrum voor Micro- en Sporeanalyse (20/08/99). A report with more details on the dataset can be found in the annex.

Research with respect to the methodology is an important component during the first years of the project. In 1998, only qualitative results were obtained using EPMA for the analysis of individual particles. The results for 1999 are all quantitave, another method is used.

Complement of data set will be sent in January!

1.4.2. Data set from Pr. W. Baeyens (K. Parmentier, VUB – Laboratorium voor Analytische Chemie)

1.a. Availability of the data set	YES
Date of delivery	10/11/1999
File name	OVER982.xls

1.b. Availability of the meta-information	General	Incomplete
	Methods	NO
	Quality Control	NO
	Meteorology	ODAS?

1.c. Availability of the expected parameters	Pr. W. Baeyens	Complete? (1997) Incomplete (1998)

1.d. Quality of the data set	Significant figures	To check
	Description	Incomplete*
	Unit	YES
	Detection limit	NO
	Quantification limit	NO
	Expected accuracy	NO

Expected parameters:

For water:

- Nutrients: NO_2^-, NO_3^-, SO_4^{2-}, PO_4^{3-} , ammonia, urea
- Metals: Cu*, Cd*, Ni*, Zn*, Pb*, As, Hg, monomethyl Hg, total Hg
- Major organic: organic nitrogen and carbon

1.4.3. Data set from Pr. H. Van Langenhove (Tom Huybrechts, UG – Laboratorium voor Organische Scheikunde)

1.a. Availability of the data set	YES
Date of delivery	30/03/1999
File name	RES97-1.xls

1.b. Availability of the meta-information	General	YES
	Methods	To receive*
	Quality Control	To receive*
	Meteorology	YES

1.c. Availability of the expected parameters	Pr. H. Van Langenhove (RUG)	YES (1997) NO (1998)

1.d. Quality of the data set	Significant figures	To check
	Description	YES
	Unit	YES
	Detection limit	To receive*
	Quantification limit	Not available
	Expected accuracy	To receive*

*These data have been discussed during the meeting held at UG and will be sent by Tom Huybrechts.

Expected / received* parameters:

For water:

- Aromatic hydrocarbons:
  - benzene*, toluene*, m/p/o-xylene*, ethylbenzene*
- Organochlorines:
  - chloroform*
  - tetrachloromethane*
  - 1,1-dichloroethane*
  - 1,2-dichloroethane*
  - 1,1,1-trichloroethane*
  - trichloroethylene*
  - tetrachloroethylene*

Questions / comments:

The visit at the laboratory (06/05/1999) has permitted us to receive more information about this data set: sampling and analyses methods, quality control (QUASIMEME), date of analyses. Data about the Scheldt estuary will be sent soon. A report is available in annex for more details about this meeting.

1.4.4. Data set from Pr. R. Wollast (N. Roevros, ULB – Laboratoire d’Océanographie Chimique)

1.a. Availability of the data set	YES
Date of delivery	31/03/1999
File name	9809ulb.wk4 9725ulb.wk4 9716ulb.wk4

1.b. Availability of the meta-information	General	YES
	Methods	YES
	Quality Control	To receive*
	Meteorology	YES

1.c. Availability of the expected parameters	Pr. R. Wollast (ULB)	Incomplete (1997) Incomplete (1998)

1.d. Quality of the data set	Significant figures	To check
	Description	YES
	Unit	YES
	Detection limit	To receive*
	Quantification limit	To receive*
	Expected accuracy	To receive*

*These data have been discussed during the meeting held at ULB and will be sent by Nathalie Roevros.

Expected / received* parameters:

For water:

- Physical:
  - temperature
  - suspended matter*
- Metals:
  - Cu*, Cd*, Ni, Zn, Pb*, Mn*, Co*, Cr*, Al*, Ca*, Fe*, Si* (* sometines missing, but will be available soon)
- Major organic:
  - organic nitrogen (total nitrogen*)
  - organic carbon*
  - total carbon*
- Pigments:
  - chlorophyl-a*
  - phaeopigment*
- Particulate phase of water column:
  - profile of the pore*
  - specific surface*
  - average pore radius*

Questions / comments:

The visit at the laboratory (11/05/99) has permitted us to receive more information about this data set: time, water depth, sampling and analyses methods, quality control (QUASIMEME), date of analyses, data missing for metals. These data will be sent by Nathalie Roevros. Some results are not yet available but will be sent when ready. A report is available in annex for more details about this meeting.

Water:

- Major organic: Nitrogen: total nitrogen instead of organic nitrogen because concentration in inorganic nitrogen is very small, so total nitrogen can be representative of organic nitrogen.
- Major inorganic: in the *IDOD* Newsletter #2, it was noted that this laboratory would provide us with data about major inorganic but they do not measure these parameters (in consequence, this reference has to be withdrawn from the list of expected parameters).
- Chlorophyll-a has to be added in the list of expected parameters.

1.5. *ICAS – The Impact on North Sea organisms of pollutants Associated with Sediments*

PDM: Pol Gosselin (UMH - Laboratoire de Biologie Marine)

1.5.1. Data set from Dr. Ph. Dubois (P. Gosselin)

1.a. Availability of the data set	YES
Date of delivery	31/03/99
File name	Results.xls MatMet.doc

1.b. Availability of the meta-information	General	YES
	Methods	YES
	Quality Control	YES
	Meteorology	Not available (see ODAS)

1.c. Availability of the expected parameters	Dr. Ph. Dubois	No data expected (1997) Incomplete (1998)

1.d. Quality of the data set	Significant figures	To check
	Description	YES
	Unit	YES
	Detection limit	YES
	Quantification limit	YES
	Expected accuracy	Not known

<u>Expected / received* parameters:</u>

<u>For water:</u>

- Physical: water depth*
- Major inorganic: salinity*

<u>For sediment:</u>

- Physical: granulometry*
- Metals: Cd*, Cu*, Hg, Pb*, Zn*

<u>For biota:</u>

- Sea urchins: embryotoxicity and metamorphosis tests
- Starfishes:
  - Metals: Cd*, Cu*, Hg, Pb*, Zn*
  - Biological effects: MFO activity, amoebocyte RO^2 species, embryo-toxocity test*, amoebocyte phagocytic activity*, metallothioneins

Questions / comments:

The visit at the laboratory (03-04/06/99) has permitted us to receive more information about this data set:

No data available for 1997 (no campaign), except the results of test experiments with larvas in known metals solutions.

In the expected parameters list, Hg has to be deleted (for sediment and biota) because it's not analysed. Granulometry is measured by ULB only.

Analysis on sea urchins are made by the UMH (Pr. M. Jangoux); so the list of parameters has to be corrected. The same for embryotoxicity test in starfishes.

Some data are missing in the data set: metal concentrations in sediment and types of development of sediment-exposed embryos are not available when the dry weight percentage of that fraction is too small to allow analysis.

Some results are not yet available due to the time needed for analysis: metal concentrations in starfish organs will be sent when ready.

Data concerning phagocytic activity of amoebocytes are available when starfishes are present at the station (during trawling). Other biological effects (MFO activity, amoebocyte reactive oxydase species, metallothioneins) will be considered in 1999 or 2000.

1.5.2. Data set from Pr. M. Jangoux (P. Gosselin)

1.a. Availability of the data set	YES
Date of delivery	31/03/99
File name	Results.xls MatMet.doc

1.b. Availability of the meta-information	General	YES
	Methods	YES
	Quality Control	YES
	Meteorology	Not available (see ODAS)

1.c. Availability of the expected parameters	Pr. M. Jangoux	No data expected (1997) Incomplete (1998)

1.d. Quality of the data set	Significant figures	To check
	Description	YES
	Unit	YES
	Detection limit	YES
	Quantification limit	YES
	Expected accuracy	Not known

Expected / received* parameters:

For water:

- Physical: water depth*
- Major inorganic: salinity*

For biota:

- Sea urchins: embryotoxicity and metamorphosis tests
- Starfishes:
  - Biological effects: embryotoxocity test*

Questions / comments:

Metamorphosis tests for 1998 in sea urchins will be sent when ready. Measures on embryotoxicity in sea urchins will begin in 1999.

1.5.3. Data set from Pr. R. Flammang (P. Gosselin)

1.a. Availability of the data set	YES
Date of delivery	31/03/99
File name	Results.xls MatMet.doc

1.b. Availability of the meta-information	General	YES
	Methods	YES
	Quality Control	YES
	Meteorology	Not available (see ODAS)

1.c. Availability of the expected parameters	Pr. R. Flammang	No data expected (1997) Incomplete (1998)

1.d. Quality of the data set	Significant figures	To check
	Description	YES
	Unit	YES
	Detection limit	YES
	Quantification limit	YES
	Expected accuracy	Not known

Expected / received* parameters:

For water:

- Physical: water depth*
- Major inorganic: salinity*

For sediment:

- PCBs: 28*, 52*, 101*, 118*, 138*, 153*, 180*, 37, 77, 81, 126, 169

For biota:

- Starfishes:
  - PCBs: 28, 52, 101, 118, 138, 153, 180, 37, 77, 81, 126, 169

Questions / comments:

Some results are not yet available due to the time needed for analysis: PCBs measurements will be sent when ready. Some PCBs data in sediment are also missing.

1.6. MARE-DASM – Marine Resource Damage Assessment and sustainable Management of the North Sea

PDM: Serge Scory

1.a. Availability of the data set	NO

Comment:

This project started in November 1998 and, to our knowledge, did not collect any data in 1998.

Reports of the meetings between IDOD and the data provi-ders for the Programme "Sustainable Management of the North Sea".

1. Meeting UG, Vakgroep Organische chemie

Held on May 6th, 1999
Present : Tom Huybrechts, Siegrid Jans, Karien De Cauwer

1.1. Data set

Normally the laboratory participates in two or three campaigns per year in the North Sea (RV Belgica) and two or three campaigns per year at the Scheldt estuary with a vessel from the Centre for Estuarine and Coastal Ecology (Yerseke). The Scheldt estuary is sampled within a short period after a campaign at sea.

Besides these sampling campaigns, more samples are taken at station 330 by other laboratories (tentatively weekly).

Due to instrumental problems only the results of one campaign are given. Results for the Scheldt estuary will be sent soon.

1.2. Meta-information

1.2.1. General

Sampling depth is more or less –5 m. The sampling depth is not recorded. Research (by J. Dewulf) on vertical profiles showed that the sampling depth has no influence.

1.2.2. Methods

The standard operational procedures will be sent. A short summary is given:

Sampling method
One Niskin bottle (5 l) is taken at every station. Special conditions are recorded : someone painting, a RIB passing near the boat, oil pollutions, ...

Sample handling
With a silicion tube, two green bottles (750 ml) are filled. The bottles were rinced with destillated water, oven dried and closed with teflon. HCl was added (pH = 2) to prevent any microbiotic activity. The bottles are completely filled with water to avoid air interaction and evaporation of substances. Known amounts of three substances are added to verify the retreival by the analysis method. The samples are stored at 4° C. Only one bottle is used, the other is only used when the results are not good, when something went wrong.

Analysis method

Gaschromatography is used to analyse the samples. The purging of the sample takes place at a different instrument with which it is possible to reduce the injection of water. Before the analysis, IS (internal standards) are added for calibration.

The date of analysis is available. Studies showed that 100% of substances were retreived after two years of storage.

1.2.3. Quality control

Participation in Quasimeme exercises (normally 2 times/year). Results are very good for 4 compounds. The results will be sent.

Control chart

Addition of known amount of three substances 'surrogaten' : % recovery.

Occasionally, the two bottles are analysed to have information on the reproducability of the analysis.

1.2.4. Meteorology

The values are extracted from the list of ODAS-measurements every 10 minutes. When other laboratories take samples at station 330, this meta-information is often missing.

For the samples gathered on board of the vessel from the Centre for Estuarine and Coastal Ecology (Yerseke), this meta-information is missing.

1.3. Parameters

For the moment only parameters in water samples were measured. Air samples are envisaged for the next campaign in June.

Sediment samples won't be taken in the near future, maybe later but only in the Scheldt where higher O.M. contents are found. (J. Dewulf found low concentrations in sediment of the North Sea.)

1.4. Quality

Number of significant figures

This information will be sent.

Description

Parameters are described by their IUPAC name. No codes are used.

SG : *surrogaat*

IS : internal standard

Detection limit

A detection limit is normally determined per campaign. A blanco is analysed, DL is the signal-to-noise ratio multiplied by three. For compounds which are

present everywhere, the DL is augmented by the signal of the blanco.
Quantification limit
Not used

Expected accuracy
This information will be sent.

1.5. *Question/Comments*

Samples
Always one Niskin bottles is taken. Sometimes the two subsamples are analysed. The samples can be stored for a long time before analysis. The samples are conserved until the data have been verified.

The data on VOC will be compared with the data on planktonic activity measured by data ulb-gmma.

1.6. *Actions*

MUMM :
To send *IDOD* Newsletter #1
To send data sets screening procedure

UG :
To send information on Scheldt estuary
To send information on quality control (quasimeme scores, thesis, control chart, quantity of *surrogaat* added)
To send information on methodology (and detection limit, accuracy)
To send date of analysis

2. Meeting ULB, Laboratoire d'Océanographie chimique

Held on May 11th, 1999
Present: Nathalie Roevros, Karien De Cauwer, Siegrid Jans

2.1. Data set

The laboratory participates in two campaigns per year in the North Sea (RV Belgica).

Due to delay in the analyses (due to other project priorities), some data are missing in the tables. They will be sent when ready. The data about the transect 98/9 will also be sent.

2.2. Metadata

2.2.1. General

Some metadata were missing in the first document sent by ULB. These data are available and will be sent soon: time of sampling, water depth, quality control information, (meteorology: data available from ODAS). The positions given in the tables are the real ones (not reference positions).

2.2.2. Methods

A description of the applied methodology (for sampling and analysis) has been given.

<u>Sampling method</u>
For metal: surface water (-1m) is collected with the Zodiac (by VUB), in polyethylene bottles, previously decontaminated.

For chlorophyll and major organic: surface water is collected with the Zodiac, or Niskin bottles. A test is in progress for water samples taken directly from tap.

For the analysis of the particles of water column: water is collected at –3m by the piping of the Belgica and immediately centrifuged.

Special conditions such as bad weather or dredging activities are recorded.

<u>Sample handling</u>
For metal analyses: water is filtered by VUB in the 'clean lab' with polycarbonate filters (0.4 µm – weight known), which are directly put into the deep freezer until analysis. The MES quantity is evaluated by weight difference of filters before and after filtration (deep-freeze filters are previously dried).

For major organic analyses: water is filtered with fibreglass filters, previously warmed up (500°C) during 2 hours (to destroy organic matter). Filters are kept into the deep freezer.

For chlorophyll analyses: water is filtered with fibreglass filters which are put in aluminium paper before being stored in the deep freezer.

For particle analyses: the collected matter (by centrifugation or filtration) is stored into the deep freezer.

Analysis method

For metal analyses: the digestion process consists in adding acids to the SM in order to degrade the different compounds (silicates, carbonates, alumino-silicates, organic matter,...). To accelerate this process, the samples are placed in clossed vessels and submitted to high pressures ("Closed vessel acid digestion with micro-..."). The resulting sample is analysed by spectrometry (electrothermic, ICP- Plasma, flame depending on substances concentration).

For major organic analyses: pieces of filters are submitted to a flash combustion (in order to isolate the substances) and analysed by gas chromatography.

For chlorophyll analyses: acetone is added to the filters (during one night) and the resulting samples are submitted to fluorimeter.

For particle analyses: the collected matter is carefully rinsed, dried and crushed. The resulting sample is placed in the "High Speed Surface Area and Spore Size Analyser", which measures the specific surface, the average pore radius, and the pore profile.

The date of analysis is available and will be sent. The chlorophyll samples have to be analysed relatively quickly after sampling. The other samples can be kept in the deep freezer without deterioration.

2.2.3. Quality control

This laboratory has participated in Quasimeme exercises but due to lack of time and money, no exercise has been done last year. The previous results will be sent.

Three measures are done for each sample to analyse one parameter.

Some reference substances are analysed during all the processes/methods in order to check the quality of the results.

Some tests are in progress in order to compare results given by centrifugation with those given by filtration.

2.2.4. Meteorology

All data are stored in ODAS. No meteorology measurement is done by the laboratory. Data available (from ODAS) will be sent.

2.3. Parameters

Major inorganic parameters (for water: t°, salinity, dissolved oxygen, pH, carbon) are not measured by this laboratory, so this category has to be withdrawn in the expected parameter list (cfr. *IDOD* Newsletter #2) for ULB-Wollast.

The data concerning the particles will be sent. The profile of the pore and specific surface is only measured for two campaigns. Later on only one or two samples will be analysed for verification.

2.4. Quality

Detection and quantification limits will be sent.
Expected accuracy is around 2-5% for metal analysis. For other parameters, information will be sent.

2.5. Questions/comments

Samples are stored after analysis in order to be able to make more analyses or to check some values (even few years later).

2.6. Actions

ULB:
To send metadata (time, water depth, time of analysis, (meteorology)).
To send new data when analyses are done.
To send data about the particulate phase of water.
To send information about quality control (Quasimeme, detection limit, quantification limit, expected accuracy).

3. Meeting UMH, Laboratoire de Biologie marine

Held on June 4th, 1999
Present: Pol Gosselin, Pascale Wantier, Karien De Kauwer, Siegrid Jans

Meeting ULB, Laboratoire de Biologie marine

Held on June 5th, 1999
Present: Geoffroy Cotteur, Philippe Dubois, Siegrid Jans

3.1. Data set

The laboratory participates in one or two campaigns per year. In 1997, any campaign has been done; in 1998, one campaign has been undertaken and in 1999, the laboratory has already participated to two campaigns. In addition to these activities, some samples are directly taken from the beach (by hand).

Due to the difficulty of the setting up of the techniques and due to the time required for analysis, some parameters have not been considered in 1997 and/or in 1998; moreover some analyses are not yet finished, data will be available when ready.

Test experiments on sea urchins and starfishes are to be considered as a special data set. They are made in order to observe the “normal” behaviour of larva and to determine/estimate the reaction due to a known concentration of contaminants. These results have to be introduced in the database as they are necessary for the interpretation of the complete data set.

3.2. Metadata

3.2.1. General

Metadata missing (time and meteorology) are not collected directly by the laboratory. However, these data can be found in the ODAS database. Concerning the trawling samples, duration and end position are not recorded.

3.2.2. Methods

A description of the methodology had been sent with the data set and is well detailed.

The date of analysis is sometimes available and will be sent soon. When it’s not available, the maximum delay between the sampling date and the analysis date is given.

Individuals used for embryotoxicity tests are those resulting from breeding. They are not related to the sample site. On the other hand, metal and PCB concentrations are measured in individuals taken at each station (if present). Extra individuals are necessary for the analysis of phagocytic activity. Thus, when there is not enough individuals, the analysis is not possible.

The different samples of sediment are taken from three boxcorers at each station. However, these three samples are recorded as one sediment sample.

3.2.3. Quality control

These laboratories don't know the QUASIMEME inter-comparison exercises. However, they analyse control referenced material (CRM).

Some standard measurements are made for each parameter in order to have a reference to compare the results obtained in the natural environment.

3.2.4. Meteorology

As said before, information about meteorology is not available from these laboratories, but it can be found in the ODAS database.

3.3. Parameters

The list of the expected parameters has to be reconsidered:

Hg is not measured by any laboratory (to be deleted from the list).

Measures concerning granulometry are made by ULB only.

Some data concerning PCBs (PCBs 37, 77, 81, 126, 169 and PCBs in starfishes) and metal concentrations in organs are missing due to the time needed to analyse them. These data will be sent when ready.

Data pertaining to metal concentrations in sediment and to type of developments of embryos are not available when the sediment fraction is too small to allow analysis. This does not mean that the data set is incomplete.

The results for metamorphosis test on sea urchins will be available for the year 1998. Embryotoxicity tests will begin in 1999.

No analysis has been done in 1997-1998 concerning MFO activity and amoebocyte reactive oxydase.

Metallothionein analysis are not expected to begin before 2000.

Other data concerning biological effects are not always available because of the slight quantity of individuals sampled (for some stations).

3.4. Quality

Detection limit (constant) is given in the methodology report. Quantification limit is considered the same as the detection limit.

Expected accuracy is not known and is difficult to be estimated. Moreover, the error due to analysis procedure (equipment and manual errors) are considered to be small compared to natural variability.

To correct for errors due to the losses resulting from pre-treatment, a known concentration of PCB 103 (not present in the natural environment) is introduced artificially in the samples. The PCB concentrations are re-calculated taking into account the percentage recovery of the PCB 103.

3.5. *Questions/comments*

Samples of sediment are stored after analysis for a long time period. Organs of starfishes are destroyed during analysis. Normally, all the samples will be eliminated at the end of the project because there is no need of keeping information on the stored samples.

Concerning the position of samples (trawling and beach stations), a decision has to be taken in order to standardise the geographic representation.

The Newsletter2 will be forwarded to each researcher of the laboratory in order to have some comments about the ‘requests to the IDOD database’.

3.6. *Actions*

UMH:

To send data concerning date of analyses.
To send missing data when ready.
To forward the Newsletter#2 to persons involved in the project.
To check the Quasimeme code list.

4. Meeting UG, Sectie Mariene Biologie

Held on 10th of March
Present : Andre Cattrijse, Mia Devolder, Karien De Cauwer

4.1. *Dataset*

The laboratory participates in one monitoring campaign in October (when benthos biomass and density are maximum) and in two campaigns (spring and autumn), during which the Flemish banks are intensively sampled, to study the structural biodiversity. During spring 1999 weekly samples are taken at three stations (330, 215 and near 115) to study the functional biodiversity.
The final set of epi- and hyperbenthos data for 1997 and 1998 will be available soon. Meiobenthos and macrobenthos is studied by other persons, the delivery date is unknown. Macrobenthos results of the westcoast will be supplied soon. Stomach analysis results will be provided by the beginning of September.
For data of KUL-Ollevier, Els Gysels can be contacted. Information on seabirds should be asked for at the '*Instituut voor Natuurbehoud*'.

4.2. *Meta-information*

4.2.1. General

The sampling date for all water samples is available. For the benthos samples, the same date and position is valid. The sampling time can be retreived via outprints of ODAS. The time is not really important, but mainly the period of the day : day/night. Sampling depth is the water depth, stored in CTD-profiles.
Position
When the banks are sampled for meio- and macrobenthos, twenty points are randomly selected along transects perpendicular to the bathymetric lines. For epi- and hyperbenthos 6 tracks per bank are done. Always the same name is given e.g. KWI1, KWI2, KWI3, ... but the position is not the same ! During a track, begin- and endposition is recorded.
It is important to know the exact position and as such also the datum (WGS84 or DG50) used.

4.2.2. Methods

A description of the methodology will be send for benthos, nutrients, pigments, sediment and stomach analysis.
Benthos
Sampling method : samples are taken with Van Veen, Reineck box core, hy-

perbenthic sledge (not standard, 3 m width, subdivision in sampling depths : 0-0,5 m and 0,5 m to 1 m) and plankton net.

Hyperbenthos should be sampled during day-period. Normally a distance of more or less 1000 m is sampled with the hyperbenthic sledge, so a surface of approximately 3000 m².

Sample handling : benthos samples are sorted and rinsed on board of the Belgica. Meiobenthos samples are first subdivided in slices. The fauna is fixed in formol. Important to record is whether the samples are fixed before or after rinsing.

Analysis method : the samples are analysed after minimum one month so that the shrinking process is complete. From the length-frequency distribution, the weight frequency distribution is calculated via a length-dry weight regression. Herefore the mean of the length class is used.

4.2.3. Quality control

The laboratory does not participate in QUASIMEME exercises nor analyses internal or certified reference material. This, as well as the availability of control charts should be checked with the person responsible for those analyses. The relative values are of main importance to study the relation between nutrients and fauna.

4.2.4. Meteorology

All relevant information on meteorology is stored in ODAS.

4.3. Parameters

Temperature, salinity, water depth, PAR and OBS is supplied by the vertical profiles available at MUMM-Oostende.

Pigments might not have been measured in February 1998 (to be checked).

Sediment (0,4 till 1000 µm): median correl size, carbonate, total organic matter, nutrients.

4.4. Quality

Description

Length is divided in 4 mm classes (the label does not appear correctly in the file).

Chla(vis) : visual

Chla(flu) : fluorimetric

Number of significant figures

Detection limit/quantification limit

Expected accuracy

To be checked with person responsible for nutrient and pigment analysis.

4.5. *Question/comments*

Date of analysis is not recorded for benthos. The availability of analysis date for nutrients should be checked.
Stomach analysis is executed by UG, not KUL-Ollevier as mentioned on the list.
Samples
During monitoring campaigns 5 replicas are taken for meio and macrobenthos as the sediment often shows a high local variation. This means five different Van Veen or Reineck box core samples. The results of benthos and granulometry (mainly grain size median) are taken on the same sample. KUL uses the same samples to study the genetic variation.
All benthos and sediments samples are stored for ever. The laboratory plans to make a list of all stored samples.

The calculated densities are the main values of interest to be stored in the database. Besides this, it should be decided whether the densities of every replica will be stored or the mean density of all replica's. The mean density is of most value, but other densities will supply information on the heterogeneity of the sediment. Together with mean densities, the number of replicas on which it is based, should be mentioned. Three replicas are considered as a minimum for meiobenthos.

Requests to the IDOD database : most important functions to be provided are regression and comparison of means.

For the moment it is not know whether the epibenthos samples are big enough. A test has been done, with three replicas, the results are not yet analysed. This could be an topic for KUL-UCS.

4.6. *Actions*

<u>MUMM</u>
To send data sets screening procedure.
To send information on QUASIMEME and the code lists.

<u>UG</u>
To send a description of methodology for benthos, nutrients, pigments, sediment and stomach analysis.
To send the report on benthos and avifauna (March, 1999).
To send a preliminary data set of meiobenthos.
To send the final data sets of benthos, sediment and stomach analysis together with meta information (date, time, position, eventually date of analysis) when finalised.
To check for availability of analysis date and quality information on nutrient analysis (see 4.4.2.3 & 4.4.4).
To check for availability of pigment results in February 1997.

5. Meeting ULg, Laboratoire d'Océanologie & Service d'Anatomie pathologique

Held on June 22nd, 1999
Present : Virginie Debacker, Thierry Jauniaux, Serge Scory, Karien De Cauwer, Siegrid Jans

5.1. *Data set*

The "samples" are not collected during campaigns at sea. Mammals and birds are usually found on the beach. Some mammals are catched accidentally by fishermen.

5.2. *Meta-information*

5.2.1. General

The sampling time is not available. It is not important for this type of data. The exact position is not available (except for certain mammals accidentally catched by fishermen). Only the name of the site is given.

5.2.2. Methods

The standard operational procedures will be sent.

<u>Sampling method</u>

For birds: the birds are collected by the *'Instituut voor Natuurbehoud'* and sent to the ULg laboratory. The birds are put in plastic bags with identification number and information on the sampling place. Another number (autopsy number) is given to every specimen by Ulg.

For mammals: a protocol is established in order to clarify the procedure to follow when arriving on the site. Animals lying on the beach are either directly autopsied in situ (large whale) or transported to the laboratory (max.: 6 meters long). Mammals catched by fishermen are sometimes "sent" to the laboratory for autopsy.

<u>Sample handling</u>

For birds: the birds are stored in a deep-freezer (-18°C). When they are autopsied a sample of each organ (whole liver, whole kidney, part of muscle) is taken. Half of these samples are sent to VUB and the other part is put into the deep-freezer until analysis.

For mammals: animals are either directly autopsied, either frozen till the autopsy.

The tissues intended to virology analysis and to cellular cultivation are stored at –80°C. The others are stored at –18°C: they are for toxicology analysis (Ulg, Oceanology), for micro-biology (bacteriology and parasitology laboratory), for virology (CERVA, Brussels). For histopathology (Ulg, “Anatomie pathologique”), tissues are stored in formalin.

The date of autopsy is noted. When mammals are fresh, the autopsy is done the day of grounding.

Analysis method

Details of the necroscopy of Sperm whales are given in the “protocol report”. “As a general rule, all organs must be carefully examined and main parenchymas should be sliced... Lesions must be described..., sampled and photographed with a ruler indicating the size. In addition to lesions, some tissues are systematically collected for histopathology, parasitology, bacteriology, virology, toxicology and life history...”. (Jauniaux, T., et al., Postmortem examination and tissues sampling of sperm whales Physeter macrocephalus).

During the autopsy, some pictures are made for each specimen in order to remember all details of the pathologies. These slides are stored and available at the service of “Anatomie pathologique”.

The service begins analysis when there are enough specimen available.

The analysis date is not noted by the “laboratoire d’océanologie”. For heavy metals a standard method is used as well as for lipids (colorimetric extraction). The method to analyse selenium is fluorometry (method used since 5 years). The method used to analyse metallothioneins is new and in development.

5.2.3. Quality control

The laboratory participates to Quasimeme exercises: results for metals in biota will be sent for 1998.

Remark: one method used by this laboratory (“bain-marie”/double boiler) is not listed in the Quasimeme list.

Certified Reference Material is also used: DORM-2 (Dogfish muscle CRM for trace metals) is the most common. Obtained results will be sent.

5.2.4. Meteorology

No meteorological information is available. This kind of data is not important in this case.

5.3. Parameters

For birds:

- No Selenium analysis
- Metallothionein concentrations: not yet available
- Total lipids: for livers and muscles only (not for kidneys)

For mammals:

- Selenium analysis: for kidneys and livers only (not for muscles)

- Metallothionein concentrations: not yet available
- Total lipids: for livers and muscles only (not for kidneys)
- Pathology and parasites: these data are available but there is a need to chose a pathology code in order to simplify the entry in the database

PAHs: no analysis from now

5.4. *Quality*

Detection limit

The given value represents the mean value; a value for each serie of analysis is available and is related to the viscosity of the sample. All detection limits will be added in the data set, although the variation is small.

A standard sample is analysed for calibration and a sample containing a mixture of all samples is analysed to determine the range of concentrations.

Quantification limit:

The quantification limit is 5 times the detection limit.

Expected accuracy

To receive from V. Debacker

5.5. *Question/Comments*

The area considered for mammal samples by the ULg laboratory is extended to "la Baie de Somme".

Stockage:

Birds: samples for liver and kidney are just big enough to perform all analyses. Consequently, those samples are only stored till analysis.

Mammals: samples are stored for a long time and a list is kept with all the frozen samples.

Parasites: the parasites found on animals are stored in alcohol.

Descriptive information:

Some extra descriptive parameters are important to understand the contaminant concentrations. These are: weight, cachexy, sex, age, species, oil coverage (Y/N), real indicator of the environment (classes giving the indication about the cause of mortality).

5.6. *Actions*

ULg :

- Send analysis methodologies (toxicology: heavy metals, lipids, metallothioneins, selenium/ Virology, bacteriology, parasitology, virology and histopahtology) + Comparison between the Quasimeme list of methods and those used in each laboratory

- Send results for Quasimeme exercices and CRM
- Send expected accuracy data
- Send results of pathology and parasites for mammals (list and autopsy reports)
- Send complete results for birds (including identification number from IN)
- Send metallothionein results when ready

6. Meeting ULB – ESA, Ecologie des Systèmes Aquatiques

Held on July 8th, 1999
Present: Véronique Rousseau, Karien De Cauwer, Siegrid Jans

6.1. *Data set*

Data are collected tentatively weekly from February to mid-December at station 330, during January and summer months sampling is less frequent (once or twice a month). Station 330 is used as reference station as it represents the average physical condition for the Belgian Continental zone.

6.2. *Meta-information*

6.2.1. General

The position of station 330 is MUMM's reference position and not the real one.

Sampling time: for logistical reasons, sampling time is independent of the tidal cycle and occurs when convenient for the scientific team. For samples collected with the Tuimelaar, the time is often missing. Sampling usually happens around noon.

6.2.2. Methods

The surface water samples are collected with an oceanographic bucket (0 m), one sample per station, and stored in decontaminated PE flasks on board. These flasks are preserved in thermostatic boxes until treatment in the ESA laboratory in Brussels.

The description of the sampling processing and analysis is given in a short report made by ULB for each parameter. For counting of phytoplankton, 10 or 50 ml of seawater is concentrated.

Nutrients and salinity are analysed by BMM-Oostende since 1997. Before that year, only nitrate + nitrite was analysed by BMM-Oostende.

6.2.3. Quality control

The consistency of the data is checked internally and is used as quality control by comparing the data amongst themselves and with existing other parameters. It is based on a comprehensive knowledge of the studied ecosystem, eco-physiology of phytoplankton. Samples are kept until consistency of the data is checked (phytoplankton samples are kept for more or less 3 years).

Concerning nutrients, analyses are made by BMM-Oostende, who participates in QC and QA exercises (Quasimeme exercises). For phytoplankton analysis, ringtests are done. 400 cells are necessary to have a precision of 5

%. ULB-ESA will contribute to the SGQAE (Steering Group on Quality Assurance of Biological Measurements Related to Eutrophication Effects) of ICES.

6.2.4. Meteorology

Meteorology information is not available from ULB-ESA; the IDOD team will recover them from the ODAS database.

6.3. *Parameters*

For nutrients:

- Nitrate and nitrite data will be sent as separate parameters.
- Suspended matter: there is no data available for 1997 because of the use of unsuitable filters. Data for 1998 (analysed by using polycarbonate filters), will be sent.
- Data concerning bacteria and protozooplankton will be sent for 1998 (no analysis for 1997).

The data set from VUB will be completed for 1998. There was no team on board in 1997, so there is no data to expect for that year.

6.4. *Quality*

Detection limit

A detection limit is given for each parameter (except for salinity, chlorophyll-a and phytoplankton).

Quantification limit

Not used

Expected accuracy

An accuracy of 5-10% is expected for the counting of phytoplankton cells and of 9% for chlorophyll-a. For nutrients, see BMM-Oostende.

6.5. *Questions/Comments*

ULB, VUB and IDOD team think it should be interesting to hold a meeting to discuss the storage of derived data. Indeed, some data are the result of a modelisation developed by the laboratory , involving subjective factors that are continuously subject to changes. To interpret these data, background information is a prerequisite. The question arises what should be stored and how ?

e.g.: Rough data = cell number of a species.
Derived data = carbon biomass (calculated with a personal conversion factor)

Rough data = chlorophyll-a
Derived data = primary production (based on a model which implies different processes such as light intensity, phytoplankton biomass, turbidity, nutrient availability...)

As remark for the quality control procedures developed by UCS : results for ammonium (MUMM) between 1977 and 1982 are not reliable.

6.6. *Actions*

ULB

To send 'suspended matter data' (for 1998)

To send nitrate and nitrite data separately

To send data concerning bacteria and protozooplankton + documentation (for 1998)

To collect and forward to IDOD data sets from VUB

To start recording analysis date

VUB

To send the complete 98 data set (no cruise in 1997) to Véronique Rousseau

IDOD

To recover water depth and meteorological data from ODAS

To organise a meeting between the three teams to discuss the problem of modelled data

7. Meeting UIA, Centrum voor Micro- en Sporenanalyse

Held on August 20th , 1999
Present : Kurt Eyckmans, Siegrid Jans, Karien De Cauwer

7.1. Data set

All data on particles in air samples are gathered on board of the RV Belgica. In 1997 few data were gathered during the monitoring of the Scheldt estuary, which were mainly measured for a first screening and testing of the methodology. In 1998, the laboratory participated in two campaigns, in 1999 three cruises are scheduled. In 1998, only qualitative results were obtained using EPMA for the analysis of individual particles. The results for 1999 are all quantitative, another method is used. Rainwater samples are collected since 1999. Research with respect to the methodology is an important component during the first years of the project.

7.2. Meta-information

7.2.1. General

The samples are taken more or less 13 m above the sealevel. The precise height is not measured and not important when studying air masses. Date, time and position are recorded at the start and end. Sampling always takes place in tracks with a duration of several hours (e.g. 12 hours). During these tracks the speed and direction of the ship can vary, for instance other laboratories take samples when passing a station (the ship should always arrive at the station with the front in the direction of the wind).

7.2.2. Methods

A complete description of the applied methodology will be sent.

Nutrients (Cl-, nitrate, nitrite, phosphate, sulphate, and potassium, Na, ammonia)

Sampling : 0.4 µm nucleopore filters and pump (type ?).

Sample handling : The particles are dissolved by using an ultrasonic bath or in a loop. The bio-available fraction is measured.

Analysis method : IC (Ion chromatography)

Individual particle analysis :

Sampling : 0.4 µm nucleopore filters and pump.

Sample handling : ?

Analysis method : EPMA (Electron Probe X-ray Microanalysis) was used in 1998. 400 particles are analysed. Every particle is irradiated during 20 s. The frequency of the returned radiation gives information on the composition. A spectrum is obtained for every individual particle. With a clustering program, similar spectra are grouped e.g. 200 spectra of sea salts and marine crystallisation products (the latter are older : chlorine replaced by sulphate), 50 spectra of Fe-rich particles (indicating continental origin), ... No clear and quantitative information on the composition can be given as the mass and size of the particles are not taken into account. In 1999 this method is not used anymore and quantitative information will be available.

Metals

Sampling : Nucleopore filter (polycarbonate) and pump

Sample handling : (not dissolved)

Analysis method : XRF (X-ray fluorescence) (a faster method but not as good as ICP)

ICP will be used for metals in the near future. Hereford another type of filter is necessary and the particles will be dissolved in an ultrasonic bath.

7.2.3. Quality control

No intercomparison exercises and control charts are executed.

Always an extra sample is taken, to verify the result. The average is given.

7.2.4. Meteorology

The meteorological information in the data set is derived from ODAS. Those parameters are recorded more or less half way the track (the time-stamp is noted) or at start and end time. It is up to IDOD to decide how this information will be stored.

7.3. Parameters

For 1998 the data set consists of qualitative results for : sea salt, marine crystallisation products, gypsum, alumino-silicates, Si-rich, Fe-rich, organic, Ag-Cu-Zn. Further nutrients and Cl^- are analysed. A complete list of parameters measured in 1999 is not yet given.

7.4. Quality

The detection limit (not variable) and the expected accuracy will be sent. All units are given.

File IDOD data.doc

The first two tables : columns ‘readings’ refer to the start and end readings of the pump, the difference is the volume air sampled. The time stamp (between start and end time) refers to the time when the meteorological information is recorded.

Results of EPMA : the presence of different types of particles is indicated with a cross.

Tables of samples for nutrients : two different lines for start and end with in 'Bulk – A' and 'Bulk – B' the readings of the pump and the next line, the volume of air sampled.

The results for nutrients in the column 'Water' are intermediate (dissolved) and do not refer to the water column. F-? : it is not clear whether it indicates fluor or organic substances, the peaks are very close. The samples A4 and A5 of campaign 98/21 were contaminated by sea spray and thus not analysed. The loads of samples I (impactor samples) of campaign 98/9 were to low, so no results are available.

7.5. Question/comments

The weight of the particles is not determined. When the filters would be dried prior to the measurement, the volatile aerosols would evaporate.

Date of analysis/samples :

The date of analysis is not recorded but this will be started. Normally the samples are analysed within two weeks till a month after sample collection. After analysis, no samples remain.

Analysis tools :

Important information is provided by comparing the analysis results with air mass backtrajectories (KMI) (information on the origin of the airmass for 24 hours).

7.6. Actions

With regards to the file format: for IDOD it would be easier to have all information about one sample on one line (e.g. start time, end time, air volume sampled, odas time stamp, etc) to minimise errors during input. Intermediate results like the readings of the pump and the data on the water column are not necessary.

UIA

- To send us the complete description of methodology of the different parameters, together with detection limit and expected accuracy.
- To record date of analysis from now on.
- To send the data for 1998 in a format like described above.

IDOD

- To check the possibility of incorporating air mass backtrajectories.

8. Meeting VUB-ECOL, Laboratoroium voorEcologie and Systematiek

Held on September 2th, 1999
Present: Elvire Antajan, Nanette Daro, Stéphane Gasparini, Allan Meyer, Siegrid
Jans

8.1. Data set

Samples of mesozooplankton are collected during spring campaigns (2-3 weeks). Samples for abundance are collected nearly each week from February to mid-June. Other campaigns (1-2) are made during the year. Water samples (for pigment analysis) are collected tentatively weekly at station 330.

8.2. Meta-information

8.2.1. General

All the stations are related to the MUMM's reference position. As the boat has to move during the sampling (to create a water current), the real sample position is around the station reference.

Sampling time and date are available.

8.2.2. Methods

The description of the sampling processing and analysis will be given by Elvire Antajan (for mesozooplankton) and by Allan Meyer (for pigments).

8.2.3. Quality control

No participation to international neither comparative exercises.

8.2.4. Meteorology

Meteorology information is not available from VUB-ECOL; the IDOD team will recover them in the ODAS database.

8.3. Parameters

Data set from VUB will be sent for 1998. There was no team on board in 1997, so there is no data to expect for this year.

Some modifications need to be made in the “List of expected parameters”:

In water: VUB does not measure suspended matter and carbon anymore. These analyses will be made by ULB from 1999. Pigments need to be had to the list (Chl-a, Chl-c, fuccoxanthine, and phaeopigment + other pigments to add when the complete data set is given).

8.4. *Quality*

Detection limit

A detection limit will be given for pigment analysis, but not for mesozooplankton.

Quantification limit

Only used for pigment analysis.

Expected accuracy

An accuracy of 30% is expected for the mesozooplankton analyses (abundance).

8.5. *Questions/Comments*

ULB, VUB and IDOD team think it should be interesting to hold a meeting to discuss some particular data. Indeed, some data are the result of an analysis specific to the laboratory (the technique has been developed by the lab). It is therefore difficult and dangerous to interpret these data without knowing the method used. Moreover, it is not rough data but analysed data (with a subjective analysis). This fact can introduce a problem of access right and property. The same type of derived date is present in the UMH's data set.

VUB has proposed to make a reference to these types of data (just to say that they exist but without put them in the database, as they are not listed in the "list of parameters expected") and to allow a link to the laboratory for further information.

Samples

There is no replicate analysis for mesozooplankton.

The data set for pigments present replicate values. The idea is to send the complete data set and not only the mean values.

All mesozooplankton samples are stored for ever. Water samples (for pigments) are destroyed during analysis.

8.6. *Actions*

VUB

To send the complete 98 data set (no cruise in 1997) to Véronique Rousseau

IDOD

To recover water depth and meteorological data from ODAS

To modify the “list of parameter expected”

To organise a meeting between the three teams to discuss the problem of derived data

9. Meeting VUB-ETOX (Laboratorium voor ecotoxicologie en polaire ecologie)

Held on 14th of October 1999
Present : Ludo Holsbeek, Mia Devolder & Karien De Cauwer

9.1. Data set

Analyses are done on all mammals and on a fixed number of guillemots per year gathered by IN and Ulg-Coignoul. When not enough guillemots are collected, other species are analysed. Only ‘fresh’ animals are analyzed.

9.2. Meta-information

9.2.1. General

See report of the visit to Ulg-Laboratoire d’Océanologie and Ulg-Service d’Anatomie pathologique.

9.2.2. Methods

A description of the methodology with the appropriate QUASIMEME codes will be sent. The sampling method is described by Ulg-Laboratoire d’Océanologie and Ulg-Service d’Anatomie pathologique.

9.2.3. Quality control

The laboratory has participated in QUASIMEME exercises, analyses certified samples and internal standards. The results and details of these quality control procedures will be sent.

9.2.4. Meteorology

Not available.

9.3. Parameters

Data exist for total Hg, monomethyl Hg (MMHg), PCBs (7 or 10 congeners), DDTs and hydrocarbons. The dry weight (dw)/fresh weight (fw) ratios and the neutral lipid weight (lw)/dw ratios are available.

The neutral or non-polar lw is determined by extracting with hexane (Remark : Ulg Bouquegneau determines total lipids). High lipophylic substances like PCBs accumulate mainly in neutral lipids. They are determined on lw basis. Total Hg (only slightly lipophylic) is determined on fw basis and MMHg on dw basis. Inorganic Hg can be retrieved by subtracting MMHg from total Hg, taking in mind that the measurement error will be doubled.

9.4. *Quality*

Detection limit

An absolute detection limit exists for the apparatus (dose that can be detected). The detection limit for a sample is variable and can be derived by calculating an average value for a certain sample handling procedure (for PCB's the result of a blank analysis is zero). The limit is currently seen as the limit above which the concentrations are quantifiable, so in fact the quantification limit. Two limits could however be defined : detection limit, above which one can assume the contaminant is present, and quantification limit, above which the values for contaminant concentrations can be accepted.

Expected accuracy

An accuracy of 10-20% is expected for PCB analyses. The reproducibility of the apparatus as well as the reproducibility of the method (including extraction, injection, ...) will be given.

9.5. *Questions/Comments*

The date of analysis is currently noted in the processing files/books. In future it will be supplied with the final data set. Especially for the measurement of dry weight/fresh weight, the storage time is important (more or less 10% loss of weight after one year in the freezer). The samples are analysed within a period of two weeks till 3 months after delivery. The extracted samples are stored till all analyses are finished. Lyophilized samples are stored in dark forever. Some fresh samples are stored 2 till 3 years in the freezer.

With respect to statistical analysis, non-parametrical test should be foreseen. As PCB's accumulate in animals, concentrations should be standardised to the age of the individuals.The statistical analysis of data of PCBs in mammals is not adequate if no information exists on the age of the individuals. The first years, the age can be guessed using a growth curve (size, weight). For older animals, the teeth of the mammals should be analysed to provide information on the age. However, no means are foreseen in this project. For seabirds, especially migrating birds, it is also important to consider the season. As the diet changes, higher concentrations are found after residing in the southern North Sea.

9.6. *Actions*

VUB-ETOX

To send complete data file with results for PCBs (individual congeners), hy-

drocarbons, total and MM- Hg, dw/fw and lw/dw.
To send information on methodology, precision and quality control.

IDOD

To send the “list of parameter” measured in the frame of the programme ‘Sustainable Management of the North Sea’.

To ask data on the weight of the organs analysed at Virginie Debacker.

10. Meeting MUMM, dataset Kevin Ruddick

Held on 19th of October, 1999
Present : Kevin Ruddick, Mia Devolder, Siegrid Jans & Karien De Cauwer

10.1. Dataset

The data received by Kevin is not only gathered in the frame of AMORE, but also for the project MULTICOLOR (1997-2000) (including Dutch and Russian data for MUMM monitoring stations).

Data was collected during four cruises in 1997 and 1998. Water samples and CTD profiles were taken. Vertical profiles of natural fluorescence were executed during cruise 97/16. Spectral reflectance data were obtained during cruise 98/8. PNF measurements and reflectance data will also be collected in the future. The PNF profiles will only be used for derivation of the vertical attenuation coefficient, not for the estimation of chlorophyll concentration.

The CTD profiles will be processed by Kevin in order to obtain down-casts after pump stabilisation with suitable smoothing and despiking every 0.5 m.

The variability of optical parameters and chlorophyll is studied along the transect 230-330-435.

10.2. Meta-information

10.2.1. General

Sampling date, time and position are available for the water samples. The PNF-profiles have start- and end time noted. Sampling time is missing for the RAS radiometer data. Sampling depth for floating radiometer : 0 m (5 cm above surface for downwelling irradiance, 5 cm below surface for upwelling radiance, reflectance deduced is thus at (below) the surface).

Start time is stored for CTD-profiles, start position can be retrieved via ODAS. As those profiles take little time and the clocks of the ODAS system and CTD software are not perfectly synchronised, it is not appropriate to store the end position.

10.2.2. Methods

The methodology is described in Word-files.

For the derivation of kPAR : the smoothing uses 25 points which typically represent 5 m, but can vary depending on the speed, between 3 and 8m. The near-surface scalar PAR is taken manually from the profile at 5 m depth, but the variation in that region should be less than 10%. See User’s Manual of PNF-300 Profiling Natural Fluorometer.

10.2.3. Quality control

Described in Word-files.

10.2.4. Meteorology

To be retrieved from ODAS.

10.3. Parameters

Definitions

Irradiance (landing on surface) E (in Wm^{-2})

Flux arriving at a surface, total flux integrated over an hemisphere.

Downwelling irradiance $E_d = \Phi_d/A$

Flux passing down into the lower part of the water column.

Upwelling irradiance $E_u = \Phi_u/A$

Flux being backscattered in the lower water and travelling towards the surface.

Radiant intensity $I=d\Phi/d\varpi$ (in Wsr^{-1})

The radiant flux per unit solid angle leaving a point source in a given direction. An isotropic source of radiant intensity I emits a total flux of $\Phi=4\Pi I$

Radiance $L = dI/d(A\cos\Theta)$ (in $Wsr^{-1}m^{-2}$)

Optical property appropriate to light energy leaving an extended source. The radiant flux per unit solid angle in a given direction per unit projected source area in that direction. A Lambertian source is one in which the radiance L is independent of the angle from which it is viewed (normal assumption in viewing ocean colour from space) : $L=E_u/\Pi$

Diffuse attenuation coefficient $K= - 1/ E(z). dE/dz$ (in m^{-1})

(Source : Robinson, I.S. (1985). Satellite oceanography an introduction for oceanographers and remote-sensing scientists)

Measurements

Seabird : E_d : downwelling irradiance 400-700 nm with derivation of $kPAR_d$

PNF : E_0 : scalar irradiance 400-700 nm with derivation of $kPAR_0$ (integrated over all directions). KPAR is function of suspended matter and chlorophyll (approximately 1/Secci).

L_uChl : radiance at 683 nm (only one direction). Lu(683) usually refers to a measurement of upwelling radiance centred spectrally at 683 nm using a sensor with a 10 nm wide spectral responsivity. An LuChl natural fluorescence detector has a responsivity function that is wider than 10 nm and follows closely the emmission spectrum for chlorophyll (see PNF User's manual p77-78).

RAS floating radiometer : E_d and L_u (only one direction) with assumption that radiance is independent of the angle (Reflectance = $E_u/E_d \sim L_u\Pi/E_d$). 25 spectroscans are converted in one by removing noise. The resolution of the instrument is 2.5 nm. Only a subset is received which is however complete enough for the research purposes. The data is used for the validation of bio-optical models.

CTD-profile : Salinity, temperature and optical backscatter. It would be more convenient to convert these data to standard depth intervals (e.g. 0.5m; not smaller than 0.1 m as the sensor does not react that fast). Downcasts should be stored as the water is still undisturbed.

Single data result from the profile to store with sample data : S top-bottom (measure for stratification).

ODAS parameters of interest for AMORE :

Salinity, temperature, fluorescence, time and position every 3 minutes.

Meta info : light intensity (clouds). Windspeed is not accurate when measured on board.

10.4. *Quality*

Expected accuracy for PNF measurements (User's manual):

Water temperature : 0.1°C
Depth : 1% full scale
PAR : ?
Natural fluorescence : ?

Accuracy of radiance and irradiance measurements of RAS floating radiometer : 5%.

10.5. *Questions/comments*

Report Burenkov (1998) p 8 : a_g is measured by Machteld Rijkeboer after filtration with a spectrophotometer.

IDOD agrees with the restriction asked by Machteld on the use of her data till a more specific bilateral contract on access rules exists. At first instance documents defining the rights and obligations of data providers and IDOD are being developed for the projects of the programme Sustainable Management of the North Sea. In a later stage they will be adapted for other data providers.

Also satellite data is collected :

- SeaWIFS : 6 bands, derivation of suspended matter (and perhaps chlorophyll)
- AVHRR : derivation of surface temperature and suspended matter

Derived products of interest :

- Countourplots of each monitoring parameter on which it is possible to select a place and see the corresponding CTD profile.
- Scatterplots (e.g. KPAR versus SUSP)

10.6. Actions

Kevin

To supply the IDOD-team with processed CTD-profiles.

To contact Joan with respect to the automatic extraction ODAS data.

Deadline : End of January

IDOD

To test storage and analysis tools for CTD data on the processed CTD-profiles.

Deadline : 15th of March

ANNEX 4

MsVBA code of the Search-form

```
Private Sub cmdSearch_Click()
Dim strQuery, viewquery, temp2 As String
Dim i, m, n, t, o As Integer
Dim vfields As String
Dim qfields As String
Dim params As String
Dim temp As String
Dim tblMyTable As TableDef
Dim rst As Recordset

DoCmd.SetWarnings 0
strQuery = ""
viewquery = ""
DoCmd.Close acTable, "Search_results", acSaveNo
DoCmd.Close acForm, "View definition of fields", acSaveNo
If Not IsNull(Me!strParam) Then

'If parameter (and category) is specified, build the query string

    strQuery = Me!strParam
    viewquery = "/  PARAMETER : " & Me!paramlist
    If Not IsNull(Me!strCat) Then
        strQuery = strQuery & " OR " & Me!strCat
        viewquery = viewquery & ";" & Me!catlist
    End If
    strQuery = "(" & strQuery & ")"
Else

'build query string in case only category is specified

    If Not IsNull(Me!strCat) Then
        strQuery = Me!strCat
        strQuery = "(" & strQuery & ")"
        viewquery = "/  PARAMETER : " & Me!catlist
    End If
End If

'Add selection criteria of time, place and meta-information in the query string

If Not IsNull(Me!sel_beginyear) Then
    If (strQuery = "") Then
        strQuery = "([Sample]![Monitoring_year]>= Forms!Search![sel_beginyear])"
    Else
        strQuery = strQuery & " AND ([Sample]![Monitoring_year]>= Forms!Search![sel_beginyear])"
    End If
    viewquery = viewquery & "/  BEGINYEAR : " & Me!sel_beginyear
End If
If Not IsNull(Me!sel_endyear) Then
    If (strQuery = "") Then
        strQuery = "([Sample]![Monitoring_year]<= Forms!Search![sel_endyear])"
    Else
        strQuery = strQuery & " AND ([Sample]![Monitoring_year]<= Forms!Search![sel_endyear])"
    End If
    viewquery = viewquery & "/  ENDYEAR : " & Me!sel_endyear
End If
If Not IsNull(Me!sel_campaign) Then
    If (strQuery = "") Then
        strQuery = "([Sample]![#Campaign] = Forms!Search![sel_campaign])"
    Else
        strQuery = strQuery & " AND ([Sample]![#Campaign] = Forms!Search![sel_campaign])"
    End If
    viewquery = viewquery & "/  CAMPAIGN :" & Me!sel_campaign
End If
If Not IsNull(Me!sel_station) Then
    If (strQuery = "") Then
        strQuery = "([Sample]![#Station] = Forms!Search![sel_station])"
    Else
        strQuery = strQuery & " AND ([Sample]![#Station] = Forms!Search![sel_station])"
    End If
        viewquery = viewquery & "/  STATION : " & Me!sel_station
End If
If Not IsNull(sel_project) Then
    If (strQuery = "") Then
        strQuery = "([Sample]![#Project] = Forms!Search![sel_project])"
    Else
        strQuery = strQuery & " AND ([Sample]![#Project] = Forms!Search![sel_project])"
    End If
    viewquery = viewquery & "/  PROJECT : " & Me!sel_project
End If
If Not IsNull(sel_labo) Then
    If (strQuery = "") Then
            strQuery = " ([Non-continuous_value]![#Service] = Forms!Search![sel_labo])"
    Else
        strQuery = strQuery & " AND ([Non-continuous_value]![#Service] = Forms!Search![sel_labo])"
    End If
    viewquery = viewquery & "/  LABORATORY : " & Me!sel_labo
End If
If Not IsNull(flag) Then
    If (strQuery = "") Then
         strQuery = " [Non-continuous_value]![Validity_flag] = " & Chr$(34) & Me!flag & """"
    Else
         strQuery = strQuery & " AND [Non-continuous_value]![Validity_flag] = " & Chr$(34) & Me!flag & """"
    End If
    viewquery = viewquery & "/  VALIDITY FLAG : " & Me!flag
End If

'Run query and put results in table 'Search_firstoutput'
'- if no criteria specified

If ((strQuery = "") And (Me!beginmonth = "Jan") And (Me!endmonth = "Dec") And (Me!beginday = 1) And
(Me!endday = 31)) Then
    If Not vbOK = MsgBox("No criteria specified.@Are you sure you want to continue ?" & _
        "@Click OK to see all records." & _
        "Click cancel to specify selection criteria", vbQuestion + vbOKCancel, "IDOD - seawater") Then
        DoCmd.Hourglass False
        Application.Echo True
    Exit Sub
    End If

    DoCmd.RunSQL "INSERT INTO Search_firstoutput SELECT Basic_query.* FROM Basic_query;"

'- if only period specified

ElseIf (strQuery = "") Then
    strQuery = "(CInt(Format([Sample]![Date], " & Chr$(34) & "mdd" & Chr$(34) & "))) Between
(Forms!Search!begindate) And (Forms!Search!enddate)"
    DoCmd.RunSQL "INSERT INTO Search_firstoutput SELECT Basic_query.* FROM Basic_query WHERE (" &
strQuery & ");"

'- if period and other criteria specified

Else
```

```
    If Not IsNull(Me!begindate And Me!enddate) Then
    strQuery = strQuery & " AND ((CInt(Format([Sample]![Date], " & Chr$(34) & "mdd" & Chr$(34) & "))) Between
(Forms!Search!begindate) And (Forms!Search!enddate))"
    End If
    DoCmd.RunSQL "INSERT INTO Search_firstoutput SELECT Basic_query.* FROM Basic_query WHERE (" &
strQuery & ");"
    End If
Me!strQuery = strQuery
o = CurrentDb.TableDefs("Search_firstoutput").RecordCount
If (o = 0) Then
    MsgBox ("No results found")
Else

'Exporting results into table with a column for every parameter
'-Execute crosstab query (Searchcross_value) for values and save results in table 'Search_results1'

DoCmd.RunSQL "SELECT Searchcross_value.* INTO Search_results1 FROM Searchcross_value;"

'- Build part of sql string by retrieving fieldnames (depending on parameters specified) to specify
'  the fields to be added for the qualifier and validity flag

vfields = ""
qfields = ""
params = ""
temp = ""
i = 10
n = CurrentDb.TableDefs("Search_results1").Fields.Count
temp = Left(CurrentDb.TableDefs("Search_results1").Fields(i).Name,
Len(CurrentDb.TableDefs("Search_results1").Fields(i).Name) - 6)
vfields = "Searchcross_vflag." & temp & "_Validity_flag"
qfields = "Searchcross_qflag." & temp & "_Qualifier_flag"
params = temp
i = i + 1
For i = 11 To n - 1
    temp = Left(CurrentDb.TableDefs("Search_results1").Fields(i).Name,
Len(CurrentDb.TableDefs("Search_results1").Fields(i).Name) - 6)
    params = params & """, """ & temp
    vfields = vfields & ", Searchcross_vflag." & temp & "_Validity_flag"
    qfields = qfields & ", Searchcross_qflag." & temp & "_Qualifier_flag"
Next i

'- add columns for validity and qualifier flag by make-table queries and using result of crosstabqueries for
resp. vflag and qflag (Searchcross_vflag and Searchcross_qflag)

    DoCmd.RunSQL "SELECT DISTINCTROW Search_results1.*, " & qfields & " " & _
    "INTO Search_results2 " & _
    "FROM Search_results1 LEFT JOIN Searchcross_qflag ON (Search_results1.[Start_date] =
Searchcross_qflag.[Start_date]) AND (Search_results1.Start_time = Searchcross_qflag.Start_time) AND
(Search_results1.sampling_depth = Searchcross_qflag.sampling_depth) AND (Search_results1.Sequence_number =
Searchcross_qflag.Sequence_number) AND (Search_results1.Monitoring_year = Searchcross_qflag.Monitoring_year);
"
    DoCmd.RunSQL "SELECT DISTINCTROW Search_results2.*, " & vfields & " " & _
    "INTO Search_results " & _
    "FROM Search_results2 LEFT JOIN Searchcross_vflag ON (Search_results2.[Start_date] =
Searchcross_vflag.[Start_date]) AND (Search_results2.Start_time = Searchcross_vflag.Start_time) AND
(Search_results2.sampling_depth = Searchcross_vflag.sampling_depth) AND (Search_results2.Sequence_number =
Searchcross_vflag.Sequence_number) AND (Search_results2.Monitoring_year = Searchcross_vflag.Monitoring_year);
"

    '-Change order of columns (if ordinal position same for all columns, columns will be ordered alphabetically

    i = 10
    t = 10
    m = CurrentDb.TableDefs("Search_results").Fields.Count
For t = 10 To (m - 1)
    CurrentDb.TableDefs("Search_results").Fields(i).OrdinalPosition = m
Next t

'Display information on field names in the search results : parameter, methodology and applied methods

    params = "(""" & params & """)"
DoCmd.RunSQL "SELECT Parameter.[ICES_parameter_code] & Left(Parameter.[Matrix],1) & Parameter.[#Method] AS Fieldname, " & _
"Parameter.Substrate, Parameter.Matrix, Parameter.Parameter_name, Parameter.Measurement_unit, Analysis_method.Description, " & _
    "Analysis_method.Detection_limit , Analysis_method.Detection_unit, Sample_handling.Sample_preservation, Sample_handling.[Sample_pre-treatment], " & _
    "Sample_handling.Sample_separation, Sample_handling.Procedure_description, Service.Service_code, Service.Service_name" & _
    " INTO Search_results_fields" & _
    " FROM (Service RIGHT JOIN (Sample_handling RIGHT JOIN Analysis_method ON Sample_handling.[ID-Sampling_handling] = Analysis_method.[#Sample_handling])" & _
    " ON Service.[ID-Service] = Analysis_method.[#Service]) RIGHT JOIN Parameter ON Analysis_method.[ID-Analysis_method] = Parameter.[#Method]" & _
    " WHERE (Parameter.[ICES_parameter_code] & Left(Parameter.[Matrix], 1) & Parameter.[#Method]) IN " & params & ";"

'Delete all records in table search_firstoutput

If Not (CurrentDb.OpenRecordset("Search_firstoutput").RecordCount = 0) Then
    DoCmd.OpenTable "Search_firstoutput", acViewNormal, acEdit
    DoCmd.SelectObject acTable, "Search_firstoutput"
    RunCommand acCmdSelectAllRecords
    RunCommand acCmdDeleteRecord
End If
DoCmd.Close acTable, "Search_firstoutput"
DoCmd.Minimize
End If
Application.Echo True
DoCmd.Hourglass False
DoCmd.SetWarnings -1
If (o = 0) Then
    DoCmd.Close acForm, "Viewresults"
    DoCmd.SelectObject acForm, "Search"
    DoCmd.Restore

Else
    DoCmd.OpenForm "Viewresults"
    Forms!Viewresults!query = viewquery
End If
End Sub
```

Integrated and Dynamical Oceanographic Data Management

IDOD

Contract MN/DD/61

Scientific report

Redaction : Fabrice MULLER

December 1999

SURFACES

Service Universitaire de Recherches Fondamentales et Appliquées
en Cartographie et en Études Spatiales

TABLE DES MATIÈRES

1. Introduction

During the 1999-year, the different aspects of data visualisation were investigated. Almost all data included in the IDOD database are point data requiring interpolation to be visualised as continuous data surfaces. For this reason, the different existing interpolations were investigated during this year. The main part of this report is devoted to a presentation of these interpolation or gridding techniques.

2. Spatial data analysis

2.1. Introduction

Spatial analysis *: quantitative study of phenomena that are located in space.*

Spatial data analysis or geostatistic involves the accurate description of data relating to a process operating in space, the exploration of patterns and relationships in such data, and the search for explanations of such patterns and relationships.

It is important to make a distinction between discrete and continuous view of spatial phenomena. A discrete view consists of a space filled with objects, while a continuous view consists of a space covered with essentially continuous surfaces. The first one is generally used to visualise all the spatial phenomena that are usually conceptualised as points, lines or areas. For example, such objects can be : buoys, coast lines, *etc*. On the other hand, for phenomena in the natural environment such as salinity, temperature, relief, and so on, the variables can be measured anywhere on the earth's surface. But in practice, data are discrete because the measures are sampled at some specific points. Thus, the relief can be sampled at a collection of sample points and represented as a series of contour lines, while temperature is sampled at a set of sites and represented also as a collection of lines or isotherms. But such variables being sampled at a set of discrete locations can also be represented as a continuously varying field. In all cases, an attempt is made to represent an underlying continuity from discrete sampling. To summarise, it is conceptually useful to keep in mind the distinction between an entity-oriented as opposed to a field-oriented view of spatial phenomena.

Data can be handled and displayed using a GIS (Geographical Information System). A GIS is a computer-based set of tools for capturing or collecting, editing, storing, integrating, analysing and displaying spatially referenced data. The data are stored in a database and the GIS is linked with a DBMS (DataBase Management System). The querying of the database is an important function of the GIS. The main GIS software available on the market are : *ARC/INFO, MGE, IDRISI, GRASS, SPANS, MapInfo, Geomedia, TransCAD, etc*.

Generally, a GIS is vector or raster oriented but not both. For example, *ARC/INFO* is essentially vector-based with good functionality for handling raster data, while *IDRISI* is fundamentally raster-based but allows vector overlay.

In data spatial analysis, there are three distinct fields of application : methods that are essentially dedicated to the *visualisation* of spatial data, those which are *exploratory*, concerned with summarising and investigating map patterns and relationships, and those which rely on the specification of a *statistical model* and the estimation of parameters. In the practice, the distinction between methods for visualising, exploring and modelling data is not always very clear-cut.

2.1.1. Visualising spatial data

It is essential in any data analysis to visualise or to see data being analysed. So, the visualisation is equivalent to mapping. Many GIS offer an environment with functionality to create such maps and explore spatial patterns and relationship interactively, easily and quickly.

2.1.2. Exploring spatial data

Exploratory methods in data analysis involve seeking good descriptions of data for helping the analyst to develop hypotheses and appropriate models for such data. The employed techniques require *a priori* assumptions about the studied data. These techniques are designed to highlight particular features such as unusual values, and so on. The results of the exploratory stage are presented in the form of maps or may involve other kinds of plots.

2.1.3. Modelling spatial data

In some cases, the answer to the asked questions can be found by a judicious choice of exploratory methods combined with appropriate visualisation methods. In other cases, it will be necessary to test some hypotheses and to consider explicit statistical models.

A statistical model for a stochastic phenomenon consists of specifying a probability distribution for the random variable(s) that represent the phenomenon. Let be a random variable Y taking values in $\Re$. A spatial process $\{Y(s), s \in \Re\}$ is *stationary* or *homogenous*, if its statistical properties are independent of absolute location in $\Re$. In particular, this would imply that the mean $E(Y(s))$ and variance $VAR(Y(s))$ are constant in $\Re$ and therefore do not depend of the absolute location s. Stationarity also implies that the $COV(Y(s_i),Y(s_j))$, between values at two sites, depends only on the relative locations of these sites, the distance and direction between them, and not on their absolute location in $\Re$. Moreover, the spatial process is *isotropic* if it is stationary and the covariance depends only on the distance between the two locations and not on the direction in which they are separated. Finally, if the mean, or variance, or covariance changes over $\Re$ then the process is *non-stationary* or *heterogeneous*.

2.2. Exploring spatially continuous data

2.2.1. Data distribution

The data distribution can be described by the data coordinates or by its density function. A well-sprayed data distribution over the studied area is fundamental to obtain good results with spatial analysis. The figure 1 shows the locations of the Belgian sampling stations in the North Sea and Scheldt estuary.

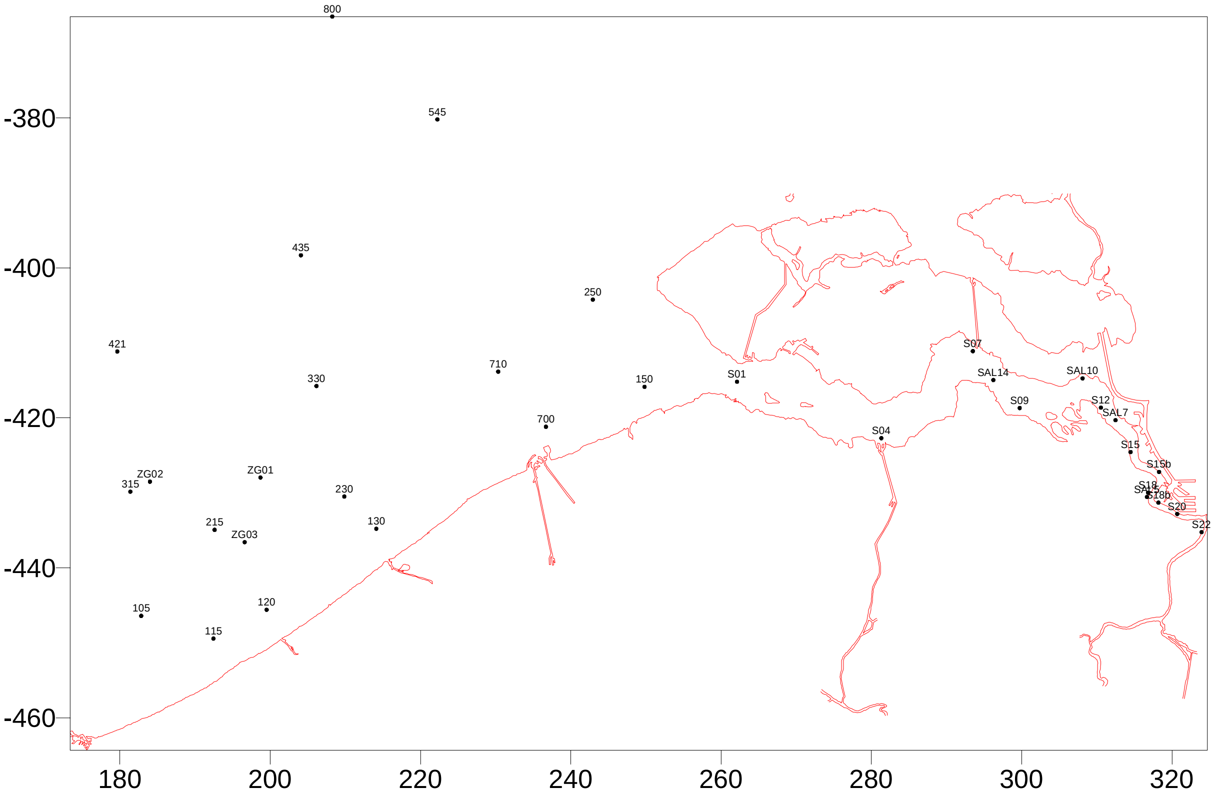


Figure 1. Sampling stations of the North Sea and Scheldt estuary.

2.2.2. Distribution described by the coordinates of the samples

The coordinates of the samples describe the location of the data points as shown in figure 1. At every station a parameter value is measured and registered in the database.

A very simple way to estimate a value $\mu(s)$ at point s is by the average of the values at neighbouring sampled data points. For example, $\mu(s)$ can be estimated as the non-weighted average of the sample values at the nearest sampling points inside a circle of radius r. The obvious problem with this approach is that it does not allow for spatial variations in the distribution of sample sites. There is no discrimination between a site that is at a large distance from its neighbours and one that is very close to them. Using a weighted average of neighbouring points as in the following formula can solve this problem:

$$\mu(s) = \sum_{i=1}^{N} w_i(s) z_i$$

where $\sum_{i=1}^{N} w_i(s) = 1$ with $w_i(s) \propto h_i^{-\alpha}$ or $w_i(s) \propto e^{-\alpha h_i}$,

and where h_i is the distance from s to s_i and α is a parameter with a value chosen to provide a suitable degree of smoothing. Usually $w_i(s)$ is chosen equal to zero beyond some appropriate maximum distance to examine only a few neighbours.

This technique is the easiest to implement but there are more sophisticated ways to perform the weighting.

The data distribution can also be formulated by using a position parameter such as the gravity centre of the data points. Here also, a weighting coefficient can be used.

$$x_G = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} \quad \text{and} \quad y_G = \frac{\sum_{i=1}^{N} w_i y_i}{\sum_{i=1}^{N} w_i}$$

Another way to define the data distribution is by use of an ellipse of dispersion. The variance and covariance along the coordinates axes define the dispersion.

$$s_x^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2 \qquad s_y^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{y})^2 \qquad C = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$$

The dispersion can also be given along an axis of $\alpha°$ to the abscises axe :

$$s_\alpha^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i \cos\alpha - x_i \sin\alpha)^2 = s_x^2 \cos^2\alpha + s_y^2 \sin^2\alpha - 2C \sin\alpha \cos\alpha$$

From all the diameters passing through the gravity centre, there are two that are orthogonal and for which the dispersion is respectively minimum and maximum. Their orientation can be found by zeroing the first derivative of s_α^2:

$$(s_y^2 - s_x^2) \sin 2\alpha - 2C \cos 2\alpha = 0$$

$$\tan 2\alpha = \frac{2C}{s_y^2 - s_x^2}$$

We can so obtain two values for α that are 90° distinct. The length of the diameters corresponding to the half-axes of the dispersion ellipse is given by s_α^2.

2.2.3. Distribution described by the density

This method requires a preliminary study of the data set to determine a reference area or surface. Generally, this reference surface corresponds to the convex hull of the data points. The density is defined as the ratio of the number of points by the surface of the reference area.

$$d = \frac{N}{S}$$

where *N* is the total number of data points and *S* is the surface of the reference area.

This reference surface *S* can be the rectangle including all the data points and is given by:

$$S = (x_{\max} - x_{\min})(y_{\max} - y_{\min})$$

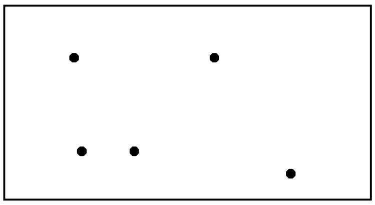
or it can be the hull of the data points. This hull is generally convex and its surface *S* is given by :

$$S = \frac{1}{2}\left|\sum_{i=1}^{N+1}(x_i y_{i+1} - x_{i+1} y_i)\right|$$

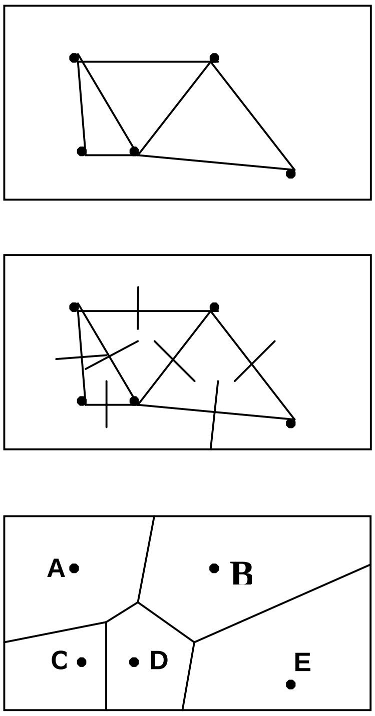
where N is the number of points of the hull and $x_{N+1} = x_1$ and $y_{N+1} = y_1$.

2.2.4. Methods based on tessellation

In the previous section, a technique was described to give spatial moving average *μ(s)*. On another way, there are a number of estimation techniques for *μ(s)*, which have been developed on the basis of tessellation or tiling of the observed sample location s_i. The most commonly used employs the *Delaunay* triangulation, also known as *triangulated irregular network* (TIN). This method consists in partitioning the planar spatial region into triangle tiles. The TIN results in connecting all points on the surface with straight lines with respect to some criteria. From this TIN, we can construct the Thiessen polygons by drawing the perpendicular bisectors of these connections. A territory or part of the planar region is assigned to each location s_i so that every point of this territory is closer to s_i than to any other of the locations. The attractive feature of Thiessen polygons is that all points on the geographic surface contained within a polygon's boundaries are closest to the given polygon's spatial data point than to any other spatial data point on the surface.



(a) Distribution of points.



(b) Delaunay triangles.

(c) Construction of perpendicular bisectors.

(d) Thiessen polygons.

Figure 2. Thiessen polygons partitioning of a points surface.

	A	B	C	D	E
A	0	1	1	1	0
B	1	0	0	1	1
C	1	0	0	1	0
D	1	1	1	0	1
E	0	1	0	1	0

Table 1. Connectivity matrix for the polygon partitioning of the surface in figure 2.

2.2.5. Spatial autocorrelation

The correlation is the measurement of the relationship prevailing between a pair of variables. The conventional statistics coefficients used to measure this relationship provide information about its nature and degree. The sign of the correlation coefficient indicates the nature of the relationship : a *plus* identifies a positive or direct relation, while a *minus* identifies a negative or inverse relation. The actual numerical value noted ρ identifies the strength of the relation. $|\rho| \to 1$ corresponds to an increasingly strong relation, whereas $|\rho| \to 0$ corresponds to an increasingly weak relation.

The spatial autocorrelation may be defined as the relationship among values of some variable that is attributable to the manner in which the corresponding area units are ordered on a planar surface.

2.2.6. Kernel estimation

If we denote by *λ(s)* the intensity of the process at location *s*, which is the mean number of events per unit area at point *s*, then the first order properties are defined by the mathematical limit :

$$\lambda(s) = \lim_{ds \to 0} \left[\frac{E(Y(ds))}{ds} \right]$$

where *ds* is a small region around the point *s*. For a stationary process, *λ(s)* is a constant over $\Re$. The second order properties or spatial dependence of a spatial point process involves the relationship between numbers of events in pairs of areas in $\Re$. Formally, this can be described by the second order intensity $\gamma(s_i,s_j)$ of the process :

$$\gamma(s_i, s_j) = \lim_{ds_i, ds_j \to 0} \left[\frac{E(Y(ds_i)Y(ds_j))}{ds_i ds_j} \right]$$

For a stationary process : $\gamma(s_i, s_j) = \gamma(s_i - s_j) = \gamma(\vec{h})$. So, the second order intensity depends on only the vector $\vec{h}$ which is the vector difference (direction and distance) between s_i and s_j and not their absolute locations. The process is isotropic if the dependence is purely a function of the length *h* of this vector $\vec{h}$ and not its orientation. In this last case, $\gamma(s_i,s_j) = \gamma(h)$.

The estimation of intensity or events per unit area over the complete area $\Re$, can be provided by kernel estimation.

$$\hat{\lambda}_\tau(s) = \sum_{i=1}^{N} \frac{1}{\tau^2} \mathrm{k}\left(\frac{(s - s_i)}{\tau} \right)$$

where k() is a bivariate probability density function (kernel) which is symmetric about the origin, and $\tau > 0$ (bandwidth) determines the degree of smoothing. The bandwidth defines a circle centred on *s*.

In the present case, we are not interested by the intensity *λ(s)* but by the mean value *μ(s)* of an attribute whose values y_i have been sampled at locations s_i. Then, it is intuitive to introduce the attribute values into the previous kernel estimate as :

$$\sum_{i=1}^{N} \frac{1}{\tau^2} \mathrm{k}\left(\frac{(s - s_i)}{\tau} \right) y_i$$

If the original estimate represents the number of observations per unit area, then this extension is the expression of the total amount of the attribute per unit area. Therefore, to obtain an estimate of the mean value of the attribute, the previous expression needs to be divided by the number of observations per unit area. An appropriate kernel estimate for *μ(s)* would be :

$$\hat{\mu}_\tau(s) = \frac{\sum_{i=1}^{N} \mathrm{k}\left(\frac{(s-s_i)}{\tau}\right) y_i}{\sum_{i=1}^{N} \mathrm{k}\left(\frac{(s-s_i)}{\tau}\right)}$$

It is important to specify that at points *s* where the denominator is 0, the numerator must also be 0 and by convention $\hat{\mu}_\tau(s)$ is set to 0 at these points.

Note that the kernel estimate $\hat{\mu}_\tau(s)$ is just a more sophisticated version of the weighted moving average schemed presented in section 2.2.

$$\mu(s) = \sum_{i=1}^{N} w_i(s) y_i \qquad \text{where} \qquad w_i(s) = \frac{\mathrm{k}\left(\frac{(s-s_i)}{\tau}\right)}{\sum_{j=1}^{N} \mathrm{k}\left(\frac{(s-s_j)}{\tau}\right)}$$

Sample observations at a given distance from *s* obtain more weight in regions of $\Re$ where sample points are sparse than where they are dense.

2.2.7. Variogram and covariogram

The covariance between two variables is the measurement of how these two variables vary together. It is estimated as the sum of the cross-products of deviations of observations from the respective means of the two variables. In spatially continuous phenomena, we know that many variables show spatial persistence and we can anticipate observing positive covariance (or correlation) at short distances and lower covariance (or correlation) at greater distances.

For a spatial stochastic process {*Y(s),* s ∈ $\Re$}, the covariance of two particular points s_i and s_j is defined as :

$$C(s_i, s_j) = E((Y(s_i) - \mu(s_i))(Y(s_j) - \mu(s_j)))$$

where *E(Y(s))* is noted *μ(s),* and the corresponding correlation is defined as :

$$\rho(s_i, s_j) = \frac{C(s_i, s_j)}{\sigma(s_i)\sigma(s_j)}$$

where *VAR(Y(s))* is noted $\sigma^2(s)$.

Note that $C(s,s) = \sigma^2(s)$.

If the process is stationary (mean and variance are independent of the location), $\mu(s)=\mu$ and $\sigma^2(s)= \sigma^2$, and we have

$$C(s_i,s_j)=C(s_i-s_j)=C(h)$$

C(h) is the *covariance function* or the *covariogram* of the process, and $\rho(h)$ is the correlogram. Note that $C(0)=\sigma^2$.

It is common to make a weaker hypothesis than that of stationarity : the *intrinsic stationarity* which is defined through a constant mean and a constant variance in the differences between values at locations separated by a given distance and direction. That is :

$$E(Y(s+h)-Y(s)) = 0$$
$$VAR(Y(s+h)-Y(s)) = 2\gamma(h)$$

The function $\gamma(\mathbf{h})$ is the *variogram* or *semi-variogram*.

For a stationary process the covariogram, correlogram and variogram are directly related by :

$$\rho(h)=\frac{C(h)}{\sigma^2}$$
$$\gamma(h)=\sigma^2-C(h)$$

In this section, the covariogram or variogram are used as exploratory devices to examine spatial dependence in the observed data. Later, they will be used in the modelling of such data.

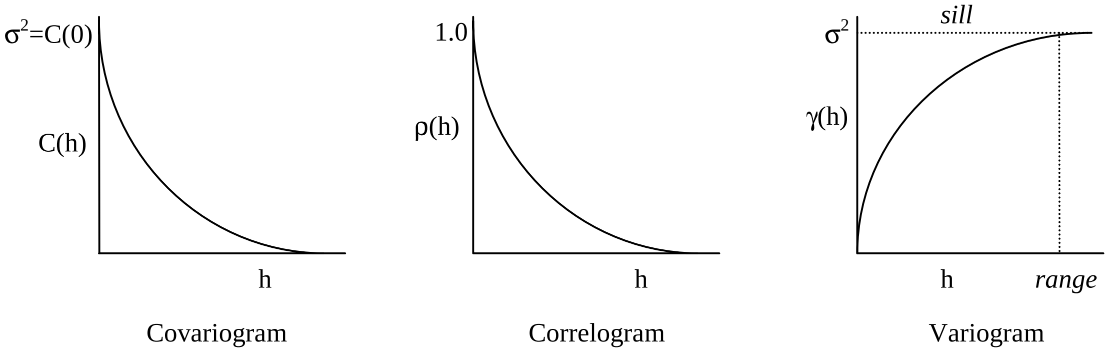


Figure 3. Covariogram, correlogram and variogram.

With the hypothesis of stationarity, the natural sample estimator of the variogram is :

$$2\gamma(h) = \frac{1}{N(h)} \sum_{s_i - s_j = h} (y_i - y_j)^2$$

where the summation is over all pairs of data points with a vector separation of *h* and *N(h)* is the number of these pairs.

Under the assumption of isotropy, the variogram is estimated over all directions for a given distance separation *h* :

$$2\gamma(h) = \frac{1}{N(h)} \sum_{|s_i - s_j| = h} (y_i - y_j)^2$$

The equivalent estimator for the sample covariogram is :

$$\hat{C}(h) = \frac{1}{N(h)} \sum_{|s_i - s_j| = h} (y_i - \bar{y})(y_j - \bar{y})$$

Generally, *N(h)* increases as *h*, the length of the vector $\vec{h}$, and the reliability of the sample estimator decreases as *h* decreases. Unfortunately, it is mainly the local behaviour that is of practical interest. The solution is to try to smooth the variability caused by small numbers of data points at short distances by using a modified estimator as :

$$\tilde{\gamma}(h) = \frac{\hat{\gamma}(h)}{\bar{\gamma}^2(h)}$$

where $\bar{\gamma}^2(h)$ is the mean of all the data values used to calculate $\hat{\gamma}(h)$ at different values of *h*. This modified estimator is called the *relative variogram*.

Theoretically *γ(0) = 0,* but due to sampling errors and small scale variability, sample values with small separations are often quite similar. This causes a discontinuity at the origin of the sample variogram and it is called the *nugget effect*. A variogram consisting of pure nugget effect, that is horizontal except at the origin, corresponds to a process with no spatial dependence.

For a stationary process, the variogram should rise to an upper bound called the *sill* and corresponding to σ^2. The distance at which this occurs is referred to as the *range*. Failure to exhibit an upper bound indicates some degree of non-stationarity in the process.

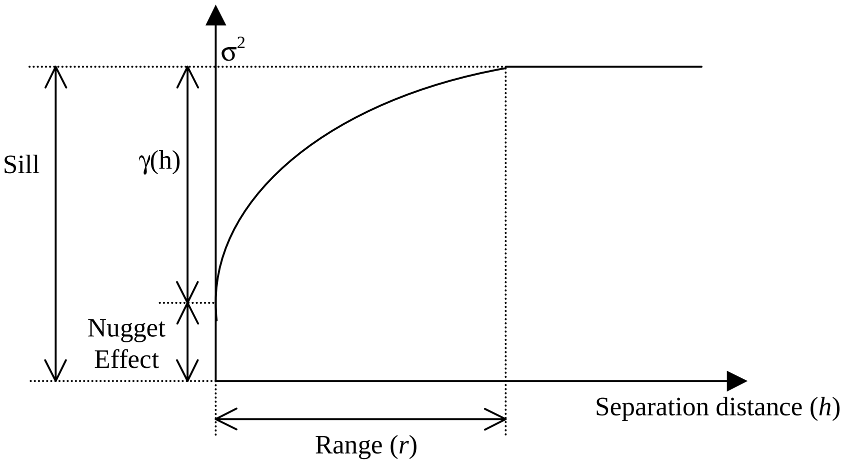


Figure 4. A typical variogram model.

2.3. Modelling spatially continuous data

After exploring spatially continuous data, it is time to consider the construction of specific models to explain the observed sample values. As in the previous section, there are models that involve first order variations in the mean value of the sample data and others that take account of second order effects.

2.3.1. Trend surface analysis

Previously, the variations in $\mu(s)$ were explored without any explicit model. One simple approach to modelling global or large scale variations in the mean value of a spatially continuous process is *trend surface analysis*. It involves the fitting of polynomial functions to the spatial coordinates (s_{i1},s_{i2}), of the sample site s_i, to the observed data values y_i at these sites by ordinary least squares regression. The variations of the data values are explained only as a function of the location : $y_i = \mathrm{f}(s_{i1}, s_{i2})$.

With the assumption that there are only first order effects involved in the process *Y(s)* and there is no spatial dependence, the regression model may be written as :

$$Y(s) = x^T(s)\beta + \xi(s)$$

where $x^T(s)\beta$ is the trend or the mean value of the random variable *Y(s)* and $\xi(s)$ is a zero mean variable that represents the fluctuations from this trend. The vector *x(s),* with dimensions (p x 1), consists of p functions of the points s. For a linear trend surface, the vector *x(s)* is simply $(1,s_1,s_2)^T$, and for a quadratic trend surface it is $(1, s_1,s_2,s_1^2,s_2^2,s_1s_2)^T$. The dimensions of the vector β are (p x 1).

With the assumption that there are no second order effects present in the process Y(s), the model may be fitted by ordinary least squares, deriving estimates $\hat{\beta}$ and their associated standard errors by :

$$\hat{\beta} = (X^T X)^{-1} X^T y$$
$$\text{VAR}\,(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

where $y = (y_1, \ldots, y_N)^T$ is the vector of observed values at the point s_i, and X is an (N x p) matrix with row vectors $x^T(s_i)$ for i = 1, …, N.

For a linear regression model and for a quadratic regression model, we have respectively :

$$X = \begin{pmatrix} 1 & s_{11} & s_{12} \\ 1 & s_{21} & s_{22} \\ \vdots & \vdots & \vdots \\ 1 & s_{N1} & s_{N2} \end{pmatrix} \quad \text{and} \quad X = \begin{pmatrix} 1 & s_{11} & s_{12} & s_{11}^2 & s_{12}^2 & s_{11}s_{12} \\ 1 & s_{21} & s_{22} & s_{21}^2 & s_{22}^2 & s_{21}s_{22} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & s_{N1} & s_{N2} & s_{N1}^2 & s_{N2}^2 & s_{N1}s_{N2} \end{pmatrix}$$

Normally, the variance σ^2 of $\xi(s)$ in the above expression for $VAR(\hat{\beta})$ would be unknown and would be estimated from the *residuals* of the model. These residuals are the differences between the observed values y_i at each sample site and those predicted by the model $\hat{y}_i = x^T(s_i)\hat{\beta}$. The appropriate estimate of σ^2 is given by :

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N - p}$$

The residuals can also help to assess the fit of the model. In practice, a measure of the goodness fitting is provided by the *coefficient of determination* calculated as :

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y}_i)^2}$$

To incorporate an adjustment taking into account the number of explanatory variables in the model, it is often preferred to use :

$$\overline{R}^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p}$$

It is important to note that the trend surface analysis is more an approach to interpolation than a formal statistical model. In most cases the model would be useless for local prediction because the simple polynomial functions typically used produce global smoothing that loses most of the local variability. When higher order terms are included in the model to retain some of these details, the regression coefficients are highly correlated and the fitted surface is very vulnerable to observational errors in y_i. Moreover, local effects in one part of the region will influence the fit of a regression surface everywhere. The results are also disrupted by a clustered distribution of sample sites and by the measurement units of the location coordinates. For this reason, it is recommended to scale the location coordinates to a unit square or to low values; otherwise, the raising of numerically large coordinates to higher powers will cause problems during the subsequent matrix calculations.

It was made the unrealistic assumption that there were only first order effects involved in the process *Y(s)* with no residual spatial dependence. Now, what happens if we relax this assumption of independence of residuals by use of *generalised least squares*. The previous model becomes :

$$Y(s) = x^T(s)\beta + U(s)$$

where *U(s)* are zero mean errors as before. However, they are not necessarily independent at different points *s* and they have some covariance function *C()*. The estimates for β and the corresponding standard errors are now :

$$\hat{\beta} = (X^T C^{-1} X)^{-1} X^T C^{-1} y$$

$$\mathrm{VAR}(\hat{\beta}) = (X^T C^{-1} X)^{-1}$$

where *C* is a (N x N) matrix of covariances $C(s_i,s_j)$ between $U(s_i)$ and $U(s_j)$ for each possible pair (*i,j*) of the N sample sites. The diagonal elements of this matrix are $VAR(U(s_i))$.

The covariance matrix *C* is required but, unfortunately, it is not possible to calculate in practice because we have only one observation at each pair of sample site. Thus, the solution is to develop models of covariance that can be estimated from the data and that allow to find the elements of *C* indirectly.

2.3.2. Models for variograms

As said in the previous section, it is rarely in practice possible to calculate the elements of the covariance matrix *C*. Fortunately, an indirect estimate of this matrix can be derived from known models and then be used in generalised least squares in order to obtain a model for the spatially continuous data incorporating both first order and second order effects.

The present section will only concern the models of the covariance structure for stationary processes. If the process is not stationary, it requires first to be smoothed by regression methods to remove the trend in the process.

Generally, the sample variogram provides more robust estimate of spatial dependence than the sample covariogram in the presence of any remaining minor departures from stationarity and the weaker assumption of intrinsic stationarity. Since stationarity is assumed, the covariogram model can be derived directly from the variogram model by the theoretical relationship : $\gamma(h) = \sigma^2 - C(h)$, where $\sigma^2 = C(0) = \gamma(\infty)$.

Moreover, only certain mathematical functions are permissible models for covariance functions and so only certain functions are valid for variograms. Necessary and sufficient conditions for the covariance function of a general spatial process are those of :

Symmetry : $C(s_i, s_j) = C(s_j, s_i)$

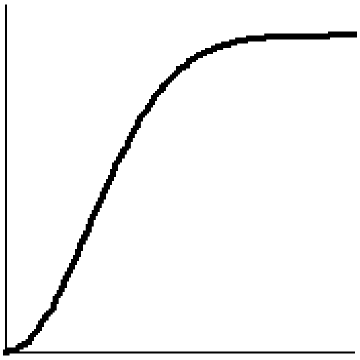
and

Non-negative definition : $\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j C(s_i, s_j) \geq 0$

for all $N, \alpha_1, \ldots, \alpha_N$ and $s_1, \ldots, s_N$.

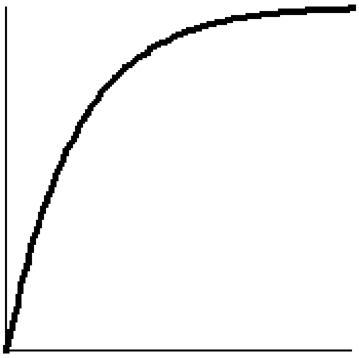
The covariance functions for stationary or isotropic processes are a restricted subset of the above functions. In practice, the appropriate functions are difficult to check and one way to guarantee that the relevant conditions are satisfied is to build models from only a few families of functions that are known to be valid.

The following valid variogram models are most commonly used for stationary processes :



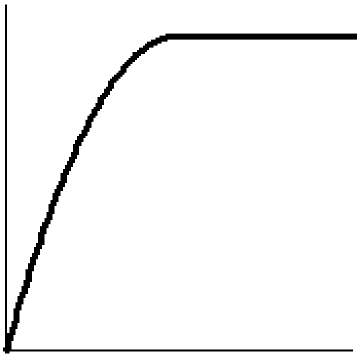
Gaussian Model

$$\gamma(h) = C\left[1 - e^{-h^2}\right]$$



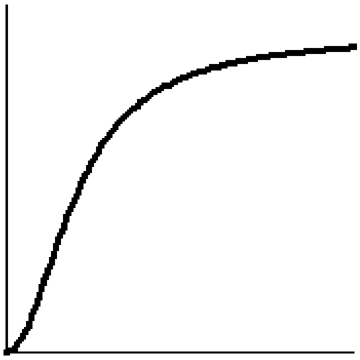
Exponential Model

$$\gamma(h) = C\left[1 - e^{-h}\right]$$



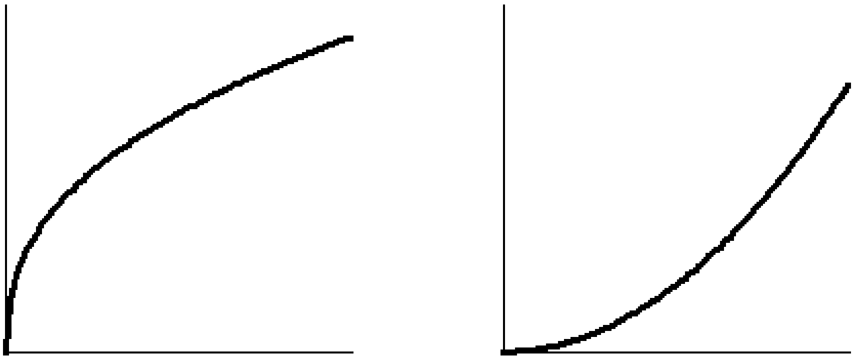
Quadratic Model

$$\gamma(h) = \begin{Bmatrix} C\left[2h - h^2\right] & h < 1 \\ C & h \geq 1 \end{Bmatrix}$$



Rational Quadratic Model

$$\gamma(h) = C\left[\frac{h^2}{1 + h^2}\right]$$

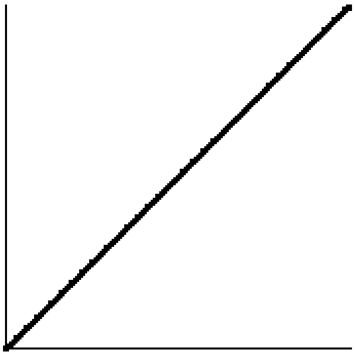


$0 < h < 1$ **Power Models** $1 < h < 2$

$$\gamma(h) = C\left|h^n\right|$$

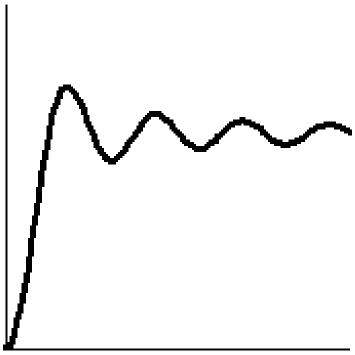
where 0 < n < 2

when n=1 it is a Linear Model



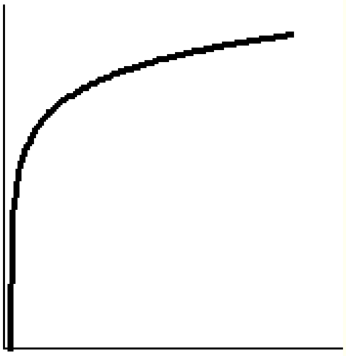
Linear Model

$$\gamma(h) = C(h)$$



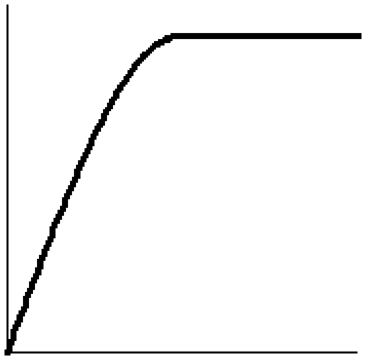
Wave (Hole Effect) Model

$$\gamma(h) = C\left[1 - \frac{\sin h}{h}\right]$$



Logarithmic Model

$$\gamma(h) = C\left[\log_e(h)\right] \quad h > 0$$



Spherical Model

$$\gamma(h) = \begin{cases} C\left[1.5h - 0.5h^3\right] & h < 1 \\ C & h \geq 1 \end{cases}$$

where C is the sill and equal to σ^2.

A nugget effect can be introduced by the addition of a term a that represents a discontinuity at the origin. For keeping $\lambda(0) = 0$, the models are adapted as :

$$\gamma(h) = \begin{cases} a + (\sigma^2 - a)\left(\dfrac{3h}{2r} - \dfrac{h^3}{2r^3}\right) & 0 < h \leq r \\ 0 & h = 0 \\ \sigma^2 & otherwise \end{cases}$$

for the spherical model, and

$$\gamma(h) = \begin{cases} a + (\sigma^2 - a)(1 - \mathrm{e}^{-3h/\ r}) & h > 0 \\ 0 & h = 0 \end{cases}$$

for the exponential model. A similar adjustment is applied to the Gaussian model.

The ratio of the nugget effect to the sill is often referred as the *relative nugget effect* and is usually expressed as a percentage. A variogram model with a pure nugget effect ($a = \sigma^2$, with a relative nugget effect of 100 %) corresponds to a complete lack of spatial dependence.

Complex variogram models can thus be constructed from basic forms. The new model is a linear combination of other models.

The previously presented models are for isotropic processes. Models for anisotropic covariance structures are similarly defined by replacing their parameters by the vector equivalents and the distance *h* by the vector separation $\vec{h}$. Generally, the range may change with direction while the sill remains constant to give a *geometrical anisotropy,* or the sill may change with direction while the range remains constant in the case of a *zoned anisotropy*. Combination of both of these effects is also possible. The various directional models would first be established by fitting the models along the identified axes of anisotropy. Then, these models need to be scaled and combined into a single vector model that is consistent in all directions.

To conclude, it is important to note that, in practice, the variogram models are often fitted by eye to the observed sample variogram.

2.3.3. Generalised least squares combined with covariance modelling

In the previous section, some theoretical models were presented for the variogram modelling. The obtained covariance matrix can be now incorporated in the model for the trend surface analysis. The proposed model for *Y(s)* is :

$$Y(s) = x^T(s)\beta + U(s)$$

where $x^T(s)\beta$ is a trend surface representing the mean *μ(s)* or first order component, and *U(s)* represents a local or second order component. The assumption was made that this second order effect is a stationary zero mean stochastic process with variance σ^2 and covariance function *C*().

In practice, this model can be fitted to the observed data, by first fitting the model using ordinary least squares regression, taking the residuals from this and using these to estimate a variogram model $\hat{\gamma}(h)$ giving rise to an equivalent covariogram model $\hat{C}(h)$. This covariogram model allows constructing an estimated covariance matrix $\hat{C}$ between sample sites with elements $\hat{C}(s_i, s_j)$. Then, the original model can be refitted using generalised least squares with the estimated covariance matrix. Doing this, the parameter estimates are corrected and the same for the standard errors of the ordinary least squares regression for second order effects. If necessary, the process can be iterated until arriving at stable estimates of $\hat{\beta}$ and $\hat{C}(h)$. The validity of the final model depends upon both the choice of an appropriate form of trend surface and the choice of an appropriate variogram model.

What is the objective ?

Two ways are now possible depending on the final objective. Firstly, if the primary interest is understanding and describing the nature of the variation in the observed values and isolating any systematic large scale trend, then knowledge of the trend component and the form of the covariance structure of variations from this trend is enough to pursue the modelling process. On the other hand, if the primary interest is in prediction or interpolation of the attribute value at points where it has not been sampled, then it is now necessary to consider how to use the derived model for prediction purposes. This second objective requires the use of the technique of *kriging* that will be detailed in the next section.

2.3.4. Simple kriging

The name of kriging derives from the South African mining geologist D.G. Krige who developed a preliminary version of the method, later refined by French geostatisticians.

As said, with kriging, the primary interest lies in the prediction of the values of a spatially continuous variable. For a process having a model as $Y(s) = x^T(s)\beta + U(s)$, where *U(s)* is a zero mean process with a covariance function *C*(), the values of *U(s)* are not entirely unpredictable. The prediction can be better than the estimated mean value $\mu(s) = x^T(s)\hat{\beta}$. The knowledge of the covariance function of the residual process *U(s)* combined with the knowledge of its observed value at the sample point s_i, could permit to add a local component to the prediction at the point *s*. This technique is called *kriging*.

In the simple kriging, the assumption is made that the first order component μ(s) in the model is known a priori and does not have to be estimated from the observed data.

This *a priori* known mean is subtracted from the original sample observations y_i to provide a set of observed residuals u_i. Moreover, it is assumed that *U(s)* has a known variance σ^2 and a covariance function *C*(). In simple kriging, the problem is reduced to find an estimate $\hat{u}(s)$ for a value *u(s)* of the random variable *U(s)* at the location s, given observed values u_i of the random variables $U(s_i)$ at the N sample locations s_i. When a suitable estimate is established, the prediction $\hat{y}(s)$ of the random variable *Y(s)* can be obtained by adding $\hat{u}(s)$ to the known trend *μ(s)* at point *s*.

The estimates can be considered as weighted linear combinations of the observed residuals u_i :

$$\hat{U}(s) = \sum_{i=1}^{N} \lambda_i(s) U(s_i)$$

The first thing to notice about $\hat{U}(s)$ is that its mean value is zero for any choice of weights, since, by assumption, the mean of each $U(s_i)$ is zero. This means that, on average, the value of $\hat{U}(s)$ will be zero, which is a desirable property given that it is wished to find this random variable to be close as possible to *U(s)* whose mean value is by definition also zero. Secondly, it is needed to study how the values of the random variable $\hat{U}(s)$ differ from those of *U(s)*. It can be measured by the *expected mean square error* between values of *U(s)* and values of $\hat{U}(s)$. Given that the random variables U(s) and U(s_i) have zero mean, this expected mean square error is :

$$\begin{aligned} E((\hat{U}(s) - U(s))^2) &= E(\hat{U}^2(s)) + E(U^2(s)) - 2E(\hat{U}(s)U(s)) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i(s)\lambda_j(s) C(s_i, s_j) + \sigma^2 - 2\sum_{i=1}^{N} \lambda_i(s) C(s, s_i) \\ &= \lambda^T(s) C \lambda(s) + \sigma^2 - 2\lambda^T(s) c(s) \end{aligned}$$

where *C* is the (N x N) matrix of covariances $C(s_i,s_j)$ between all possible pairs of the N samples sites, and *c(s)* is a (N x 1) column vector of covariances $C(s,s_i)$ between the prediction point *s* and each of the N samples sites. Since $\hat{U}(s)$ must be close to $U(s)$, the weights *$\lambda_i(s)$* are chosen in order to minimise this mean square error. Differentiating with respect to the vector *$\lambda(s)$* gives the solution :

$$\lambda(s) = C^{-1}c(s)$$

from which we have for $\hat{U}(s)$:

$$\hat{U}(s) = \lambda^T(s)U = c^T(s)C^{-1}U$$

Using these weights, the minimised expected mean square error is then :

$$E((\hat{U}(s) - U(s))^2) = \sigma^2 - c^T(s)C^{-1}c(s)$$

where $\sigma^2 = \mathrm{VAR}(U(s)) = C(s,s)$. This minimised mean square error is the *kriging variance* noted σ_e^2 .

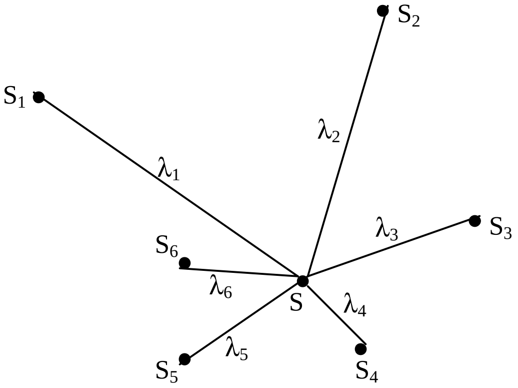


Figure 5. Example of kriging weights at location *s*.

A linear combination $\hat{U}(s)$ of the random variables *$U(s_i)$* has been obtained, which has the same mean value as *U(s)* and where the expected mean square error between the values of *U(s)* and $\hat{U}(s)$ is minimised amongst all such linear combinations. This means that we can estimate a value of $\hat{U}(s)$ instead of the needed *U(s)*. An estimated value $\hat{u}(s)$ of the random variable $\hat{U}(s)$, can be obtained simply by replacing of *U* in the above expression by the vector of observed residuals *u_i*. This estimate $\hat{u}(s)$ is then added to the known trend *μ(s)* to obtain the final prediction $\hat{y}(s)$. Since *μ(s)* is assumed known subject to no prediction error, the standard error associated with $\hat{y}(s)$ is simply that associated with $\hat{u}(s)$.

Unfortunately, in practice, the simple kriging method is not very useful because it depends on the prior knowledge of both the trend *μ(s)* of the process *Y(s)* and its covariance function *C*(). In reality, both are unknown and have to be estimated from the data.

2.3.5. General spatial prediction methodology

When the first order or trend component $\mu(s) = x^T(s)\beta$ of the process *Y(s)* is not known *a priori* and has to be estimated from the data, simple kriging cannot be applied. We know that this estimate can be derived through generalised least squares. In simple kriging, it was also assumed that the covariance function *C*() was known while, in practice, this function has also to be estimated from the data. Fortunately, it can be done using techniques for fitting variogram models. All these require the assumption of stationarity, and if the first component is not a constant over $\Re$, this component must be firstly removed from the observations for working with the residuals to estimate the variogram.

From a practical point of view, how can we resume and combine all these techniques ?

Let be *y* a (N x 1) column vector of observations, on the corresponding vector of random variables *Y(s$_i$)* at each of the N sample sites *s$_i$*. We can now put together all the techniques for variogram modelling, trend surface estimation, and simple kriging to give an overall strategy for spatial prediction.

1. The first step is the fitting of a suitable trend surface model : $Y(s) = x^T(s)\beta + \varepsilon(s)$ by ordinary least squares to obtain $\hat{\beta}$ and the residuals $y - x^T(s_i)\hat{\beta}$.

2. The sample variogram is estimated from these residuals and a suitable model is fitted to this. It is then easy to obtain the corresponding covariogram model $\hat{C}()$ from which the matrix $\hat{C}(s_i, s_j)$ is derived.

3. The trend surface model $Y(s) = x^T(s)\beta + U(s)$ is estimated by generalised least squares using the matrix $\hat{C}$, and revised $\hat{\beta}$ and revised residuals $\hat{u} = y - x^T(s_i)\hat{\beta}$ are obtained.

4. If necessary, steps 2 and 3 are iterated until reasonable stability is obtained for $\hat{\beta}$ and $\hat{C}()$. The final residuals $\hat{u}$ are derived and the final estimate of $\hat{C}$ is calculated. Then, the covariance matrix is inverted to obtain $\hat{C}^{-1}$.

5. For any location *s*, at which prediction is required, the estimated vector $\hat{c}(s)$ of covariances $\hat{C}(s, s_i)$ between location *s* and the N sample locations *s$_i$*, is constructed. Then, *Y(s)* is predicted by analogy with simple kriging :

$$\hat{y}(s) = x^T(s)\hat{\beta} + \hat{u}(s) = x^T(s)\hat{\beta} + \hat{c}^T(s)\hat{C}^{-1}\hat{u}$$

6. The kriging variance or mean square prediction error is calculated by :

$$\sigma_e^2 = ((x(s) - X^T\hat{C}^{-1}\hat{c}(s))^T (X^T\hat{C}^{-1}X)^{-1}(x(s) - X^T\hat{C}^{-1}\hat{c}(s))) + (\hat{\sigma}^2 - \hat{c}^T(s)\hat{C}^{-1}\hat{c}(s))$$

with $\hat{\sigma}^2 = \hat{C}(s,s)$. In this expression, the second term is just the kriging variance of the simple kriging. The first term is an extra component that arises because now the trend surfaces has also been estimated from the data.

7. Steps 5 and 6 may be repeated for as many locations *s* as required.

2.3.6. Ordinary kriging

In this section, contrary to the simple kriging, methods will be focused on optimal local prediction without specific identification of any global first order component of spatial variation. Instead, first order effects are simultaneously and implicitly estimated as part of the prediction process. *Ordinary kriging* is mathematically equivalent to the previous approach when it is assumed that the process *Y(s)* has a constant mean value $\mu(s) = \mu$. This corresponds to the trend model when *x(s)* is the scalar 1 and $\beta_1=\mu$. However, rather than estimate μ by generalised least squares and then predict the zero mean residual process by simple kriging, adding back $\hat{\mu}$ at the final stage, ordinary kriging instead forms the prediction $\hat{y}(s)$ in one step by use of a weighted linear combination of the observed values y_i.

Most of the discussion is the same as for the simple kriging. The weighted linear combination of random variables is :

$$\hat{Y}(s) = \sum_{i=1}^{N} \omega_i(s) Y(s_i)$$

The values of the weights are then chosen so that the mean value of $\hat{Y}(s)$ is constrained to be μ. So, as before, it is required to minimise the expected mean square error between *Y(s)* and $\hat{Y}(s)$. The expression is similar to that used for simple kriging :

$$\begin{aligned} E((\hat{Y}(s) - Y(s))^2) &= E(\hat{Y}^2(s)) + E(Y^2(s)) - 2E(\hat{Y}(s)Y(s)) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} \omega_i(s)\omega_j(s)C(s_i,s_j) + \sigma^2 - 2\sum_{i=1}^{N}\omega_i(s)C(s,s_i) \\ &= \omega^T(s)C\omega(s) + \sigma^2 - 2\omega^T(s)c(s) \end{aligned}$$

where $C(s_i,s_j)$ is the (N x N) matrix of covariances, and *c(s)* is a (N x 1) vector of covariances $C(s,s_i)$.

With the constraint $\sum_{i=1}^{N} \omega_i(s) = 1$ or $\omega^T 1 = 1$, the previous expression must be minimised.
To do this, a Lagrange multiplier is introduced, let say *v(s)* :

$$\omega^T(s)C\omega(s) + \sigma^2 - 2\omega^T(s)c(s) + 2(\omega^T 1 - 1)v(s)$$

Differentiating with respect to both $\omega(s)$ and *v(s)* leads to the two simultaneous equations :

$$\omega^T 1 = 1$$
$$C\omega(s) + 1v(s) = c(s)$$

This expression can be reformulated on the form of a single equation by using an augmented C_+ matrix and augmented vectors $\omega_+(s)$ and $c_+(s)$:

$$C_+ \qquad \omega_+(s) \quad = \quad c_+(s)$$

$$\begin{pmatrix} C(s_1,s_1) & \cdots & C(s_1,s_N) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ C(s_N,s_1) & \cdots & C(s_N,s_N) & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} \omega_1(s) \\ \vdots \\ \omega_N(s) \\ v(s) \end{pmatrix} = \begin{pmatrix} C(s,s_1) \\ \vdots \\ C(s,s_N) \\ 1 \end{pmatrix}$$

and the required weights are given by :

$$\omega_+(s) = C_+^{-1} c_+(s)$$

With this choice of weights, the minimised expected mean square error is :

$$\sigma_e^2 = \sigma^2 - c_+^T(s) C_+^{-1} c_+(s)$$

The prediction is then : $\hat{y}(s) = \omega^T(s) y$.

2.3.7. Universal kriging

With the ordinary kriging, it was assumed a constant mean value $\mu(s)=\mu$ for the process $Y(s)$. The natural extension to ordinary kriging is *universal kriging*. Universal kriging is mathematically equivalent to the general case of the previous section where a first order trend component $\mu(s)=x^T\beta$ is included. As before, $x(s)$ is a (p x 1) vector $(x_1(s), \ldots, x_p(s))^T$.

The universal kriging, as the ordinary kriging, calculates the prediction $\hat{y}(s)$ directly in one step by use of a linear combination of the observed values y_i at the sample locations s_i. The discussion is similar as for the ordinary kriging, except that here p Lagrange multipliers are involved and the matrix C and vector $c(s)$ are augmented by p rows and columns as follows :

$$\begin{array}{cccc}
C_+ & \omega_+(s) & = & c_+(s) \\
\begin{pmatrix}
C(s_1,s_{1)} & \cdots & C(s_1,s_N) & x_1(s_1) & \cdots & x_p(s_1) \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
C(s_N,s_1) & \cdots & C(s_N,s_N) & x_1(s_N) & \cdots & x_p(s_N) \\
x_1(s_1) & \cdots & x_1(s_N) & 0 & \cdots & 0 \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
x_p(s_1) & \cdots & x_p(s_N) & 0 & \cdots & 0
\end{pmatrix}
&
\begin{pmatrix}
\omega_1(s) \\ \vdots \\ \omega_N(s) \\ v_1(s) \\ \vdots \\ v_p(s)
\end{pmatrix}
& = &
\begin{pmatrix}
C(s,s_1) \\ \vdots \\ C(s,s_N) \\ x_1(s) \\ \vdots \\ x_p(s)
\end{pmatrix}
\end{array}$$

Then the solution for the weights is :

$$\omega_+(s) = C_+^{-1} c_+(s)$$

and the mean square prediction error is :

$$\sigma_e^2 = \sigma^2 - c_+^T(s) C_+^{-1} c_+(s)$$

which provides a standard error and confidence interval for the prediction.

The required prediction is given by :

$$\hat{y}(s) = \omega^T(s) y$$

2.3.8. Co-kriging

This section and the following intend to consider a different problem from those previously investigated. It refers to the case where several variables are measured at each sample site and where the prediction of the value of one of them is not taken in isolation of the others but taking account of these other variables.

In the previous sections, only one variable was considered but, in practice, more than one can be sampled at the location s_i. The secondary variables may even be available at a more extensive set of locations than the set applied for the primary variable. The different sets of locations are not always the same. All the cases can be represented mathematically as having observations on a primary variable y_i at N sample sites and additional observations on p possible covariates (x_{i1}, …, x_{ip}) at a more extensive set of N + M sites. If M = 0, both of the primary variable and the covariates exist at all sites. The problem addressed in co-kriging is to know how the covariate information can improve the prediction of the primary variable at an unsampled location *s*. The developed technique is simply an extension to ordinary kriging. The case of p = 1 will only be detailed in this section, extension to the case where there is more than one secondary variable can be obtained by analogy with the single variable case.

The concept of variogram and covariogram presented in the previous sections is extended to a cross-covariogram and cross-variogram between two variables. Formally, if *Y(s)* is the process relating to the primary variable and *X(s)* that relating to the secondary variable, and both these processes are assumed stationary, then the cross-covariogram is defined by :

$$C_{YX}(h) = E((Y(s+h) - \mu_Y)(X(s) - \mu_X))$$

where *h* is an arbitrary vector separation. Similarly, the concept of variogram is extended to a cross-variogram defined as :

$$2\gamma_{YX}(h) = E((Y(s+h) - Y(s))(X(s+h) - X(s)))$$

It is important to note that in general, the cross-covariogram need not be symmetric, that is

$$C_{YX}(h) \neq C_{XY}(h)$$

whereas the cross-variogram is always symmetric, so that

$$\gamma_{YX}(h) = \gamma_{XY}(h)$$

In practice, the cross-covariogram is often modelled as a symmetric function and this model is usually derived from an estimated cross-variogram model by the relationship :

$$C_{YX}(h) = \gamma_{XY}(\infty) - \gamma_{XY}(h)$$

The natural sample estimator of the cross-variogram, given N pairs of observations (y_i,x_i) at sample sites s_i, is :

$$2\hat{\gamma}_{YX}(h) = \frac{1}{N(h)} \sum_{s_i - s_j = h} (y_i - y_j)(x_i - x_j)$$

where *h* is the vector separation between pairs of points, and *N(h)* is the number of pairs.

For the ordinary kriging, it was assumed that the process *Y(s)* had a constant mean value *μ(s)* = *μ*. In the co-kriging extension, it is assumed that both the primary and secondary processes have constant mean values μ_Y and μ_X. The prediction $\hat{y}(s)$ is calculated in terms of a linear combination of both values of the processes y_i (with i = 1, …, N) and x_i (with i = 1, …, N+M). The appropriate optimal weights are derived by analogy with the ordinary kriging weights.

$$\hat{Y}(s) = \sum_{i=1}^{N} \omega_{yi}(s) Y(s_i) + \sum_{j=1}^{N+M} \omega_{xj}(s) X(s_j)$$

The weights $\omega_{yi}(s)$ and $\omega_{xj}(s)$ are chosen to obtain an $\hat{Y}(s)$ with a mean value equal to μ_Y, and to minimise the mean square error between the values of $Y(s)$ and $\hat{Y}(s)$. It is also assumed that the mean of *Y(s)* and of each *Y(s_i)* are all μ_Y, and that the mean of the $X(s_j)$ are all μ_X. One way to obtain such requirement for the mean value of $\hat{Y}(s)$, is to ensure that

$$\sum_{i=1}^{N} \omega_{yi}(s) = 1 \quad \text{and} \quad \sum_{j=1}^{N+M} \omega_{xj}(s) = 0$$

As in the ordinary kriging, the expected mean square error has to be minimised. This can be obtained by resolving a system of equations where now two Lagrange multipliers *$v_1(s)$* and *$v_2(s)$* are involved :

$$C_+ \omega_+(s) = c_+(s)$$

where the augmented matrix C₊ is :

$$C_+ = \begin{pmatrix} C_Y(s_1,s_1) & \cdots & C_Y(s_1,s_N) & C_{YX}(s_1,s_1) & \cdots & C_{YX}(s_1,s_{N+M}) & 1 & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ C_Y(s_N,s_1) & \cdots & C_Y(s_N,s_N) & C_{YX}(s_N,s_1) & \cdots & C_{YX}(s_N,s_N) & 1 & 0 \\ C_{XY}(s_1,s_1) & \cdots & C_{XY}(s_1,s_N) & C_x(s_1,s_1) & \cdots & C_X(s_1,s_{N+M}) & 0 & 1 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ C_{XY}(s_{N+M},s_1) & \cdots & C_{XY}(s_{N+M},s_N) & C_X(s_{N+M},s_1) & \cdots & C_X(s_{N+M},s_{N+M}) & 0 & 1 \\ 1 & \cdots & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & 0 & 0 \end{pmatrix}$$

and the augmented vectors are :

$$\omega_+(s) = \begin{pmatrix} \omega_{y1}(s) \\ \vdots \\ \omega_{yN}(s) \\ \omega_{x1}(s) \\ \vdots \\ \omega_{xN+M}(s) \\ v_1(s) \\ v_2(s) \end{pmatrix} \quad \text{and} \quad c_+(s) = \begin{pmatrix} C_Y(s,s_1) \\ \vdots \\ C_Y(s,s_N) \\ C_{YX}(s,s_1) \\ \vdots \\ C_{YX}(s,s_{N+M}) \\ 1 \\ 0 \end{pmatrix}$$

where $\omega_+(s) = C_+^{-1} c_+(s)$.

Then, the prediction is given by $\hat{y}(s) = \omega_y^T(s) y + \omega_x^T(s) x$,

And the kriging variance is given by $\sigma_e^2 = C_Y(s,s) - c_+^T(s) C_+^{-1} c_+(s)$.

2.3.9. Multivariate method : principal components analysis

With observations ($y_{i1}, \ldots, y_{ip}$) on p attributes at each of the N sample sites and following the conventional notation in multivariate analysis, the (N x p) matrix Y is used to refer these observations collectively. In this matrix, each row is a (1 x p) vector y_i^T of observations on the p attributes at the *i*th site. The matrix Y can be viewed as a set of N points in a p-dimensional data space where ($y_{i1}, \ldots, y_{ip}$) are the p coordinates of the *i*th observation in this space.

The principal components analysis identifies a small number of linear combinations of variables that account for a large proportion of the variability of observations in the data space. The first few principal components are a projection of the original data space onto a subspace of lower dimension that retains the most important separation of observations. The principal components allow interpreting the subspace in terms of the original variables.

With p original variables grouped into a (N x p) matrix Y as defined above, we can define variables of the subspace of dimension q ≤ p as linear combinations of the original variables as :

$$\begin{aligned}
u_1 &= a_{11}y_1 + a_{12}y_2 + \ldots + a_{1p}y_p \\
u_2 &= a_{21}y_1 + a_{22}y_2 + \ldots + a_{2p}y_p \\
\vdots &= \vdots \\
u_q &= a_{q1}y_1 + a_{q2}y_2 + \ldots + a_{qp}y_p
\end{aligned}$$

where the a_{jk} are constants. These new co-ordinate axes must be orthogonal and the (q x p) matrix of coefficients a_{jk} is also orthogonal. This condition implies that the row vector a_j^T satisfy $a_j^T a_j = 1$ for $j = 1, \ldots,$ p, and $a_j^T a_k = 0$ for $j \neq k$. The required solution of the problem is then to find the vector of weights a_1 that makes the variance of the transformed observations u_{i1} as large as possible, then to choose a_2 so that u_{i2} has the next largest variance, with a_2 being orthogonal to a_1, and so on through to the choice of a_q. Mathematically, the solution of this problem can be resume to find the eigenvectors of a matrix. The a_j are the p eigenvectors, ranked in decreasing order of their eigenvalues, of the (p x p) matrix $S = Y^T Y / (N\text{-}1)$. The matrix S is the matrix of sample covariances between the p original variables. The corresponding eigenvalues λ_j are the variances of the observations in each of the new directions. Since each component is by definition the best linear summary of variance left in the data after the previous components are accounted, the first q components may explain most of the variance in the data. The proportion explained by the q first components is given by :

$$\frac{\sum_{j=1}^{q} \lambda_j}{\sum_{j=1}^{p} \lambda_j}$$

where $\sum_{j=1}^{p} \lambda_j$ is the overall original variance in the data.

Finally, it is important to note that the principal components analysis is not invariant to standardisation of the variables. Generally, a standardisation is recommended before a principal components analysis is performed. The eigenvalues and eigenvectors are then derived from the correlation matrix between the p original variables.

2.4. Summary

Spatial data analysis or geostatistic involves the analysis and prediction of spatial or temporal phenomena, such as metal grades, pollutant concentrations, *etc*. Geostatistics is a name associated with a class of techniques used to analyse and predict values of a variable distributed in space or time. Although the primary interest is often the prediction of values at unsampled locations, the early sections aimed to study and decompose the spatial variation into two components : first order variation or trend component; and second order effects. The first order components were analysed by techniques such as spatial moving averages, TIN interpolation, kernel smoothing. The sample variogram and covariogram were introduced for the second order components. After that, it was tried to model these variations. First order variations were investigated and modelled by trend surface analysis with the use of the standard ordinary least squares regression model. Then, the generalised least squares were introduced to allow second order variations. Such second order variations are implicitly assumed to be correlated with each other, and the study of such correlation is called variogram modelling. After structural analysis, predictions at unsampled locations are made using kriging or simulations.

To summarise, the steps in spatial data analysis are :

1. exploratory data analysis ;
2. structural analysis (calculation and modelling of variograms) ;
3. predictions (kriging or simulations).

2.5. Applications

Some available data extracted from the present IDOD database to produce spatio-maps for different parameters such as temperature and salinity using the techniques presented in the previous sections. Some of the obtained results are presented hereafter.

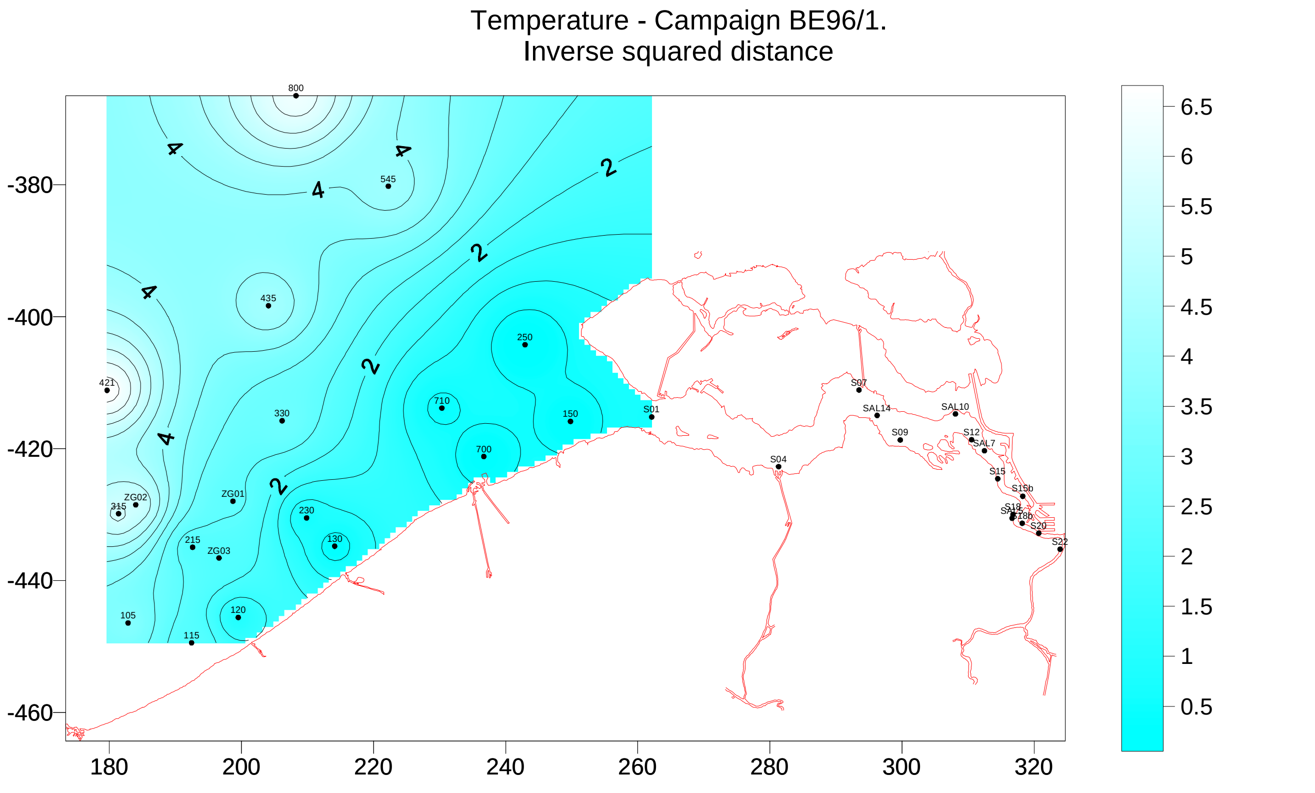
2.5.1. Sea temperature data

The following example use the temperature measurements collected on the sea stations during the first campaign BE96/1 in February 1996.

The following table show the data collected during this campaign and used for the interpolation procedures.

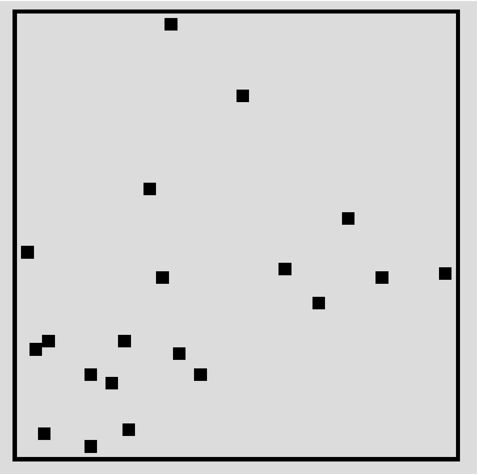
Latitude	Longitude	Coord. X	Coord. Y	Campaign name	TEMP Value	Station name	Sampling depth
51,37667	3,22	236,692765	-421,1828507	BE96/1	0,18	700	3,15
51,41617	3,402167	249,8220078	-415,8591627	BE96/1	0,27	150_a	4,38
51,30833	2,85	209,8755814	-430,4824466	BE96/1	0,66	230	4,15
51,27083	2,905	214,1217562	-434,7722493	BE96/1	0,27	130	4,61
51,185	2,701167	199,5334563	-445,5788462	BE96/1	0,99	120	4,43
51,155	2,603333	192,4544916	-449,4209383	BE96/1	2,28	115_a	4,02
51,26167	2,666667	196,6126666	-436,5607969	BE96/1	2,15	ZG03	3,74
51,33333	2,7	198,7143425	-427,9498631	BE96/1	2,56	ZG01	4,61
51,27667	2,613333	192,6113726	-434,9261394	BE96/1	2,24	215_a	3,83
51,18333	2,475	182,845078	-446,3956636	BE96/1	3,49	105	4,56
51,33333	2,5	184,0044782	-428,4970329	BE96/1	5,43	ZG02	4,52
51,32283	2,464	181,4038228	-429,8389262	BE96/1	5,63	315	4,24
51,4805	2,45	179,6664101	-411,1440489	BE96/1	6,71	421	5,25
51,58067	2,790333	204,0936362	-398,3195215	BE96/1	4,33	435	3
51,84717	2,866667	208,2722718	-366,4939336	BE96/1	6,26	800	5,21
51,72667	3,05	222,2534307	-380,2114217	BE96/1	4,19	545	5,25
51,51667	3,316667	242,9405073	-404,2320558	BE96/1	0,05	250	4,75
51,44083	3,138667	230,351596	-413,835963	BE96/1	0,39	710_a	4,93
51,43333	2,808333	206,1670052	-415,7573017	BE96/1	2,89	330_a	4,66
51,41667	3,57	262,127184	-415,1855986	BE96/1	0,60	S01	4,7

A mapping is firstly realised by the common *Inverse Squared Distance* interpolation method. This method has the advantage to be fast to run but has also artefacts as the "bull eyes" effect easily visible on the resulting gridding map.

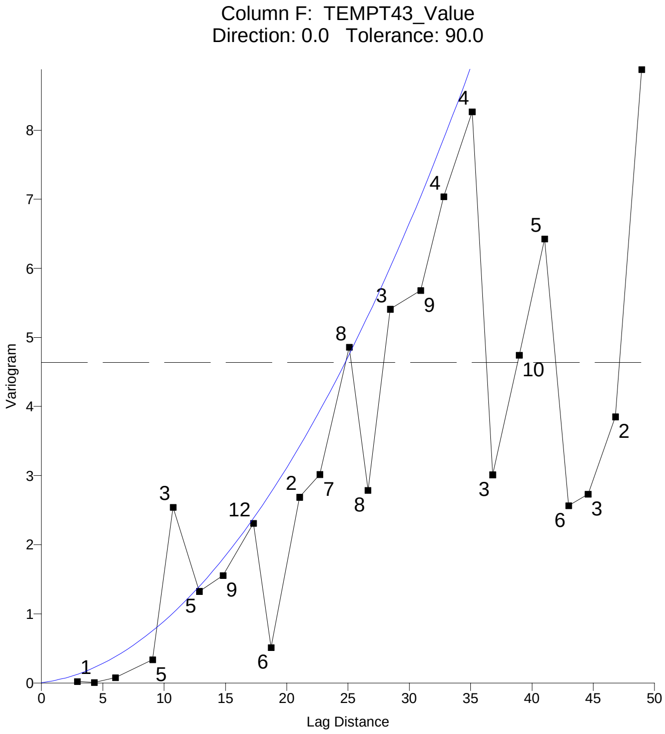


A second technique was then applied to obtain better results. This technique is kriging. At the first step, a statistical analysis was calculated on the data. These statistics are given at the end of this section.

In the sea area, the data are distributed as shown in the XY scatterplot :



A variogram is calculated and a theoretical variogram model is chosen to obtain the best fit.



When a model is selected with the appropriated parameters values, the kriging process can be run to produce a grid. The selected parameters are :

Component Type:	Nugget
Error Variance:	0
Micro Variance:	0

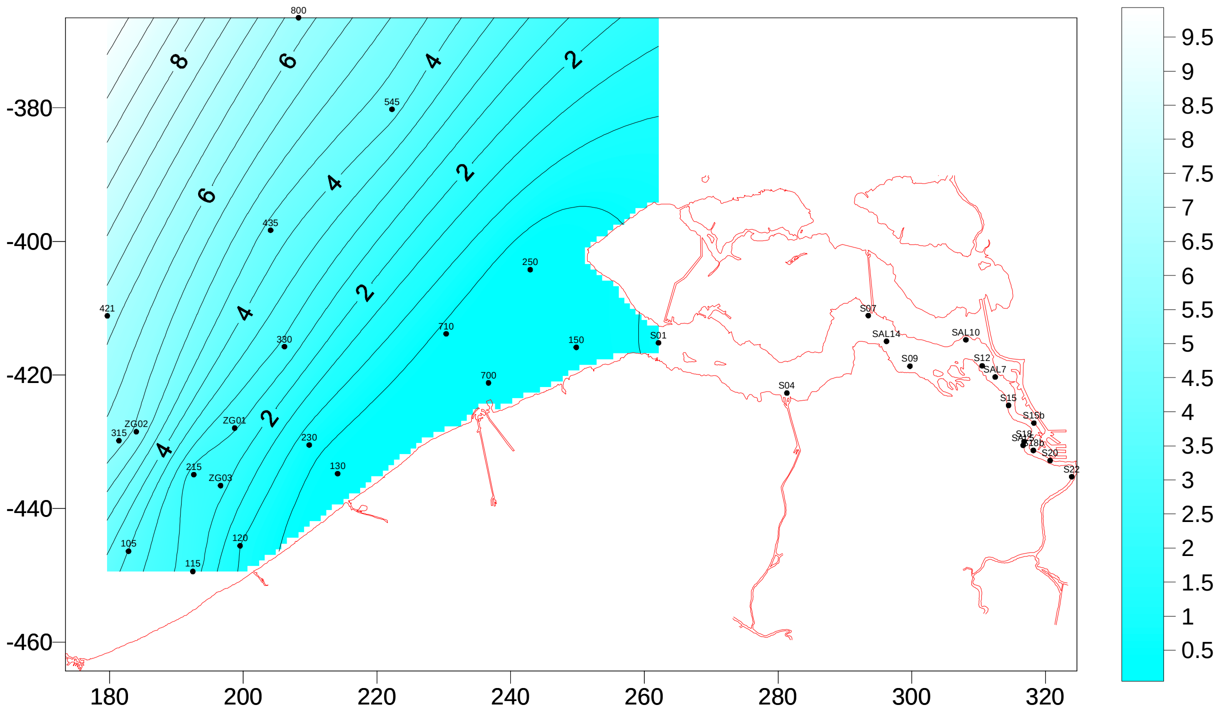
Component Type:	Linear
Variogram Slope:	0.0125
Anisotropy Angle:	59.9
Anisotropy Ratio:	2

Component Type:	RationalQuadratic
Variogram Scale:	1290
Anisotropy Angle:	52.25
Anisotropy Length:	746.1
Anisotropy Ratio:	2

Polynomial Drift Order:	0
Kriging standard deviation grid:	no

The kriging technique applied to the data lead to a grid used in the next mapping.

Temperature - Campaign BE96/1.
Kriging



The calculated statistics required to apply kriging are :

Variogram Grid Report

Detrending method:	None
Maximum Grid Lag Distance:	50
Number of Angular Divisions:	180
Number of Radial Divisions:	100

Data Filter Report

Source Data File Name:	D:\Mer\AnSp\Temperature-96-1-mer.xls
X Column:	C
Y Column:	D
Z Column:	F

Data Counts

Number of Active Data:	20
Number of Original Data:	20
Number of Excluded Data:	0
Number of Deleted Duplicates:	0
Number of Retained Duplicates:	0
Number of Artificial Data:	0

Filter Rules

Duplicate Points to Keep:	All
X Duplicate Tolerance:	0
Y Duplicate Tolerance:	0
Exclusion Filter String:	Not In Use

Data Statistics Report

Data Counts

Number of Active Data:	20
Number of Original Data:	20
Number of Excluded Data:	0
Number of Deleted Duplicates:	0
Number of Retained Duplicates:	0
Number of Artificial Data:	0

X Variable Statistics

X Range:	82.4608
X Midrange:	220.897

X Minimum:	179.666
X 25%-tile:	192.454
X Median:	204.094
X 75%-tile:	222.253
X Maximum:	262.127

X Average:	209.728
X Standard Deviation:	23.3564
X Variance:	545.52

Y Variable Statistics

Y Range:	82.927
Y Midrange:	-407.957

Y Minimum:	-449.421
Y 25%-tile:	-434.926
Y Median:	-427.95
Y 75%-tile:	-413.836
Y Maximum:	-366.494

Y Average:	-420.332
Y Standard Deviation:	20.7516
Y Variance:	430.63

Z Variable Statistics

Z Range:	6.66
Z Midrange:	3.38

Z Minimum:	0.05
Z 25%-tile:	0.39
Z Median:	2.24
Z 75%-tile:	4.19
Z Maximum:	6.71

Z Average:	2.5785
Z Standard Deviation:	2.15358
Z Variance:	4.6379

Z Coef. of Variation:	0.835206
Z Coef. of Skewness:	0.492366

Inter-Variable Correlation

	X	Y	Z
X:	1	0.363517	-0.668106
Y:		1	0.341225
Z:			1

Inter-Variable Covariance

	X	Y	Z
X:	545.52	176.191	-33.6056
Y:		430.63	15.2494
Z:			4.6379

Planar Regression Statistics (AX+BY+C = Z)

A:	-0.0841616
A Standard Deviation:	0.00961777
B:	0.0698463
B Standard Deviation:	0.010825
C:	49.5882
C Standard Deviation:	5.61148
Mean square error:	0.875866
Coef. of multiple determination:	0.839478

Inter-Parameter Correlation

	A	B	C
A:	1	0.363517	-0.654222
B:		1	-0.941526
C:			1

Nearest Neighbour Statistics

Avg. distance to nearest neighbour:	10.4048
Min. distance to nearest neighbour:	2.92645
Max. distance to nearest neighbour:	19.5868
Gamma for nearest neighbours:	0.452647

Data Filter Report

Source Data File Name:	D:\Mer\AnSp\Temperature-96-1-mer.xls
X Column:	C
Y Column:	D
Z Column:	F

Data Counts

Number of Active Data:	20

Number of Original Data:	20
Number of Excluded Data:	0
Number of Deleted Duplicates:	0
Number of Retained Duplicates:	0
Number of Artificial Data:	0

Filter Rules

Duplicate Points to Keep:	First
X Duplicate Tolerance:	0
Y Duplicate Tolerance:	0

Exclusion Filter String:	Not In Use

Data Statistics Report

Data Counts

Number of Active Data:	20

Number of Original Data:	20
Number of Excluded Data:	0
Number of Deleted Duplicates:	0
Number of Retained Duplicates:	0
Number of Artificial Data:	0

X Variable Statistics

X Range:	82.4608
X Midrange:	220.897

X Minimum:	179.666
X 25%-tile:	192.454
X Median:	204.094
X 75%-tile:	222.253
X Maximum:	262.127

X Average:	209.728
X Standard Deviation:	23.3564
X Variance:	545.52

Y Variable Statistics

Y Range:	82.927
Y Midrange:	-407.957

Y Minimum:	-449.421
Y 25%-tile:	-434.926
Y Median:	-427.95
Y 75%-tile:	-413.836

Y Maximum:	-366.494
Y Average:	-420.332
Y Standard Deviation:	20.7516
Y Variance:	430.63

Z Variable Statistics

Z Range:	6.66
Z Midrange:	3.38
Z Minimum:	0.05
Z 25%-tile:	0.39
Z Median:	2.24
Z 75%-tile:	4.19
Z Maximum:	6.71
Z Average:	2.5785
Z Standard Deviation:	2.15358
Z Variance:	4.6379
Z Coef. of Variation:	0.835206
Z Coef. of Skewness:	0.492366

Inter-Variable Correlation

	X	Y	Z
X:	1	0.363517	-0.668106
Y:		1	0.341225
Z:			1

Inter-Variable Covariance

	X	Y	Z
X:	545.52	176.191	-33.6056
Y:		430.63	15.2494
Z:			4.6379

Gridding Report

Search Rules

Use All Data:	true

Gridding Rules

Gridding Method:	Kriging

```
Kriging Type:                       Point
```

Semi-Variogram Model

```
Component Type:                     Nugget
Error Variance:                     0
Micro Variance:                     0

Component Type:                     Linear
Variogram Slope:                    0.0125
Anisotropy Angle:                   59.9
Anisotropy Ratio:                   2

Component Type:                     Rational Quadratic
Variogram Scale:                    1290
Anisotropy Angle:                   52.25
Anisotropy Length:                  746.1
Anisotropy Ratio:                   2

Polynomial Drift Order:             0
Kriging standard deviation grid:    no
```

Grid Summary

```
Grid File Name:                     D:\Mer\AnSp\Temp 96-1 - test1\Kriging.grd

Minimum X:                          179.666
Maximum X:                          262.127

Minimum Y:                          -449.421
Maximum Y:                          -366.494

Minimum Z:                          -0.266265
Maximum Z:                          9.93411

Number of Rows:                     100
Number of Columns:                  99

Number of Filled Nodes:             9900
Number of Blanked Nodes:            0
Total Number of Nodes:              9900
```

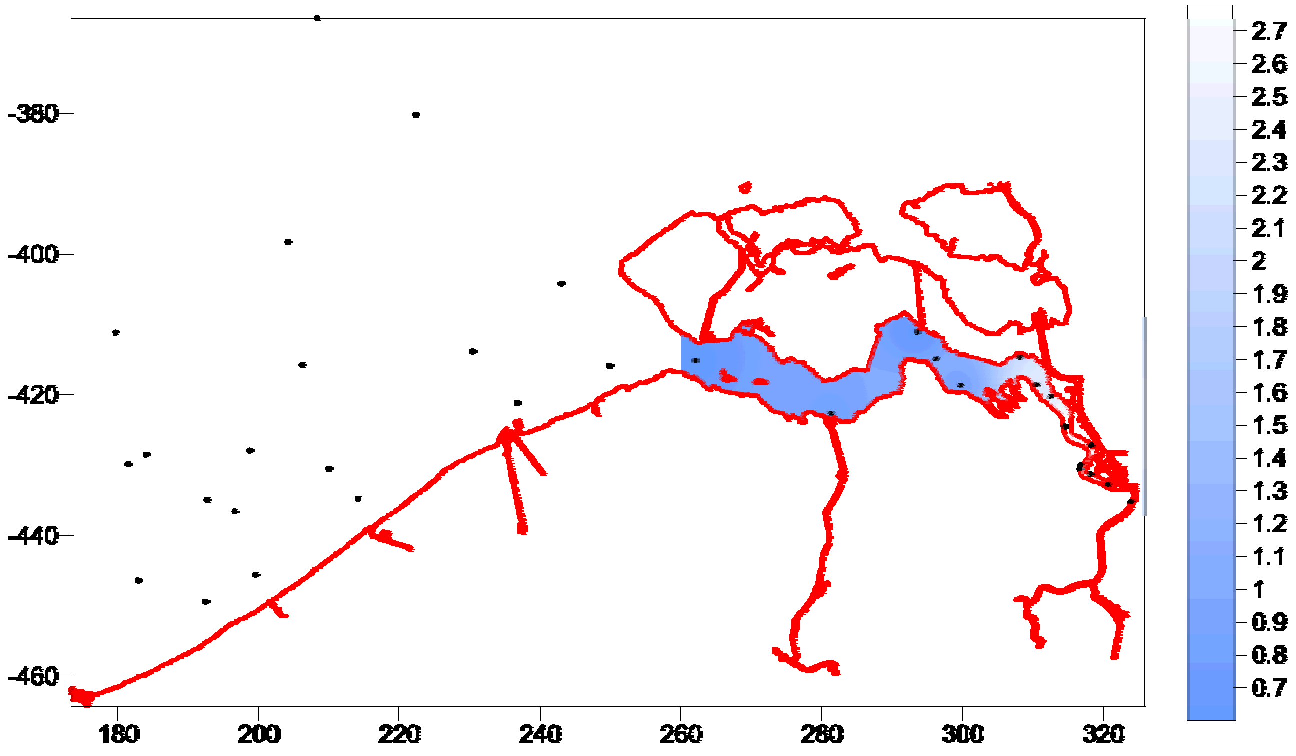
2.5.2. Scheldt temperature data

The following example use the temperature measurements collected on the sea stations during the first campaign BE96/1 in February 1996.

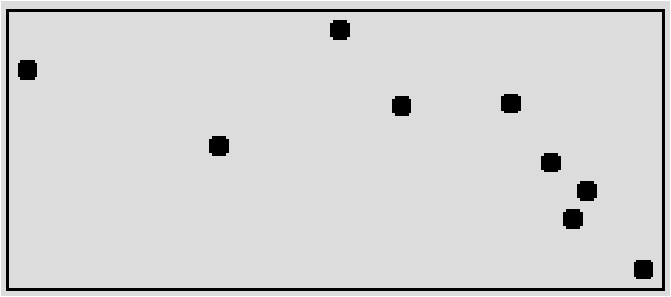
The following table show the data collected during this campaign and used for the interpolation procedures.

Latitude	Longitude	Coord. X	Coord. Y	Campaign name	TEMPT43 Value	Station name	Sampling depth
51,4167	3,57	262,127184	-415,1855986	BE96/1	0,6	S01	4,7
51,345	3,825	281,3253826	-422,7030699	BE96/1	0,82	S04	4,46
51,4367	4	293,504954	-411,1041658	BE96/1	0,61	S07	4,47
51,37	4,07833	299,7407329	-418,6832549	BE96/1	0,89	S09	3,64
51,365	4,225	310,5396049	-418,6367806	BE96/1	2,31	S12	4,88
51,3133	4,27333	314,4896098	-424,5530709	BE96/1	2,78	S15_a	3,83
51,2892	4,32233	318,2798492	-427,199211	BE96/1	2,75	S15b	3,92
51,2667	4,3	316,8154101	-429,9716357	BE96/1	2,57	S18	2,73
51,2188	4,39167	323,9404216	-435,2318489	BE96/1	2,22	S22	2,64

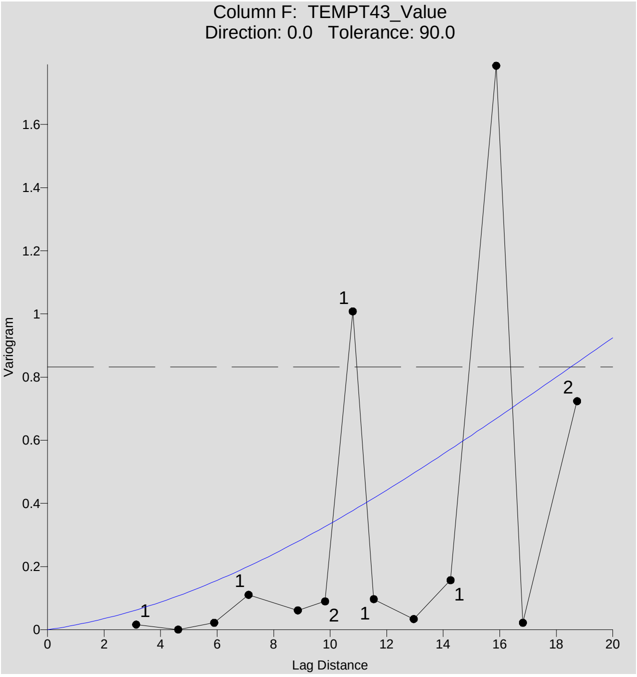
A mapping is firstly realised by the common *Inverse Squared Distance* interpolation method. This method has the advantage to be fast to run but has also artefacts as the "bull eyes" effect easily visible on the resulting gridding map.



The data distribution on the Scheldt is sketched by the XY scatterplot :



A variogram is calculated and a theoretical variogram model is chosen to obtain the best fit.



When a model is selected with the appropriated parameters values, the kriging process can be run to produce a grid. The selected parameters are :

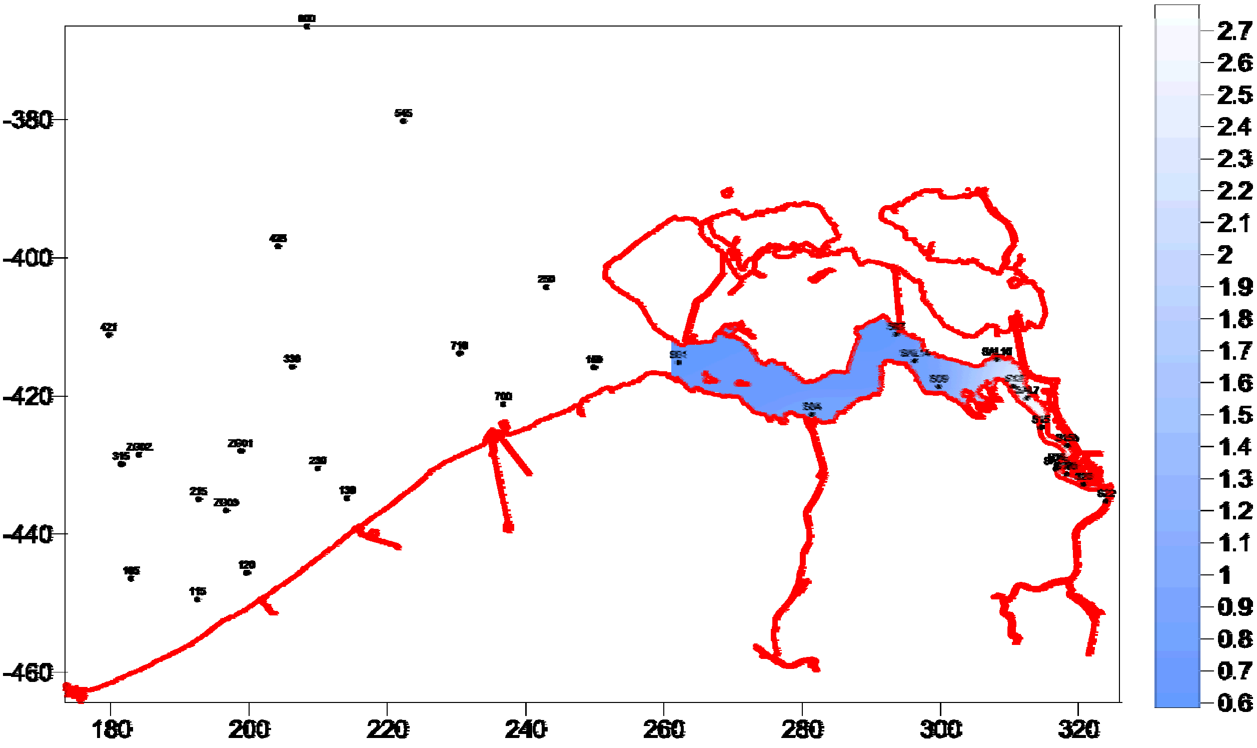
Component Type:	Nugget
Error Variance:	4.387E-11
Micro Variance:	0

Component Type:	Linear
Variogram Slope:	0.006913
Anisotropy Angle:	69.25
Anisotropy Ratio:	2

Component Type:	RationalQuadratic
Variogram Scale:	2.619
Anisotropy Angle:	65.93

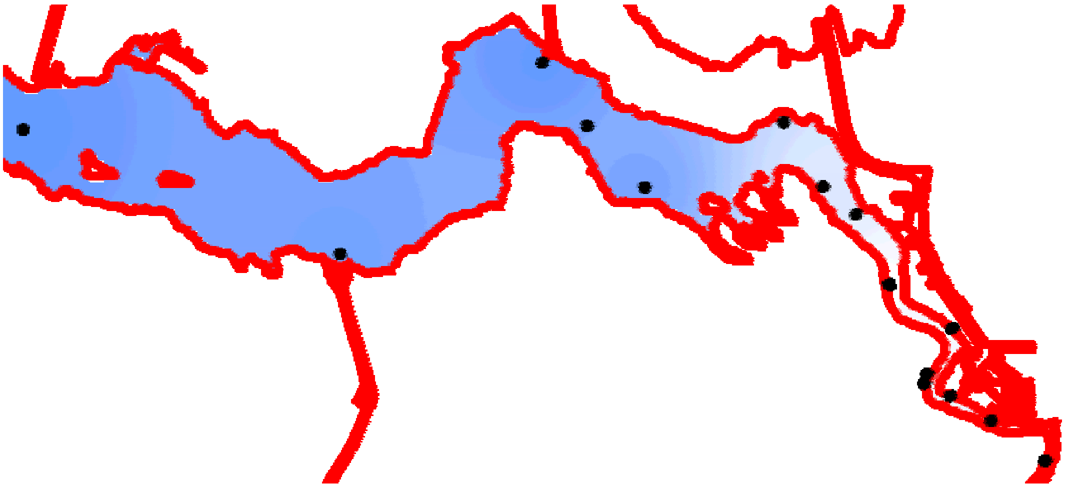
Anisotropy Length:	64.36
Anisotropy Ratio:	2
Polynomial Drift Order:	0
Kriging standard deviation grid:	no

The kriging technique applied to the data lead to a grid used in the next mapping.

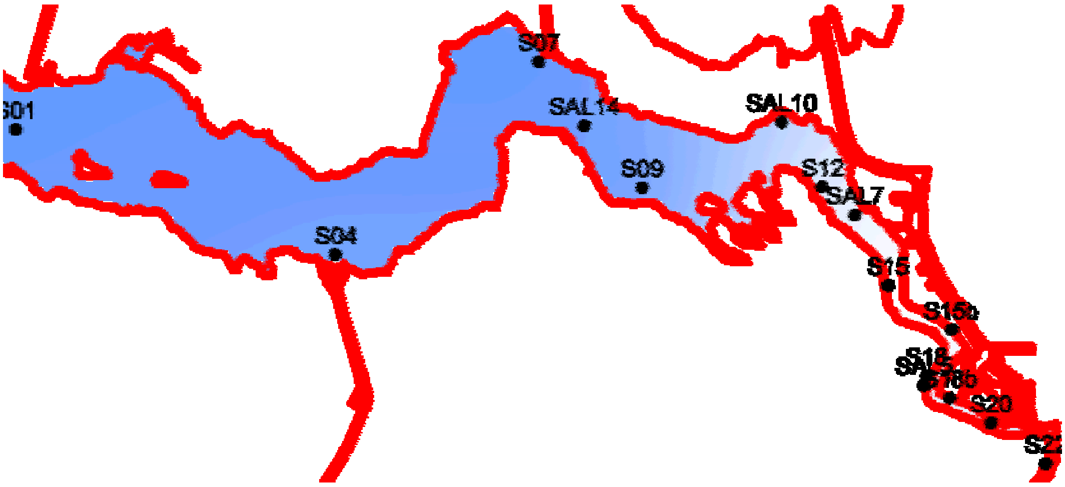


Here also, it is obvious that the result obtained by kriging is better than the one obtained by the inverse distance to a power interpolation.

The two considered methods can be examined with more details to better see all the differences and to assess which one produce the best mapping result. To do this, let consider the Scheldt area with kriging and inverse distance to a power.



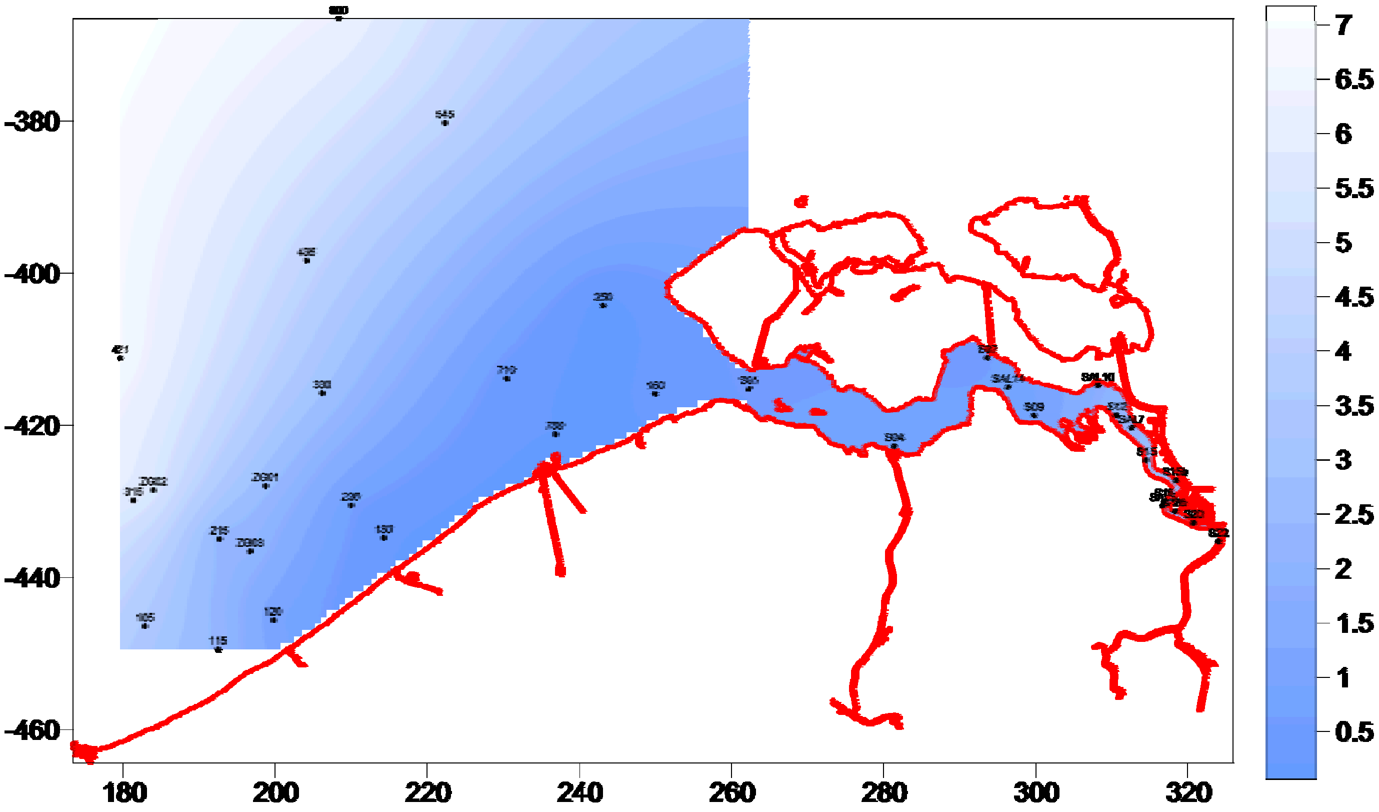
Inverse distance to power 2.



Kriging.

2.5.3. Sea and Scheldt temperature data

The results can now be merged to obtain a complete map of the area. This mapping is calculated by kriging.



3. The IDOD database

During the 1999-year, the prototype of the IDOD database widely presented in the last scientific report was adapted following to the new data and meta-data received by MUMM from the data providers. The general scheme was preserved and does not require a new complete description. When all the meta-data concerning biota and sediments will be available, some new entities will be added to the existing scheme and a new user's guide will be printed with the data dictionary describing all the entities of the database. At the present time, the scheme was modified according to the new specifications inherent to the data received during the year. These adaptations are minor and affect only the attributes inside the entities. No new entity was really created ; the new ones are only resulting of splitting of some of them.

As previously mentioned, the available geographical data were imported in the GIS software. They are not part of the alphanumerical IDOD database. The general structure involves two distinct databases. The GIS software directly manages the geographical database. To do this, the MUMM team purchased *ArcView*. This software was already available at SURFACES Laboratory. More complete and accurate geographical data are wanted to obtain good results in the spatial analysis and mapping procedures.

Bibliography

BAILEY Trevor & GATRELL Anthony (1995). *Interactive spatial data analysis*. Longman Scientific & Technical, Longman Group Limited, Essex, England.

BEGUIN Hubert (1979). *Méthodes d'analyse géographique quantitative*. Librairies Techniques (LITEC), Paris, France.

BENNET Spencer & BOWERS David (1976). *An introduction to multivariate techniques for social and behavioural sciences*. The MacMillan Press Ltd, London, England.

CRESSIE Noel (1991). *Statistics for spatial data*. John Wiley & Sons, USA.

DAVIS Bruce (1996). *GIS, a visual approach*. OnWord Press, Santa Fe, USA.

EMERY William & THOMSON Richard (1997). *Data analysis methods in physical oceanography*. Pergamon, Elsevier Science Ltd, Oxford, England.

GRIFFITH Daniel (1987). *Spatial autocorrelation. A primer*. Association of American Geographers, Washington, USA.

LEBART L., MORINEAU A. & FÉNELON J.-P. (1982). *Traitement des données statistiques. Méthodes et programmes*, Dunod, Paris.

UCS

Universitair Centrum voor Statistiek

Integrated and Dynamical Oceanographic Data Management - IDOD

Scientific Report
January 2000

Contribution of KUL-UCS

KATHOLIEKE UNIVERSITEIT
LEUVEN

University Center of Statistics
Katholieke Universiteit Leuven
de Croylaan 52b
B-3000 Leuven
Belgium

1. Introduction

During the third year of the IDOD project the work of UCS has concentrated primarily on developing the Statistical Analysis Tool (SAT program) and integrating this program as well as the SQC ("Statistical Quality Control") program with the IDOD database. This work is described in further detail in Section 3 (development of SAT program) and Section 4 (integration of various components). In addition, a method for the estimation of the variogram, which is to be used for the quality control of the data, has been developed. This method is described in further detail in Section 2 and has been presented at a workshop at the start of the year for the benefit of Ulg, who is responsible for the spatial mapping of the data and for whom the issue of spatial correlation is also of concern.

The work performed so far is summarized in Section 5. Handouts of the workshop on kriging and spatial interpolation are shown in Appendix 1. Screenshots of the various functions of the SAT program are reproduced in Appendix 2.

2. Methodology of spatial testing

As part of the statistical quality control, it is anticipated that for some of the variables testing may proceed by comparing the measured value with other (already qualified) measurements in the neighborhood. If there is sufficient spatial correlation, the already qualified measurements can be used to produce an estimate of the new measurement value and its possible variance. This in turn allows to verify whether the actual measurement value falls within reasonable bounds of the prediction.

The method to perform such interpolation is referred to as Kriging and is well known in the literature (i.e. Cressie, 1991). A key element in the application of this method is the knowledge of the so-called variogram which describes how the variance of the difference between two measurements increases with increasing distance between the two measurements. In the context of the IDOD application, this variogram will need to be repeatedly estimated for the different types of measurements available in the database and therefore a robust and reliable estimation procedure is needed. Furthermore, a particular complication due to the nature of the data is that these data may be irregularly spaced.

An extensive literature search has been made to better understand the various issues involved in a proper estimation of the variogram. After critical review, it is concluded that while some issues that are stressed in the literature are of less importance for the data in the IDOD database, some other issues are neglected. Issues that are thought to be less important are: 1. the presence of outliers (since the data in the IDOD database are extensively screened, both prior to entry as well as by the SQC program), 2. the actual shape of the variogram (since the interpolation method will be limited to data within a relatively small neighborhood and the various plausible variogram shapes can all be well approximated by a power-law model), 3. the pairwise correlation between differences used in the estimation (because data are obtained at difference time instances and these data are less or not at all correlated). Issues that on the other hand are considered important to a proper estimation of the variogram are: 1. the non-gaussian distribution of the squared differences between two measurements, 2. the sparsity and non-regularity of the data, which makes the more traditional estimation method based on grouping of the data into distance bins difficult to apply.

To address these problems, an estimation method is developed where, instead of grouping the data, the individual squared differences are fitted by a power-law model using a generalized least-squares method. The latter method allows to account explicitly for the non-gaussian distribution of the data. A first version of this estimation method has been developed and leads to satisfactory results. Further work is however needed to appropriately choose the region over which the model can be considered valid. This work will be completed in the next year.

Appendix 1 shows a copy of the transparencies that have been handout during the workshop organized by UCS in February, 1999, on the use of the Kriging method for interpolation of the data and the quality control.

3. The Statistical Analysis Tool

During this year an operational prototype version of a statistical analysis tool has been developed and implemented in the form of the SAT program. Paragraph 3.1 summarizes the role of the SAT program with respect to the overall Quality Assurance approach and with respect to the analysis of the data in the IDOD database. Paragraph 3.2 briefly explains how the SAT program is developed. Functionalities that have been or will be implemented in the future are described in Paragraph 3.3.

3.1 Role of the SAT program for the overall QA-approach and data analysis

The overall dataflow and the function of the SAT program in this dataflow are illustrated in Figure 1. The raw data delivered by the data providers is directly put into the IDOD database. At this stage, a first level of quality assurance is executed which consists in checking whether all data acceptance requirements are fulfilled (i.e. proper documentation of the instruments, completeness of the data, etcetera).

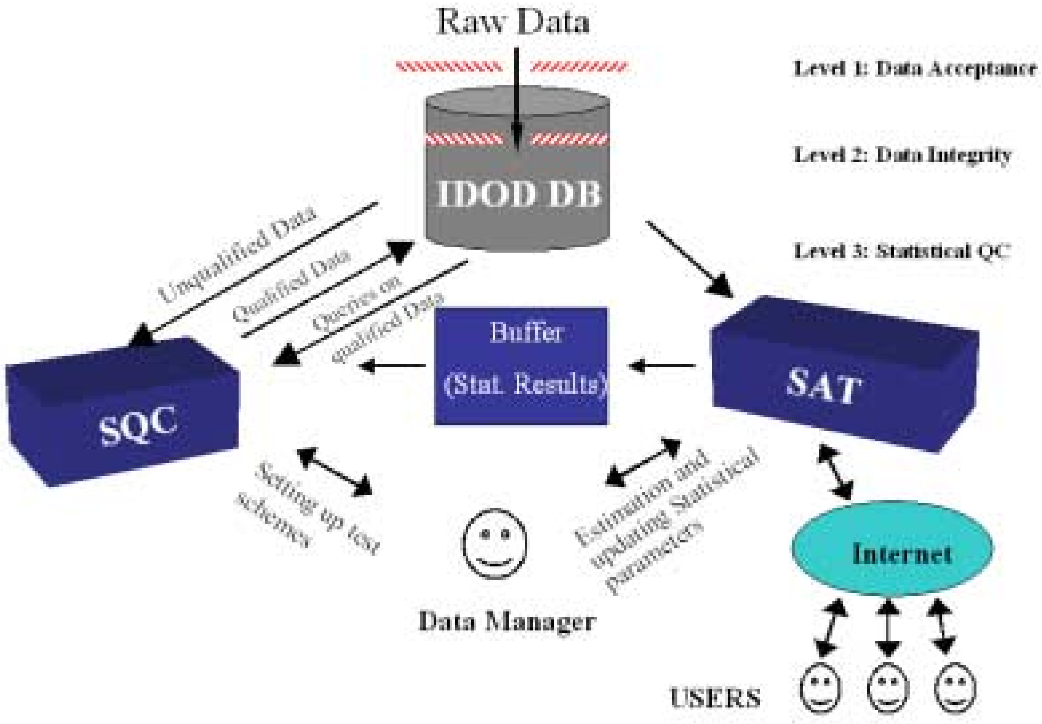


Figure 1: Data flow and position of SAT in the 3 level QA-approach

Immediately upon entrance into the database a second level of quality assurance is executed which consists of checking the data integrity of the incoming data (i.e. do reference locations exist, do references to measurement methods exist, are the data of a valid numeric form, etcetera).

Once the data are accepted into the database, they are considered valid but "unqualified" data, meaning that their statistical consistency with other data or with expected distributions has not been verified. This is the task executed by the SQC program which is activated by the data manager. The SQC program will for instance check the value of the different measurements against upper- and lower-bound values that are temporal and spatially dependent, it will check the values against similar measurements made at a distance and it will check the values against measurements of another type made at the same location and time. The relationships used in this checking are statistically defined and can be determined by use of the SAT program, which therefore includes specific algorithms for the estimation of the statistical relationships used in the SQC program. A primary and privileged user of the SAT program is therefore the data manager who can use the SAT program to update statistical relationships as more data become available and store them in a buffer, where they become accessible to the SQC program. Storage of the results is enforced, partly to shorten the CPU time of execution of the quality control procedure and partly to enforce the visual examination and documentation of such results and prevent the fully automated estimation of relationships.

Apart from this specific role in the quality assurance approach, the SAT program can be also directly used by the user to view and analyze the data in the IDOD database. For this purpose, not only the specific statistical tools but also more general statistical tools are included that are tailored to the IDOD database and to the problems associated with oceanographic data. Within this context, special attention is paid to allow for a user-friendly interface that requires only an intermediate level of statistical knowledge from the user. A specific requirement of the SAT program is that it can be used on line on the internet by multiple users.

3.2 The implementation of the SAT program

The SAT program is based upon S-Plus. This general purpose statistical analysis tool has been tailored by UCS to perform statistical analyses for the IDOD program and more general for oceanographic data as described in this paragraph. The connections between the user, the SAT program and S-Plus are visualized in the following figure.

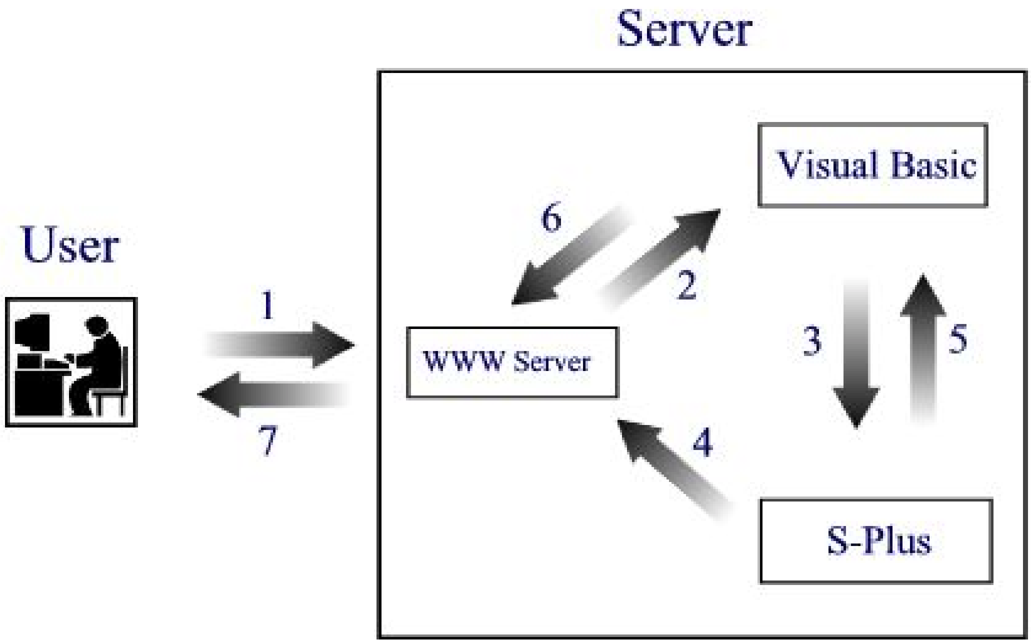


Figure 2: Implementation of the SAT program

A user, or the data manager, uses an internet browser to access the main site of the SAT program.

To make a statistical analysis the user sends a request to the www-server to execute a chosen analysis (step 1). The analysis is started by means of CGI (Common Gateway Interface, a server-side scripting language which makes it possible to activate programs on the server (step 2). This CGI command is written in the development language Visual Basic. In the next step (3) Visual Basic programs are used to identify the user and to make the connection to S-Plus. The statistical analysis is performed in S-Plus and the results are put in a web page on the server (step 4). Next S-Plus sends a command (step 5) to the Visual Basic program that the analysis is finished and the results are available. The Visual Basic program gives an indication to the www server (step 6) that the user can be redirected to the result page of his statistical analysis. Finally (step 7) the results are send to the user.

The same procedure is used by the data manager to derive models for use in the SQC program. After examination of the results the manager can use a separate off-line program (written in Visual Basic) to send the data to the buffer with statistical results. Once the data are stored there, it can be used in the SQC program. The different steps to be followed by the Data manager are summarized in the following figure.

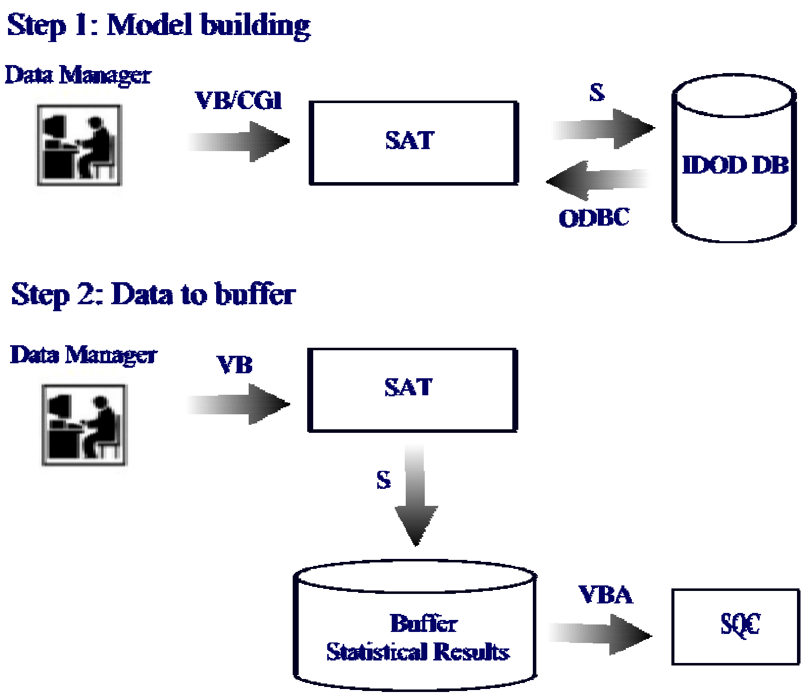
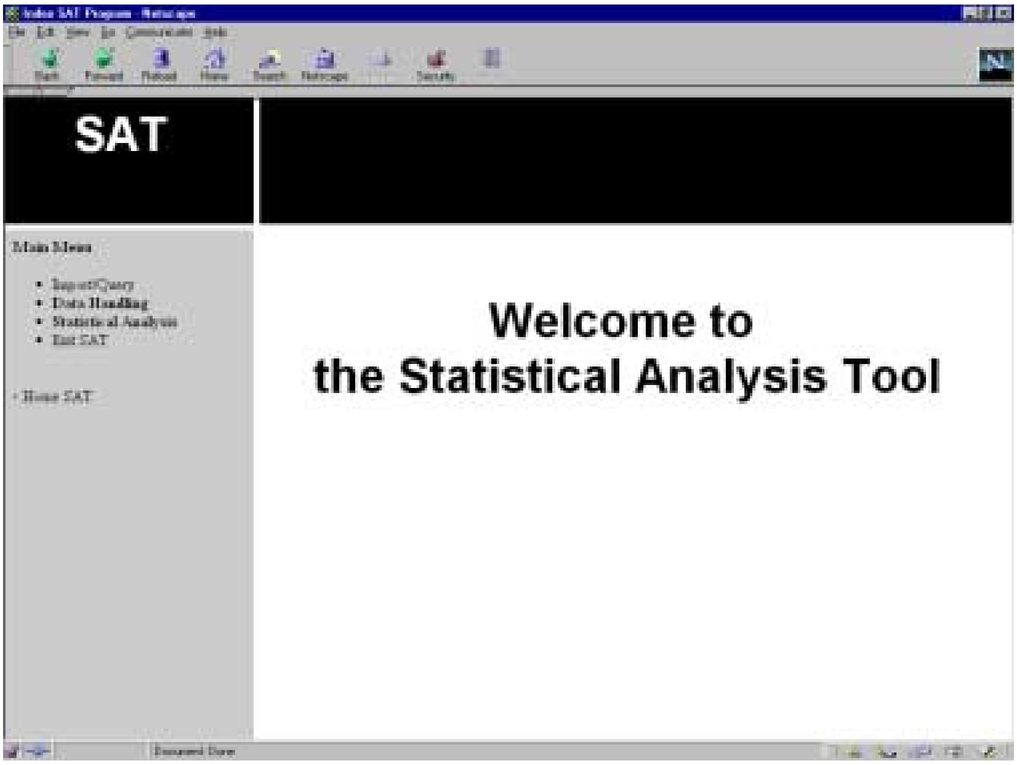


Figure 3: Methodology for the data manager to derive models

3.3 Functionalities of the SAT program

In this paragraph the different functionalities implemented in the SAT program are briefly described. Screenshots showing details of the different options are reproduced in Appendix 2.

The main menu of the Statistical Analysis Tool is shown in the following picture.


Index SAT Program - Netscape
File Edit View Go Communicator Help
Back
Forward
Reload
Home
Search
Netscape
Security
Document: Done

SAT

Main Menu

- Import/Query
- Data Handling
- Statistical Analysis
- Exit SAT

- Home SAT

Welcome to
the Statistical Analysis Tool

Figure 4: Main Menu of SAT

The 3 main functionalities of the program are indicated in the menu.

- Import/Query

  This option launches a menu (see appendix 2A) with an option to select data from the IDOD database. Implementation of this option is done by MUMM/Ulg and involves a query of the IDOD database. The second option “send data to SAT” is used to retrieve the selected data and make them available to SAT for statistical analyses.

- Data Handling

  This option gives a menu (see appendix 2B) with following general functions: 1. variables can be transformed (appendix 2C), 2. the variable type can be changed and 3. the data can be shown on screen or downloaded as a Microsoft Excel Workbook (appendix 2D) .

- Statistical Analysis

This option gives a menu with the available statistical analyses in the SAT program.

The following functions are currently operational:

- Summary Statistics.

  This function is split into Summary Statistics and Summary Plots (see appendices 2E and 2F for details). With these functions the most commonly used summary statistics can be calculated and various summary plots can be shown.

- Trend fitting

  The trend analysis is used to look for the relationship between a variable of interest (the response variable, i.e. salinity) and another variable (the repressor variable, i.e. distance from coast). In SAT a linear, quadratic and nonparametric function can be fitted to the data (see appendix 2G).

- Correlation Analysis

  The function is split into Correlation/Covariance Matrix and Scatterplotmatrix (see appendix 2H and 2I for details). The correlation/covariance matrix function calculates the correlation between the different variables in a dataset. The scatterplotmatrix visualizes the correlation between the different variables of a data set.

Following additional functionalities are under construction and are expected to be operational in the first half of the next year:

- Regression Analysis (under construction)
- Spatial Analysis (under construction)
- Time Series Analysis
- PCA/Factor Analysis
- Distribution Fitting
- Model Validation

4. Integration of the programs in the IDOD project

A substantial part of the work this year has been devoted to the integration of the different software components. As can be seen in Figure 1 there are following links between the programs

- IDOD db - SQC (to get the data from the IDOD database to make a quality control and to send the qualified data back);
- IDOD db - SAT (to retrieve data from the database into SAT);
- SAT - buffer (to store SAT model results in a buffer for later use in SQC);
- SQC - buffer (to retrieve model results of SAT in SQC).

The last link (SQC to buffer) was already established in the previous year in a prototype version, but during the development of the SAT program, it became clear that various modifications were necessary to the design of the buffer. As a consequence also this linking had to be (re)programmed.

At this stage programming of the links is nearly complete. Testing of the software will be performed in the next year.

6. Summary

During this year, effort has been devoted to 1. the development of an estimation method for the variogram that is to be used in the statistical quality control, 2. to the development of a statistical analysis tool to be used by general users to analyze the data in the IDOD database and by the data manager to derive results for the statistical quality control, and 3. to the implementation and linking of all software components.

A first version of the estimation method for the variogram is operational but some further modifications are necessary to arrive at a robust and reliable estimation method. This work will be completed in the first half of next year.

The statistical analysis tool for the IDOD database and more general for oceanographic data is under development. A first incomplete version is nearly operational. In particular, problems related to the accessibility of the tool on the internet have been solved, but their solution has required a substantial effort. With the basic framework available, it is anticipated that the development and implementation of other functionalities will proceed at a faster pace.

Furthermore, the linking between the different programs has been programmed. This task required re-engineering of the existing programs, in particular to make the linking between SAT en SQC via a buffer operational. This link has yet to be tested and possibly further modifications will need to be made to support the storage and exchange of all model results.

In the upcoming year the various programs will be further implemented, extended, tested and, when there exists specific needs for the data providers, adjusted.

References

UCS Workshop, "Spatial Statistics", February, 1 1999.

UCS part Website on “Symposium Développement Durable - Duurzame Ontwikkeling :

"A la recherche d'un dialogue durable entre science et politique", "Op zoek naar een duurzame dialoog tussen onderzoek en beleid" , Brussels - Palais des Congrès/Paleis voor Congressen - 24-25 November 1999.

Cressie Noel, 1991, Statistics for Spatial Data, John Wiley & Sons, New York

UCS

Universitair Centrum voor Statistiek

Integrated and Dynamical Oceanographic Data Management - IDOD

Scientific Report: Appendices

January 2000

Contribution of KUL-UCS

University Center of Statistics
Katholieke Universiteit Leuven
de Croylaan 52b
B-3000 Leuven
Belgium

KATHOLIEKE UNIVERSITEIT
LEUVEN

Appendix 1: Handouts Spatial Workshop

IDOD PROJECT

Workshop on Spatial Statistics
1 February, 1999
UCS

USE OF THE KRIGING METHOD

FOR INTERPOLATION OF DATA

AND QUALITY CONTROL

PROBLEM STATEMENT

Verify the consistency of a new measurement value by comparing its value against the value that is expected on the basis of neighboring measurements (if any are available).

1. Obtain an interpolator.

2. Obtain an estimate of the standard deviation of the difference between the interpolator and the actual value.

INTERPOLATION TECHNIQUES

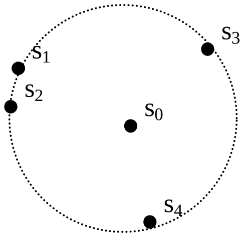
DETERMINISTIC/HEURISTIC

1. Distance-weighted method (DW-method)

$$Y_0 = \sum_{i=1}^{n} w_i Y_i \text{ where } \sum_{i=1}^{n} w_i = 1 \text{ and } w_i \propto \frac{1}{d_{i0}^{\theta}}$$

3. Splines

Problems: 1. no estimate of standard deviation of difference between interpolator and actual value (could be obtained through jacknifing or crossvalidation); 2. in DW-method, datapoints that are spatially close receive too much weight



STATISTICAL METHODS

2. Stochastic interpolator (classic random field theory)

3. Kriging interpolator (incremental random processes)

Typical of these methods: 1. assumptions are made regarding the statistical distribution of the variable to be interpolated; 2. given those assumptions, the best interpolation formula (under some constraints) is derived and the distribution of the difference between the interpolator and the actual value is derived.

ASSUMPTIONS AND INTERPOLATORS

1. CLASSIC INTERPOLATOR (SECOND-ORDER STATIONARY RANDOM FIELDS)

(local) stationarity of the mean value: $E[Y_0] = \mu$

(local) stationarity of the covariance function: $E[(Y_i - \mu)(Y_j - \mu)] = C(d_{ij})$

best (minimum-variance) linear unbiased estimator

$$\hat{Y}_0 = \mu + \left[C(d_{01}) \quad \ldots \quad C(d_{0n})\right] \begin{bmatrix} C(d_{11}) & \ldots & C(d_{1n}) \\ \ldots & \ldots & \ldots \\ C(d_{n1}) & \ldots & C(d_{nn}) \end{bmatrix}^{-1} \begin{bmatrix} Y_1 - \mu \\ \ldots \\ Y_n - \mu \end{bmatrix}$$

$$\hat{Y}_0 = \mu + \underline{A}\underline{B}^{-1}\underline{Y}$$

$$\sigma^2_{\hat{Y}_0} = C(d_{00}) - \underline{A}\underline{B}^{-1}\underline{A}^T$$

KRIGING INTERPOLATOR (INTRINSIC STATIONARY RANDOM FIELDS)

(local) zero expected increment: $E[Y_0 - Y_i] = 0$

(local) stationarity of variance of increments: $E[(Y_i - Y_j)^2] = 2\gamma(d_{ij})$

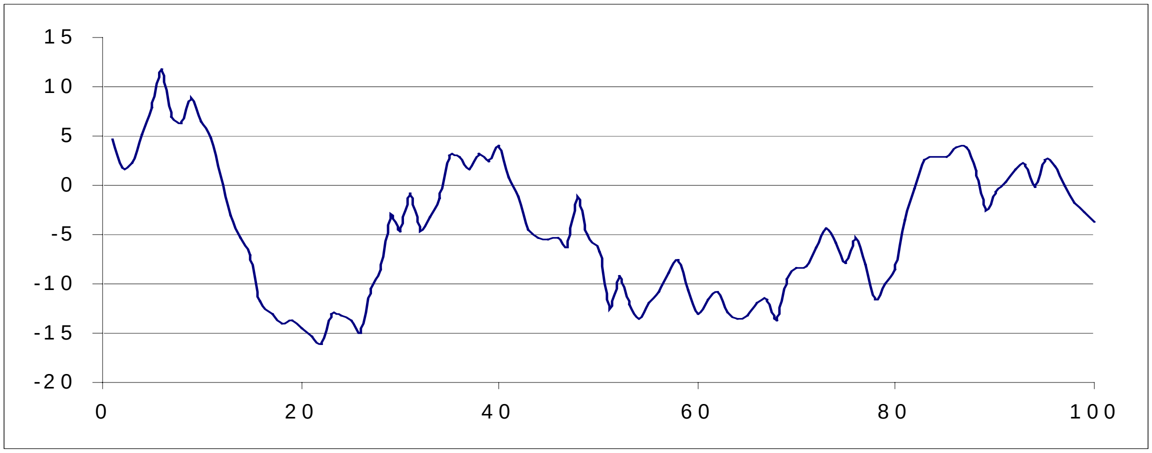
best (minimum-variance) linear unbiased estimator (Ordinary Kriging)

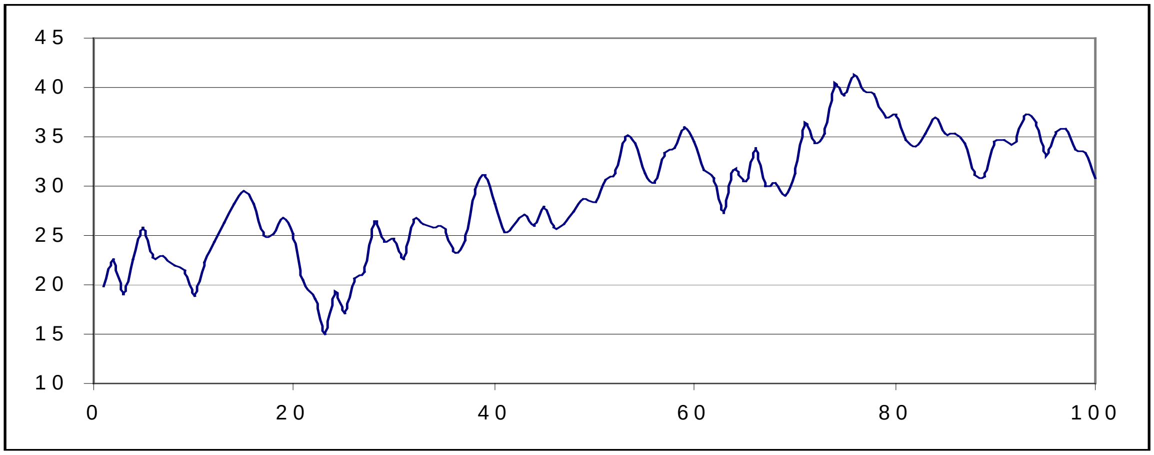
$$\hat{Y}_0 = \sum_{i=1}^{n} w_i Y_i \text{ where } \sum_{i=1}^{n} w_i = 1 \text{ and}$$

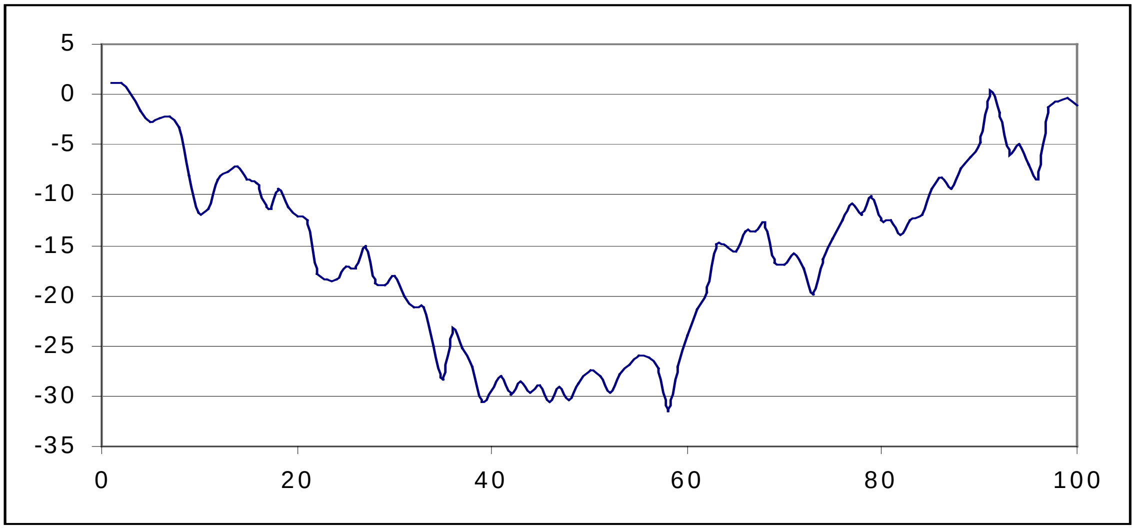
$$\begin{bmatrix} \gamma(d_{11}) & \ldots & \gamma(d_{1n}) \\ \ldots & \ldots & \ldots \\ \gamma(d_{n1}) & \ldots & \gamma(d_{nn}) \end{bmatrix} \begin{bmatrix} w_1 \\ \ldots \\ w_n \end{bmatrix} = \begin{bmatrix} \gamma(d_{01}) - \kappa \\ \ldots \\ \gamma(d_{0n}) - \kappa \end{bmatrix} \text{ or } \underline{C}\underline{w} = \underline{D} - \kappa \underline{I}$$

$$\sigma^2_{\hat{Y}_0} = 2\underline{w}^T\underline{D} - \underline{w}^T\underline{C}\underline{w}$$

SIMULATIONS OF INCREMENTAL PROCESSES







VARIOGRAM

Variance of the increment of the physical variable going from A to B as a function of the distance between the two points A and B.

Useful things to know:

4. If the random field is second-order stationary, then

$$\gamma(d) = C(0) - C(d)$$

4. Permissibility: $\gamma(d)$ cannot be any shape but must satisfy certain requirements (that depend on the dimension of the location coordinates on which d is based). Specifically, $\gamma(d)$ must be a conditionally semi-negative function, meaning that for any arbitrary choice of locations and arbitrary coefficients $\underline{w}$

$$\underline{w}^T \underline{C}\,\underline{w} \leq 0 \text{ when } \sum_{i=1}^{n} w_i = 0$$

If this condition is not satisfied then the variance of the interpolator may become negative.

5. If the random field is second-order stationary, then the variogram will converge to a constant value (the sill) at large distances;

6. If the measurements include measurement error in addition to spatial variation, then the variogram is not zero at distance zero (if non-zero, the value at d=0 is called the nugget effect).

CLASSIC (EMPIRICAL) ESTIMATOR OF THE VARIOGRAM

$$2\hat{\gamma}(d) = \frac{1}{N(d)} \sum_{i=1}^{N(d)} (Y_i - Y_j)^2$$

N(d) is the number of pairs of measurements at distance d.

Complications

5. Grouping of distances is necessary to obtain a reasonable number of pairs unless the data are sampled on a regular grid;

6. How does one fit a function γ(d) to the previous empirical estimates obtained for different distance-bins with midpoints d_1, d_2, ... d_K. The empirical estimates within a given bin-width are clearly non-Gaussian distributed when N(d) is small.

Other issues to be considered

7. $\hat{\gamma}(d)$ itself will not be necessarily permissible. Fitting of a permissible function γ(d) is necessary.

8. The empirical estimate is highly sensitive to outliers;

9. What happens if Y does show a systematic trend?

10. The pairs used in the empirical estimate are dependent: a. because the same measurement variable (i.e. Y_1) is used several times (i.e. in Y_1-Y_2, in Y_1-Y_3, ...); b. because also different increments are correlated (i.e. Y_1-Y_2 is dependent on Y_3-Y_4 if the data are mutually close). How should one deal with this dependence?

11. What is the uncertainty of the estimated variogram? This issue is particularly important, if one wishes to compare different models of different complexity (i.e. different permissible functions, different definitions of distance, fits over different neighborhood ranges, ...)

Solutions in the literature (2-step procedure)

12. Robustify empirical estimate and make it more Gaussian distributed;

13. Use (generalized) least-squares to fit a function γ(d) (eventually accounting for correlation between the pairs in an approximate manner)

Rarely addressed: 1. choice of bins (irregular sampling), 2. comparison of alternative models

PERMISSIBILITY ISSUE

2. Nonparametric (Shapiro and Botha, ...)

Permissibility constraint is formulated in terms of the frequency decomposition of the variogram and constrained least-squares can be used to obtain a fit. “Non-parametric” because for each frequency component the contribution is estimated.
Disadvantage: complicated, “strange” shapes

Parametric (others)

Use one of the many permissible models: i.e. exponential model, spherical model, Gaussian model, power-law model, ...

Comment:
For the purpose of interpolation, only the shape of the variogram at “close” distances (the typical distance to the N-nearest neighbors) is of practical importance. The power-law model can approximate reasonably well the other models near the origin:

$$\gamma(d) = \alpha + \beta\, d^{\theta} \quad \text{where } \alpha\geq 0,\ \beta\geq 0 \text{ and } 0\leq\theta\leq 2 \quad \text{for } d<D$$

where D is the distance over which a reasonable fit is obtained.

Using the fact that the sum of two permissible variograms is also permissible one might eventually consider an extension of the previous model by using the expontial variogram function as an alternative model or as an additive term:

$$\gamma(d) = \alpha + \beta\left(1 - \exp\left(-\frac{d}{\theta}\right)\right) \quad \text{where } \alpha\geq 0,\ \beta\geq 0,\ \theta\geq 0$$

OUTLIER ISSUE

Cressie and Hawkins

Empirical estimate based on the sum of the square roots (instead of the squares). Idea is to obtain a more normally distributed variable and to decrease the influence of large increments.

Genton

Empirical estimate based on a quantile value of the absolute increments (the order corresponds to the number of pairs that could be generated if only half the sample size would have been observed).

Comment

1. The issue of a non-normal distribution and the issue of the existence of outliers could be treated separately;
2. Outliers should be less frequent in the IDOD database than in typical geostatistics applications.

EXISTENCE OF A SYSTEMATIC TREND

Universal Kriging

Generalization of Kriging interpolator, where the mean is assumed to exist and estimated from the data. Estimation of the variogram from the residuals leads to bias problems (the mean value has been estimated on the basis of the data and the data are correlated). Genton proposes a generalized least-quares method that corrects for this problem.

Comment

1. Notion of intrinsic stationarity is partly lost (mean is assumed to exist). A trend should be only removed if it is physically logical and statistically significant.
2. In the IDOD project, data are sampled over the same space in different independent campaigns. This is very different from the situation in geostatistics where a single sample is available and the question always remains if one is looking at a trend or a random drift. In the IDOD project, the trend issue could be treated more or less separately from the spatial correlation issue and the bias-issue should be less important.

CORRELATION BETWEEN THE DATA

Generalized least-quares

Both Cressie and Genton propose this method. Complicated!

Comment

1. Correlation will in general not bias the estimator if it is of an additive form (i.e. classic estimator). It is an issue with respect to the uncertainty of the variogram estimate and if the estimate is based on residuals (trend removal).
2. Because in the IDOD project data are available from different campaigns, it should be less of an issue.
3. One might further reduce its importance by considering only mutually exclusive pairs of points.

REMINDER ON OTHER ISSUES

1. The IDOD data are typically on a non-regular sampling grid. Choosing an appropriate bin-width is an important issue.
2. For some of the variograms the isotropic model may not apply. More general models can be constructed by allowing for so-called geometric anisotropy, where distance is defined as a positive-definite function of the spatial coordinates. However, then one should be able to compare the goodness-of-fit of alternative distance formulations.
3. In the IDOD project, the emphasis lies on interpolation and thus on accurate modelling in the neighborhood. Determining in an objective manner the range D over which the fit is sufficiently accurate is an important issue.

OUR PROPOSAL

1. Ordinary Kriging (not universal Kriging is used): trend estimation is handled separately (in the trend estimation one could eventually take spatial correlation of the data in account). If a trend has been identified (as statistically significant and physically plausible), then the variogram is estimated on the residuals neglecting uncertainty on the trend estimate.

2. Because of the irregularity of the grid, each pair of datapoints is treated separately. Thus, no grouping in distance bins is applied.

3. The non-Gaussian distribution of the squared increments is explicitly accounted for by using general linear models: the squared increment has a chi-square distribution with 1 dof if the underlying variable is Gaussian distributed. The original variable may be transformed to satisfy this assumption if necessary (separate problem).

4. Permissibility is enforced by fitting the power-law model which should be satisfactory for interpolation purposes (eventually we could consider an extended model allowing for the exponential model as an alternative or a combination of both models).

5. Correlation between the different increments is neglected in the estimation of the variogram.

6. We assume that the percentage of outliers (if any) is small. Outlier removal (if any) would be done by verification of residuals from the fit and comparison with the chi-square distribution (iterative procedure).

7. The choice between different alternative models (i.e. power-law with or without nugget-effect, power-law versus exponential, alternative distance definitions, alternative ranges that are fitted) should be made considering goodness-of-fit and uncertainty on the estimated variogram. Perhaps resampling (i.e. blocked bootstrapping where different campaigns represent blocks) could be used as a basis.

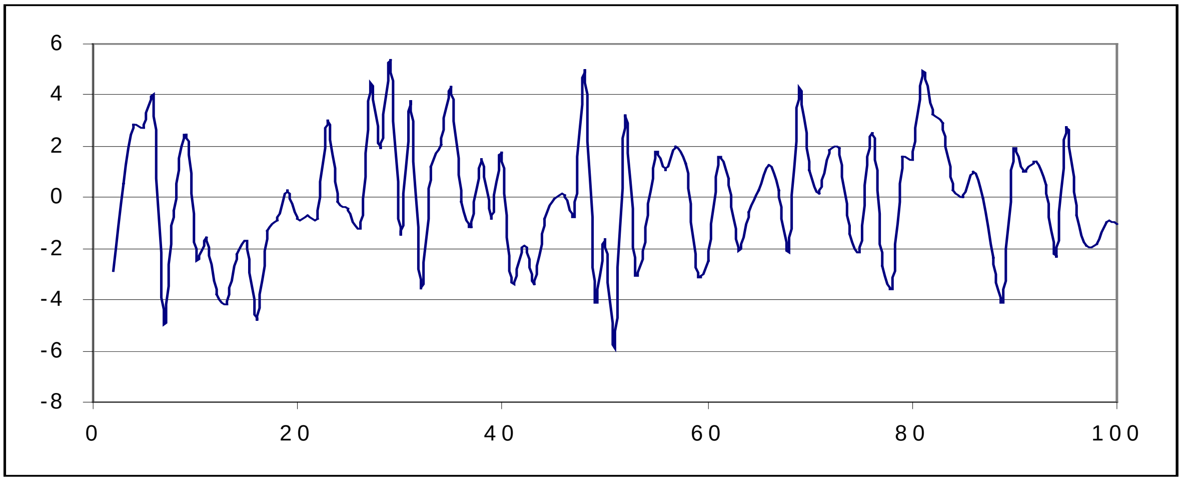
8. The theorethical Kriging variance is verified against the data and, if necessary, a correction is applied.

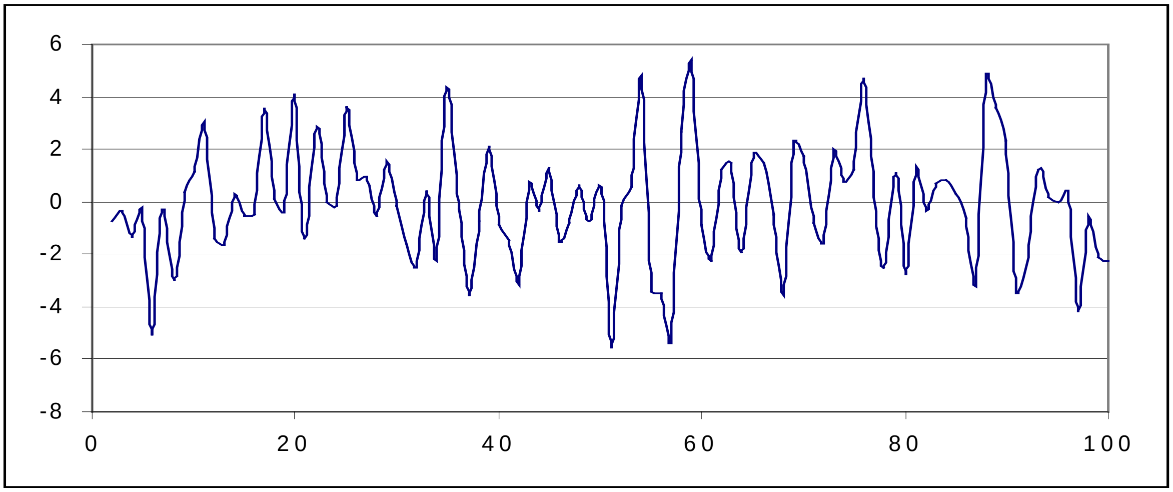
OUTSTANDING PROBLEMS:

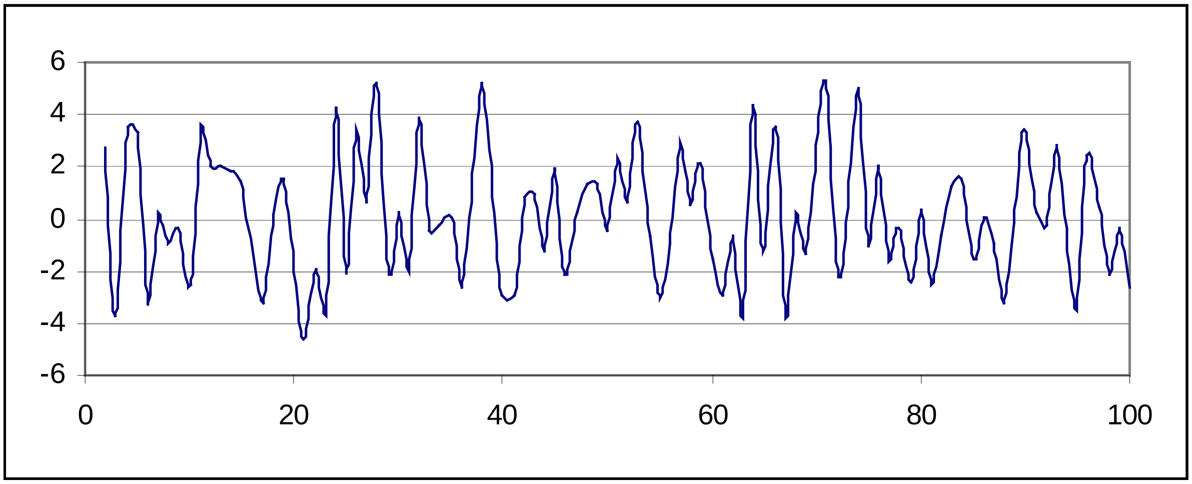
9. General linear models can accomodate the non-Gaussian distribution of the datapoints, but the models cannot always be linearized in the parameters.

10. How to compare different models and make an objective choice.

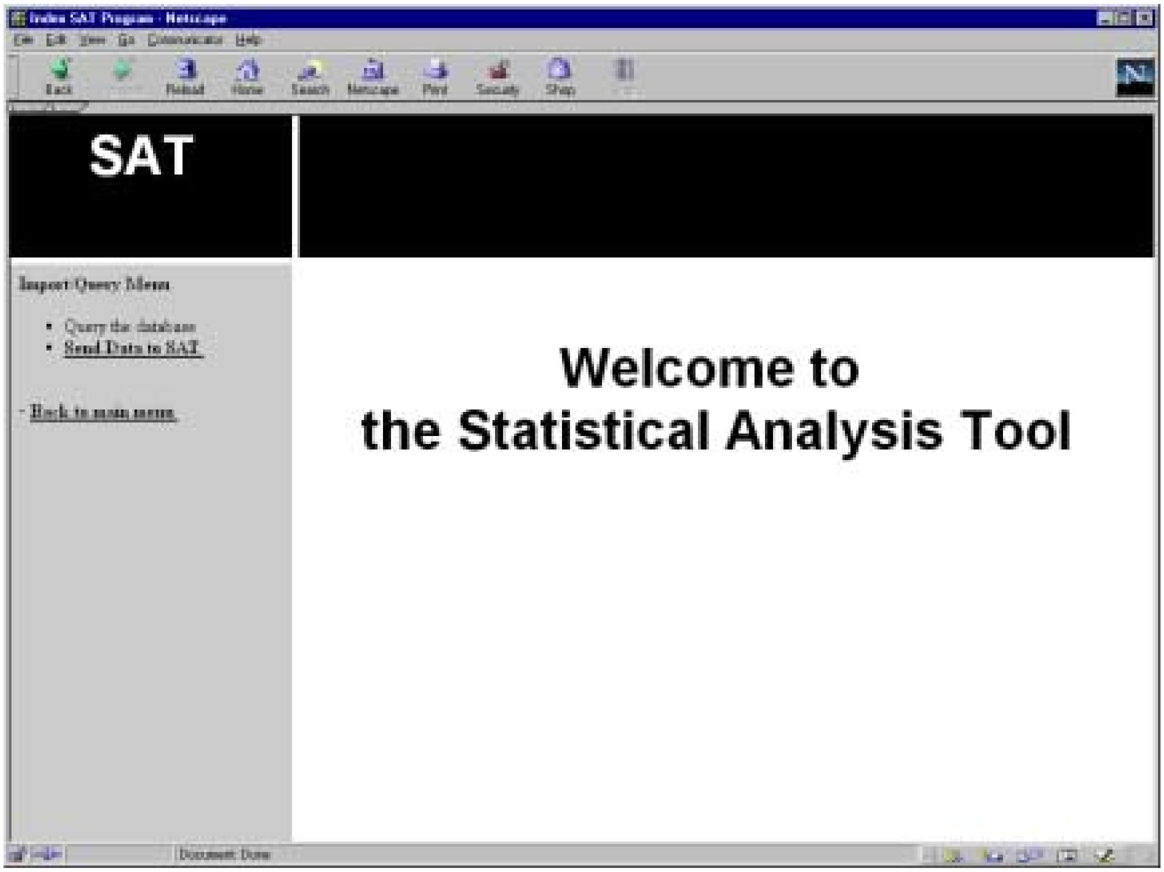
INCREMENT VALUES







Appendix 2A: Import/Query Menu


Index SAT Program - Netscape
File Edit View Go Communicator Help
Back
Forward
Reload
Home
Search
Netscape
Print
Security
Shop
Stop

SAT

Import/Query Menu

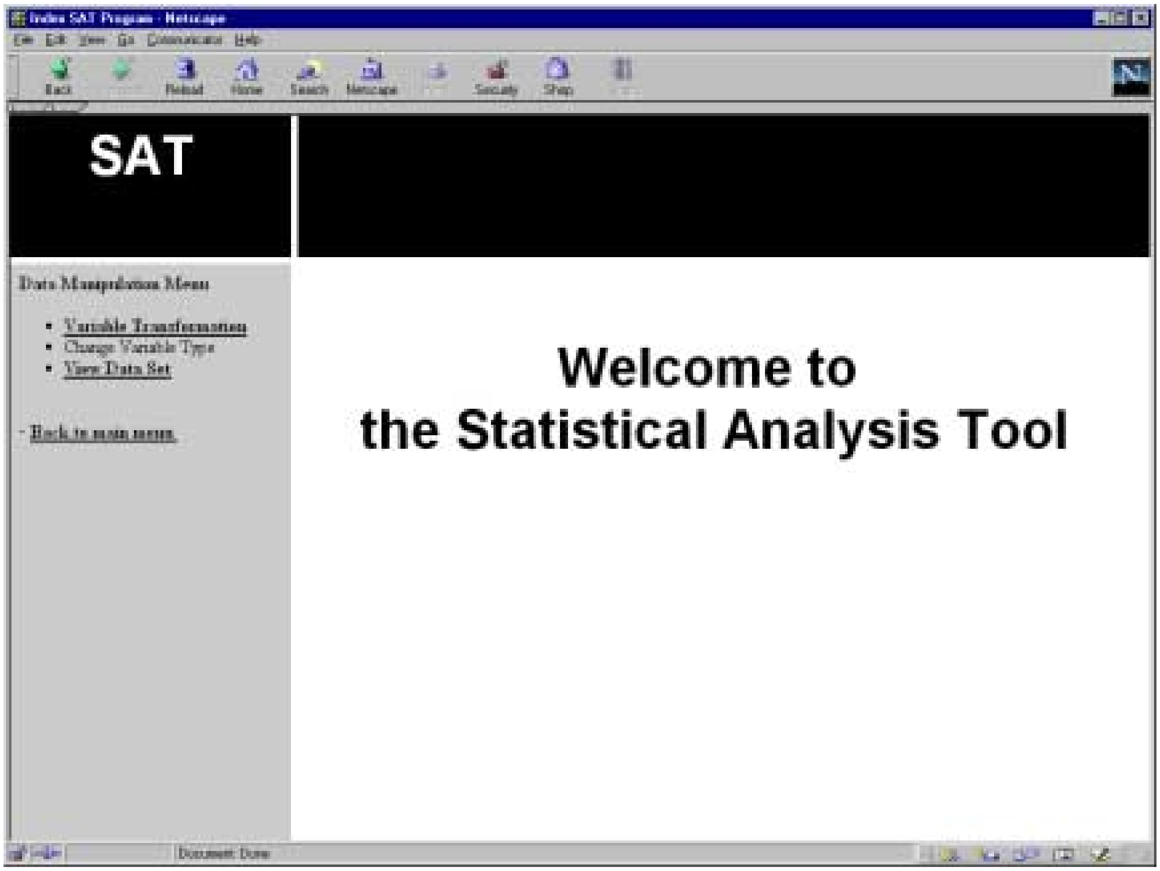
- Query the database
- Send Data to SAT

- Back to main menu

Welcome to
the Statistical Analysis Tool

Document: Done

Appendix 2B: Data Handling Menu


Index SAT Program - Netscape
File Edit View Go Communicator Help
Back
Reload
Home
Search
Netscape
Security
Shop
Document: Done

SAT

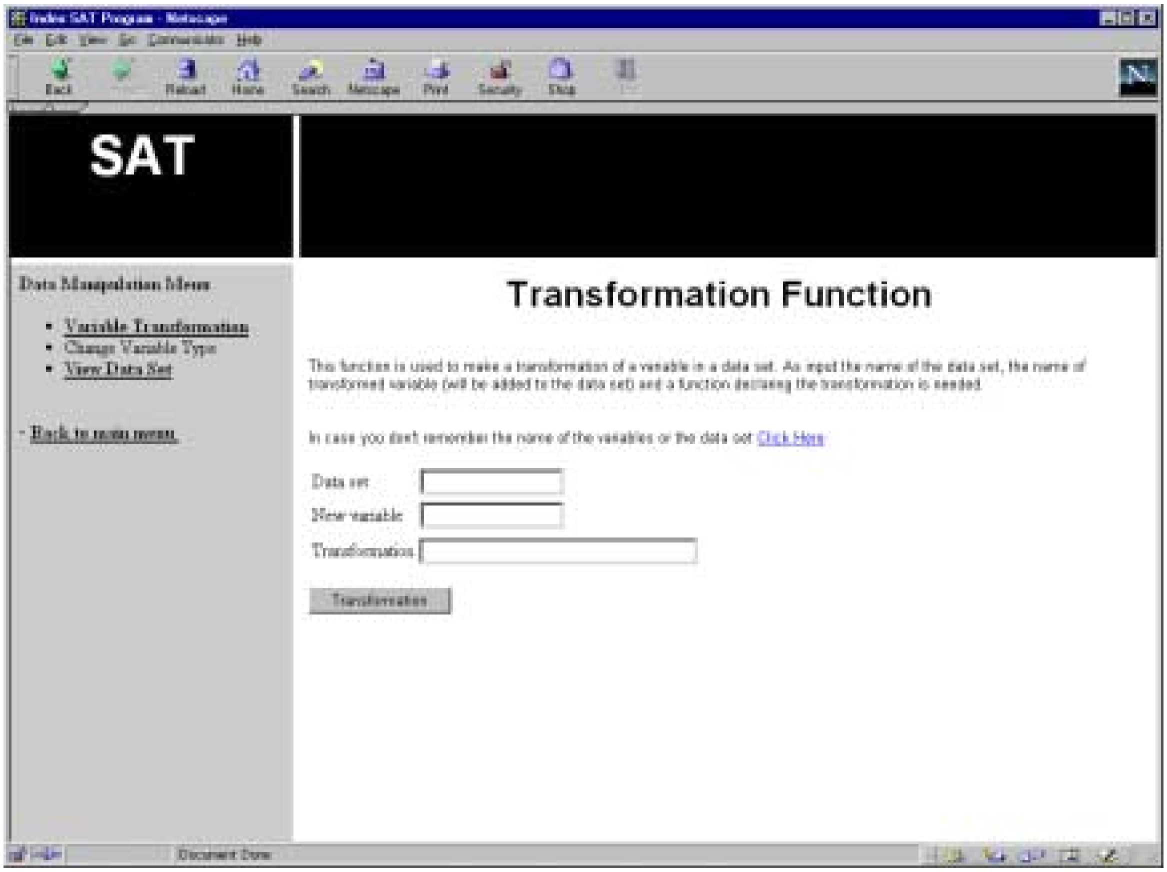
Data Manipulation Menu

- Variable Transformation
- Change Variable Type
- View Data Set

- Back to main menu

Welcome to
the Statistical Analysis Tool

Appendix 2C: Variable Transformation Function



Transformation Function

This function is used to make a transformation of a variable in a data set. As input the name of the data set, the name of transformed variable (will be added to the data set) and a function declaring the transformation is needed.

In case you don't remember the name of the variables or the data set Click Here

Data set

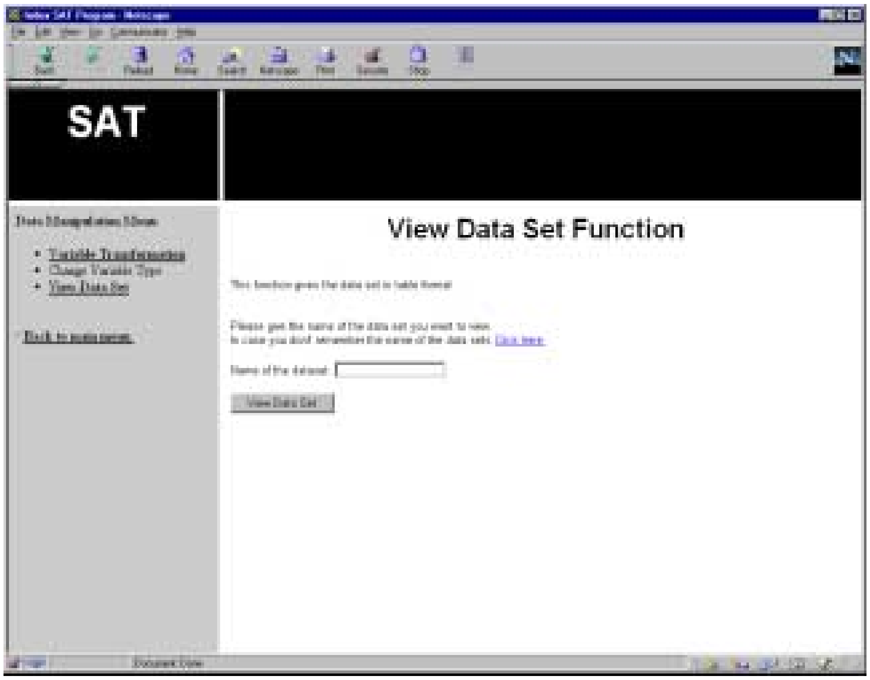
New variable

Transformation

Transformation

Appendix 2D: View Data Set Function

Screenshot of the function



Data Manipulation Menu

- Variable Transformation
- Change Variable Type
- View Data Set

Back to main menu.

View Data Set Function

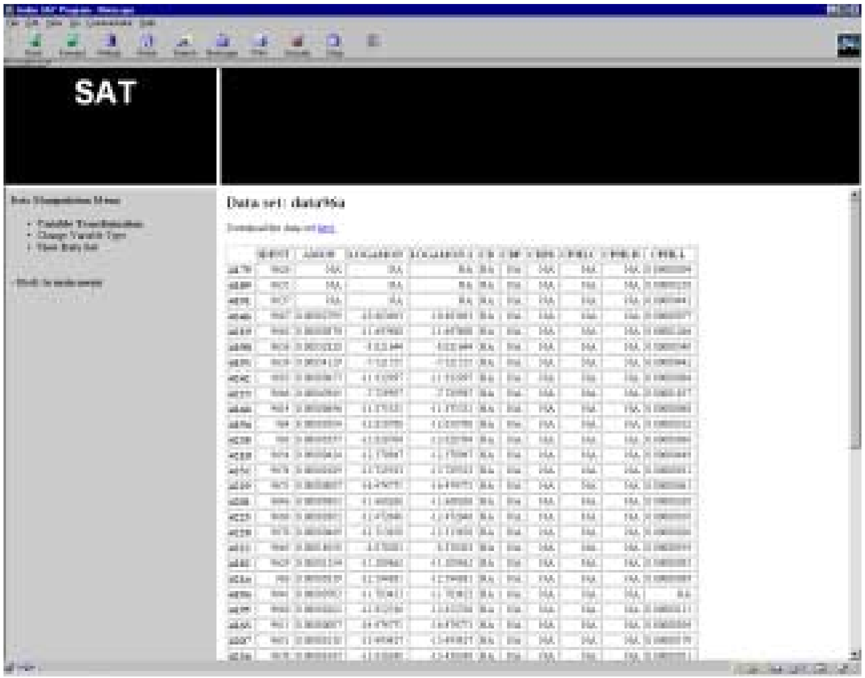
This function gives the data set in table format

Please give the name of the data set you want to view.
In case you don't remember the name of the data sets Click here.

Name of the dataset:

View Data Set

Example of a result

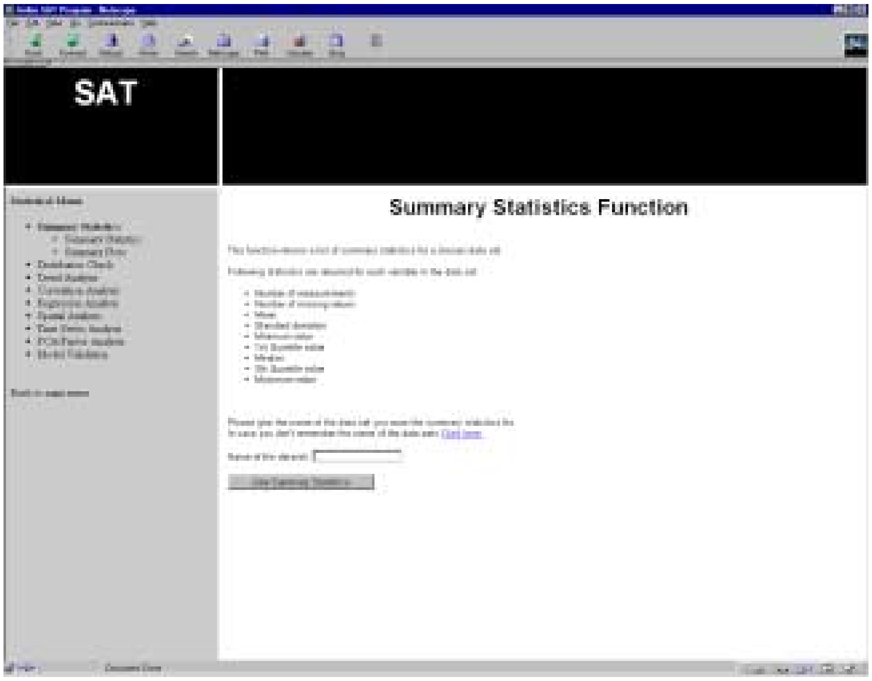


Data set: data96a

Download the data set here.

Appendix 2E: Summary Statistics Function

Screenshot of the function



Summary Statistics Function

This function returns a list of summary statistics for a chosen data set.

Following statistics are returned for each variable in the data set:

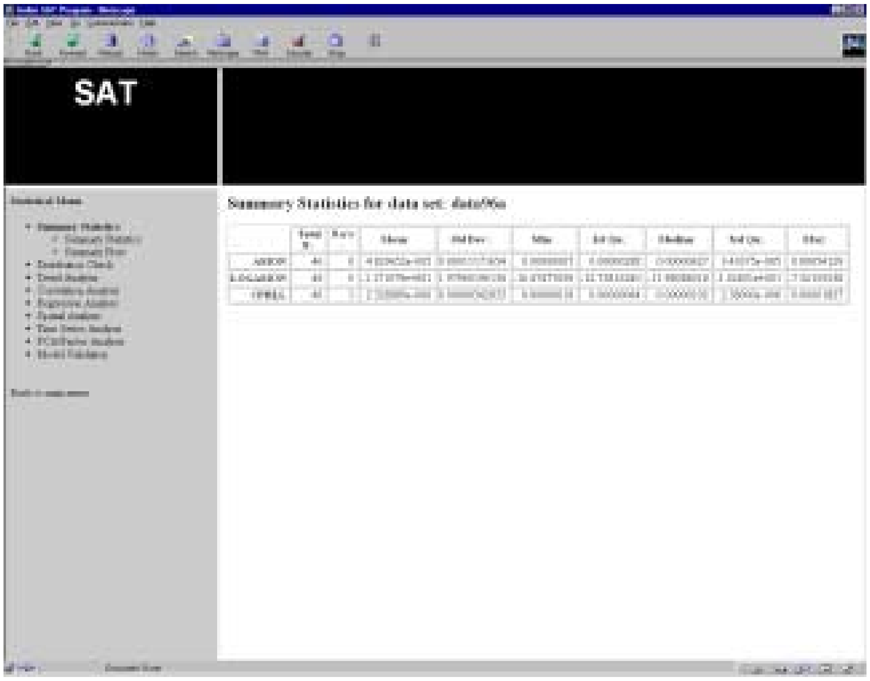
- Number of measurements
- Number of missing values
- Mean
- Standard deviation
- Minimum value
- 1st Quartile value
- Median
- 3th Quartile value
- Maximum value

Please give the name of the data set you want the summary statistics for.
In case you don't remember the name of the data sets Click here.

Name of the data set:

Give Summary Statistics

Example of a result

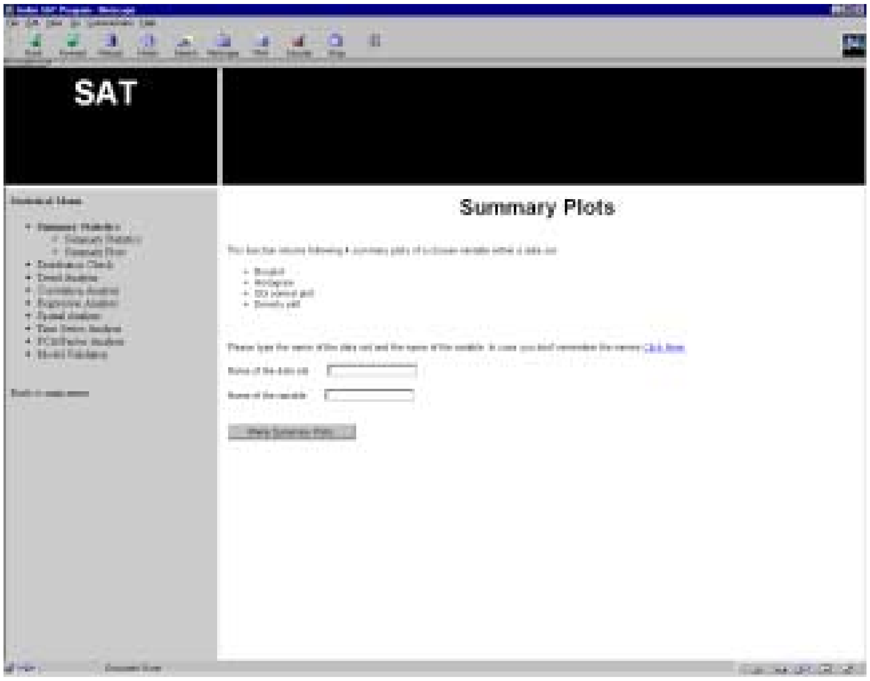


Summary Statistics for data set: data96a

	Total N	N miss	Mean	Std Dev	Min	1st Qu.	Median	3rd Qu.	Max
ARSEN	40	0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
LOGARSEN	40	0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
OPPEL	40	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Appendix 2F: Summary Plots Function

Screenshot of the function



Statistical Home

- Summary Statistics
  - Summary Statistics
  - Summary Plots
- Distribution Check
- Trend Analysis
- Correlation Analysis
- Regression Analysis
- Spatial Analysis
- Time Series Analysis
- PCA/Factor Analysis
- Model Validation

Back to main menu

Summary Plots

This function returns following 4 summary plots of a chosen variable within a data set:

- Boxplot
- Histogram
- QQ normal plot
- Density plot

Please type the name of the data set and the name of the variable. In case you don't remember the names: Click here

Name of the data set

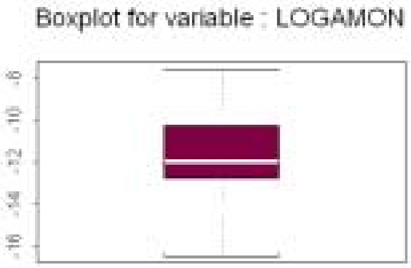
Name of the variable

Example of a result

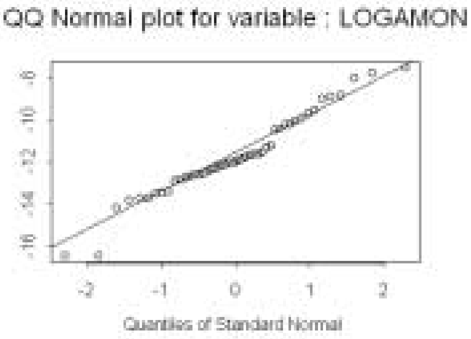
Summary Plots for Exploratory Data Analysis

Plots for dataset: data96a and variable LOGAMON
Number of observations : 46

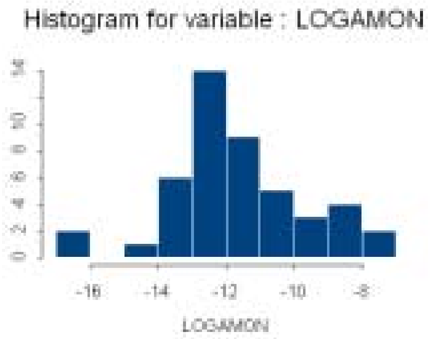
Boxplot



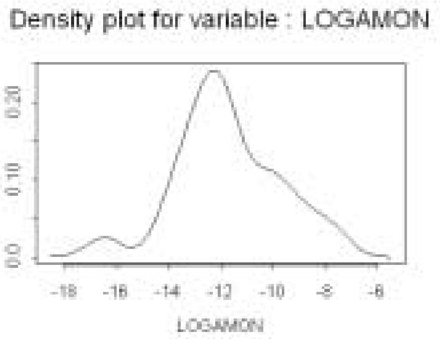
QQplot



Histogram

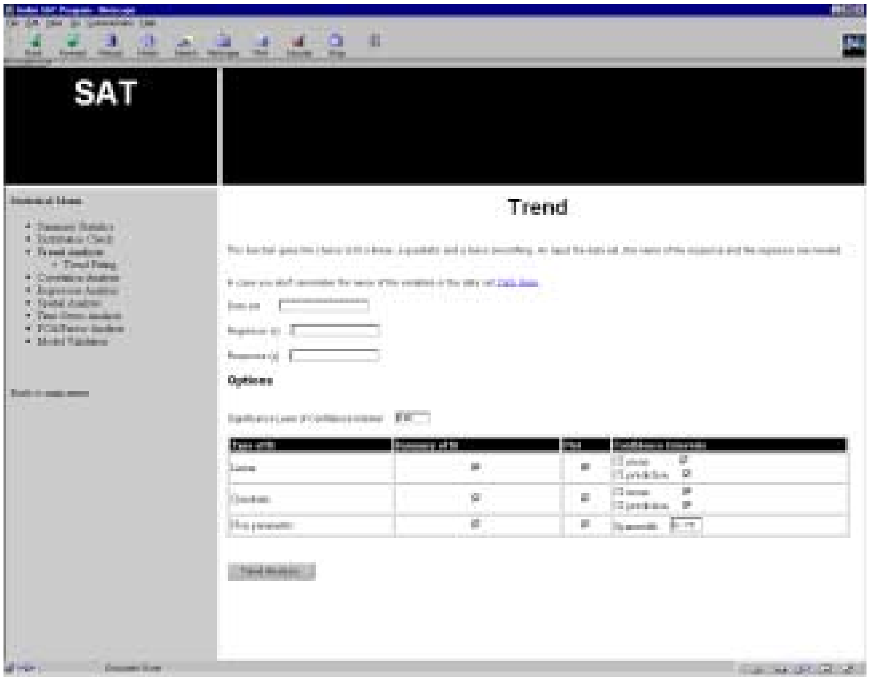


Density plot



Appendix 2G: Trend Fitting Function

Screenshot of the function



SAT

Statistical Home

- Summary Statistics
- Distribution Check
- Trend Analysis
    - Trend Fitting
- Correlation Analysis
- Regression Analysis
- Spatial Analysis
- Time Series Analysis
- PCA/Factor Analysis
- Model Validation

Back to main menu

Trend

[illegible]

[illegible] Help

Data set

Regressor (x)

Response (y)

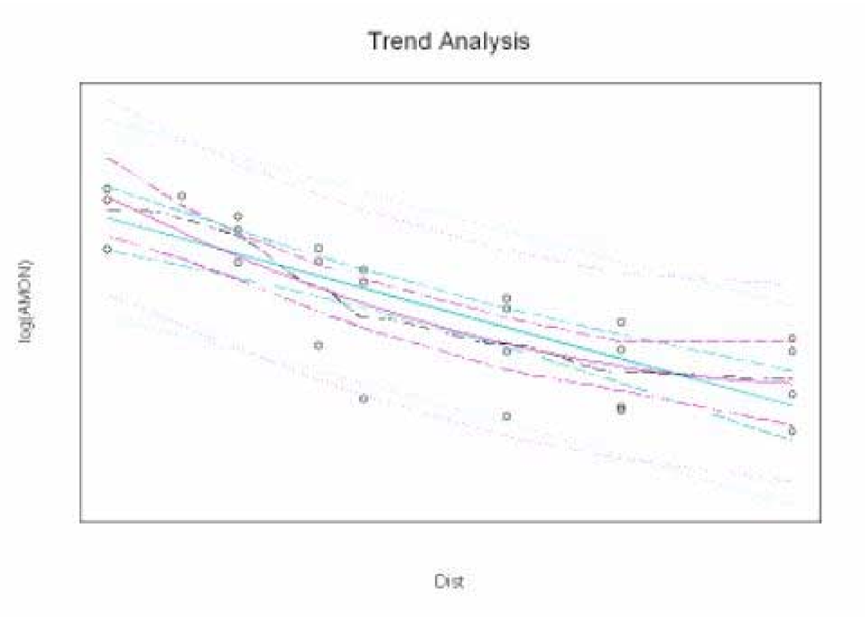
Options

Significance Level of Confidence Interval

Type of fit	Summary of fit	Plot	Confidence Intervals
Linear	☑	☑	☐ mean ☑ ☐ prediction ☑
Quadratic	☑	☑	☐ mean ☑ ☐ prediction ☑
Free parameter	☑	☑	Quantile

Trend Analysis

Example of a result



Linear Regression Results

```
Call: lm(formula = resp ~ regr, na.action = na.omit, x = T)
Residuals:
       Min          1Q      Median         3Q        Max
 -0.0001711 -0.00006701 -0.00002341 0.00008245 0.0002518

Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept) -0.0001  0.0000    -1.9888  0.0566
       regr  0.0000  0.0000    -5.9628  0.0000

Residual standard error: 0.0001091 on 28 degrees of freedom
Multiple R-Squared: 0.5594
F-statistic: 35.55 on 1 and 28 degrees of freedom, the p-value is 2.023e-006

Correlation of Coefficients:
     (Intercept)
regr 0.8657
```

Quadratic Regression Results

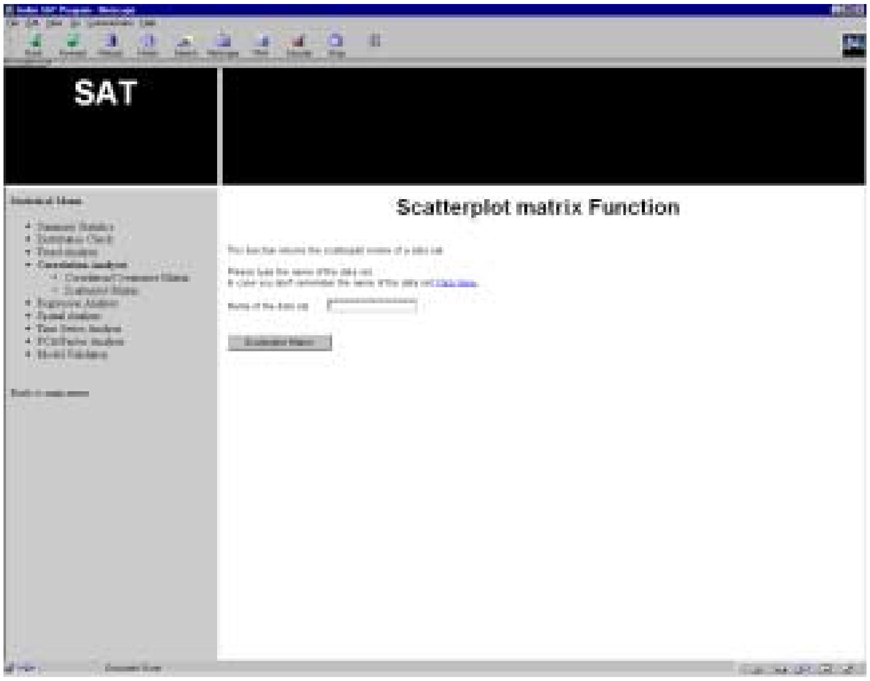
```
Call: lm(formula = resp ~ regr + regr^2, na.action = na.omit, x = T, y = T)
Residuals:
       Min          1Q     Median         3Q        Max
 -0.0002178 -0.00002917 0.00001086 0.00004008 0.0001706

Coefficients:
            Value Std. Error t value Pr(>|t|)
(Intercept) 0.0000 0.0000     0.9167  0.3674
       regr 0.0000 0.0000     2.5229  0.0178
  I(regr^2) 0.0000 0.0000     5.0287  0.0000

Residual standard error: 0.00007983 on 27 degrees of freedom
Multiple R-Squared: 0.7725
F-statistic: 45.84 on 2 and 27 degrees of freedom, the p-value is 2.085e-009

Correlation of Coefficients:
          (Intercept)   regr
     regr 0.7792
I(regr^2) 0.6104      0.9593
```

Loess Results

```
Call:
loess(formula = resp ~ regr, na.action = na.omit, span = lspan)

 Number of Observations:          30
 Equivalent Number of Parameters: 4.8
 Residual Standard Error:         0.00007749
 Multiple R-squared:              0.81
 Residuals:
       min       1st Q     median      3rd Q       max
-0.0002448 -0.00001031 7.478e-006 0.00002452 0.0001668
```

Appendix 2H: Correlation/Covariance Matrix Function

Screenshot of the function



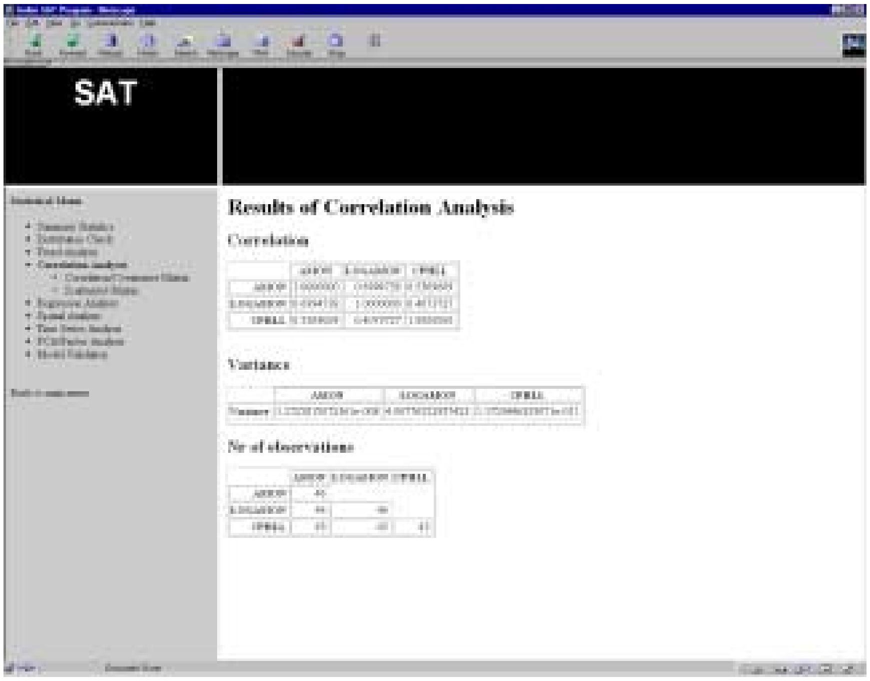
SAT

Scatterplot matrix Function

[illegible]

[illegible]

Example of a result



SAT

Results of Correlation Analysis

Correlation

	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]

Variance

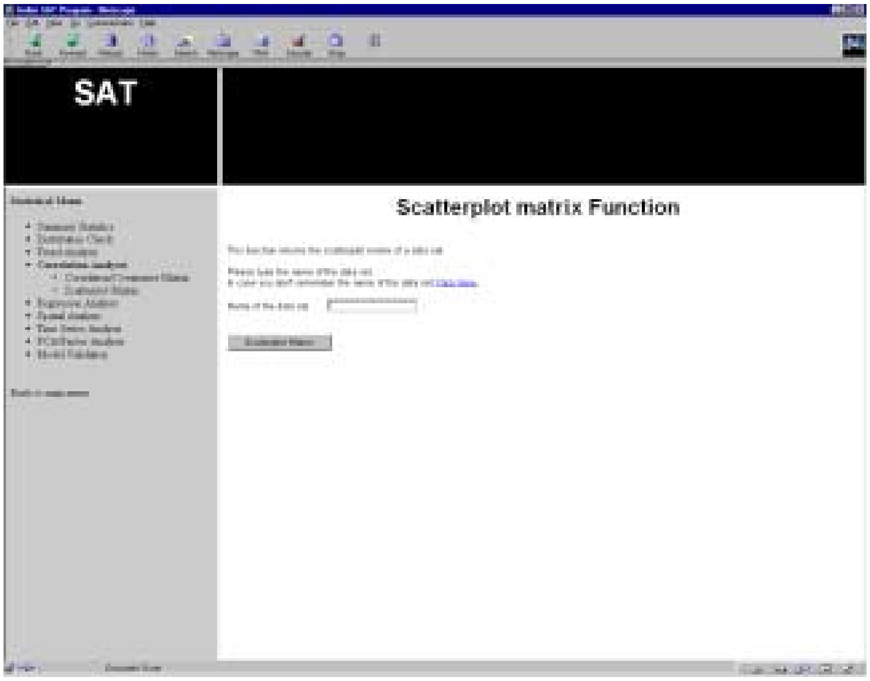
	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]

Nr of observations

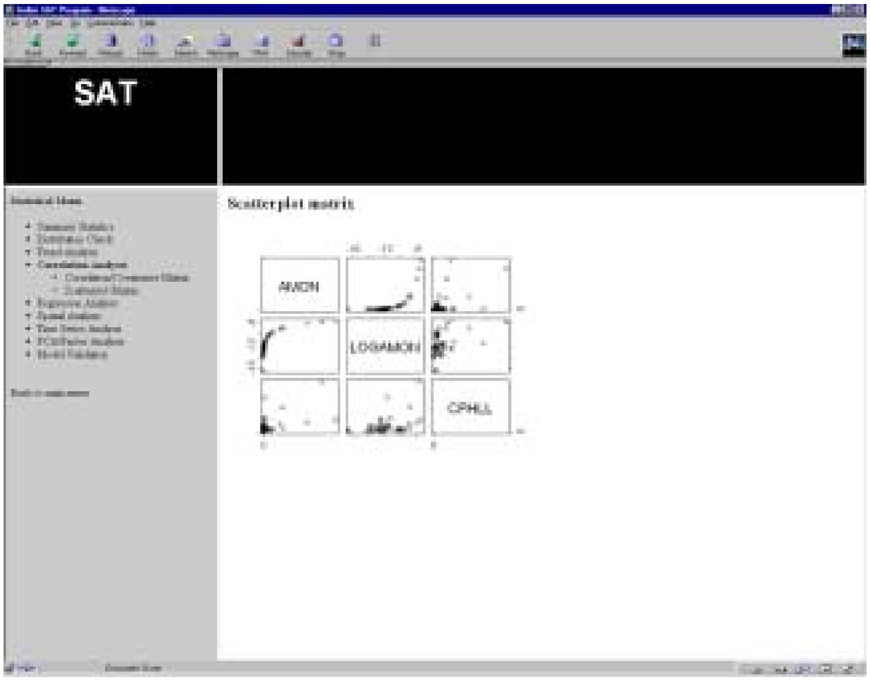
	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]		
[illegible]	[illegible]	[illegible]	
[illegible]	[illegible]	[illegible]	[illegible]

Appendix 2I: Scatterplotmatrix Function

Screenshot of the function



Example of a result



PRIME MINISTER'S SERVICES
Federal Office for Scientific, Technical and Cultural Affairs

SCIENTIFIC SUPPORT PLAN
FOR A SUSTAINABLE DEVELOPMENT POLICY

«SUSTAINABLE MANAGEMENT OF THE NORTH SEA»

RESEARCH CONTRACTS MN/DD/60, 61 & 62

INTEGRATED AND DYNAMICAL OCEANOGRAPHIC DATA MANAGEMENT

JOINT SCIENTIFIC REPORT
for the year 2000

January 2001

The present document gathers the yearly scientific reports of the three partners to the *Integrated and Dynamical Oceanographic Data Management* project, performed on behalf of the Federal Office for Scientific, Technical and Cultural Affairs.

It covers the activities performed during the year 2000. It contains, in sequence the contribution (and their annexes) of :

- the Management Unit of the Mathematical Model of the North Sea (MUMM),
- the SURFACES laboratory (ULg),
- the *Universitair Centrum voor Statistiek* (*KUL*).

MANAGEMENT UNIT OF THE
NORTH SEA MATHEMATICAL MODELS

DATA MANAGEMENT GROUP

Integrated and Dynamical Oceanographic Data Management - IDOD

Karien De Cauwer, Mia Devolder, Siegrid Jans, Serge Scory

DMG/0/KDC,MD,SJ,SS/200101/EN/IR

Joint Scientific Report n°4 (January-December 2000) Prepared for OSTC, contract MN/DD/60

MUMM
100 Gulledelle
B–1200 Brussels
Belgium

List of Annexes

1.Introduction

In the beginning of 2000, it was decided to move from the Access prototype towards an Oracle relational database that would better fit the needs of the IDOD–project. Therefore, during the elapsed year, much effort focused on the in–depth analysis of the full specification of the system and on the training of the people dedicated to it. MUMM also decided to get external assistance to fulfil its tasks in due time.

Firstly, a high–level analysis has been performed. The objectives, the constraints and the successive deployment phases have been reviewed. The conclusion of this *strategical* study are given in this report.

The development and implementation phase of the full–scale information system started in November and is due to finish in May 2001.

In parallel, the other IDOD tasks devoted to MUMM were ongoing: collection and inventory of the datasets (both “true” data, *i.e.* the scientific datasets and the “meta–data”, f.i. geographical information), co–ordination with the other teams of the network, co–ordination with the other teams of the Programme (most of the conventions defining the rules of data use are now approved), international duties (OSPAR/ICES reporting, participation in international joint actions of oceanographic data centres, *etc.*)

The development of these tasks is briefly described hereafter.

2. Data sets

2.1. Inventory

The inventory of the datasets is continually adapted (see 2.1.1.)

At the same time, an inventory of the geographical information has been undertaken (see 2.1.2).

2.1.1. Data collected in the frame of the programme "Sustainable Development of the North Sea"

Only 50% of the 1999 data have been received by January 2001, while some data sets for 1997-98 are still incomplete or missing. The present situation for 1997, 1998 and 1999 is given in Figure 1.

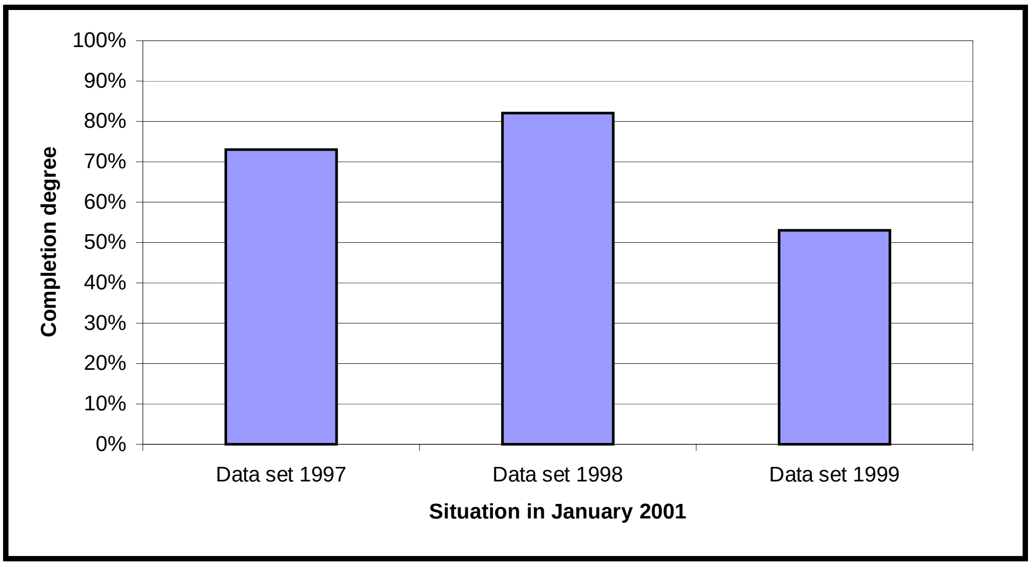


Figure 1. Evaluation of the data sets received by January 2001.

As already mentioned in previous reports, this delay in the data submission induces difficulties for the IDOD team in the organisation of the screening procedure and has led to a delay in the set-up of the project.

As shown in Annex 1 (Figures 1 and 2), some laboratories are able to meet their obligation for the agreed deadline, while others need more time to send their information. The reasons why data are not received in time are diverse and when known listed in Annex 1. It must be kept in mind to evaluate the extent of the problem. We are aware that some analyses are time consuming and induce a delay in the data transmission. Only exceptionally, it seems that a lack of goodwill is responsible for the situation.

2.1.2. “Geographic data”

Contacts took place with geographical data providers in order to collect useful information for visualisation and interpretation of the oceanographic data.

Geographic datasets have been received from mainly three different sources : Eurostat, Institute of Nature Conservation and Afdeling Waterwegen Kust.

Other data available from various internal and external sources were also collected. Contacts are still ongoing with the Euronav navigation firm.

The descriptive inventory of the available datasets is given in Annex 2. Maps representing the contents of these datasets are also provided.

The pertinence/degree of precision of these data is being evaluated using the GIS software.

2.1.3.Meta data inventory

Information about marine research projects executed by Belgian laboratories was gathered by MUMM in a database. This database was established according to the new format (EDMERP : *European Directory of Marine Environmental Research Projects*) defined during the Sea-Search action.

Besides information on projects, an inventory of Belgian marine datasets exists (EDMED). The information of both meta-databases was printed in a report for every laboratory concerned and sent for verification, update and insertion of new entries. The update of the information is actually ongoing.

2.2. Quality Control : at-source and at-time data entry

The incentive to make a proposal for systematically gathering information on samples is based on some demands from data providers and data users. Moreover, as data managers, we often experience difficulties in finding correct information on the samples.

Therefore, to improve the quality of the meta–information related to samples collected onboard of the Belgica, an application has been developed to allow a continuous tracking of the sampling process.

Some data providers formulated the need to registrate the sampling time and position automatically on board of the research vessel.

The data users have the wish to consult complete and correct datasets on all samples collected on board of the research vessel in the IDOD-database. To fulfil this demand, the data managers had until now to browse through four documents to locate a sample (campaign programme, cruise report, ROSCOP forms and the reported data). From time to time, these four information sources contain contradictory information and give no straight answer on which samples were in reality taken.

The application for sample registration will make it possible to follow a sample from the beginning (sample taken) to the end (data input in the database) and would allow a much more complete and easy data transfer from the data provider to the data manager. We are convinced that this procedure will enhance the quality of the information in the IDOD-database.

Other advantages :

- Cruise programme : by clicking a button, a sampling programme is generated that can be sent to the chief scientist for inclusion in the cruise programme.
- Link with ODAS : scientists will be able to leave the Belgica with a file containing the information on their samples including the actual position, meteorological parameters, ...
- Cruise report : by clicking a button, a report is generated containing the information on samples actually taken.

Future advantages/possibilities :

- Linkage of sample information with a barcode system,
- Storage of in situ data : values could be entered on board *e.g.* for dissolved oxygen and pH.
- ROSCOP forms can be produced automatically, based on the information entered in the application.

This programme has been installed and tested onboard during monitoring campaigns. Some functions and improvements still have to be developed. The different steps of the procedure are described in Annex 3.

3. Database development

At the end of 1999, MUMM had a prototype of the database in Microsoft Access, already containing some actual data. A draft conceptual scheme for seawater, plankton and sediment as described in the Scientific Report 1999 was also available.

In the beginning of 2000, MUMM decided that the IDOD information system should move towards an ORACLE relational database to meet the full–scale specifications.

The objective was to use the prototype as a basis for the new information system since it already contained much of the needed entities (tables) and attributes (field in a table). Some of the required functionalities for the import, retrieval, update and deletion of data were also available but not fully operational. Other functionalities were missing. The functionalities of the new system had also to be reviewed with respect to the kind of data coming from the Programme “Sustainable development of the North Sea”, the requests for data we regularly receive and the (evolving) international reporting obligations.

In order to have an information system that covers all needs, MUMM decided also to hire external assistance from a consultancy company specialised in Oracle based solutions.

As a first step, we carried out together a high-level analysis to get a precise view of the scope, the workload and the budgetary implications of the new system.

3.1. High-level analysis (strategy phase)

The objectives for the new system were determined taking into account : the central storage of data (one central database system containing all data), the stability (having a system as stable as possible), the distribution of data (distribution as flexible as possible and making the data quickly available), the integration of new technologies and the possibility to have a global access- and security-approach so that groups of users can be defined and linked to user-profiles.

In this perspective, the following aspects are essential in the characteristics of the new system :

- making use of the most advanced software components available, both on database and tools level;
- the system must be fully documented and maintained by using an integrated CASE tool;
- the system must be based on the principal of Client-Server and using Personal Computers and a GUI (graphical user interface) like Microsoft Windows;
- When using other packages, as much of the existing functionality offered by the package should be used;
- Reconsider and maybe reuse of existing hardware infrastructure.

The most important *business objective* as defined by MUMM, is the possibility to capture and process different sample data and results. These data and results will be further used by the statistical and spatial analysis programs.

In September 2000, the high-level analysis was finished. The conclusion addresses the various aspects of the implementation phase :

- Process models;
- Function hierarchy;
- Entity relationship diagram;
- Proposal for external packages;
- Overview of the different phases and tasks to perform;
- Estimation of budget.

3.1.1. Process modeling

Several sessions were organised between MUMM and the consultant in order to determine the different processes and look at them in detail. A process model was created and discussed. This approach allowed to identify corrections and/or supplementary steps.

The following important processes were identified :

- Import of sample data and their results;
- Maintenance of sample data and results;
- Maintenance of reference information;
- Feeding of external systems;
- Exploit sample data and results;
- Consult information.

For each process, a descriptive overview was written down containing the different steps within the process.

During this phase, several documents have been gathered that play an important role within the different processes. Most of these documents describe the feeding of external systems (e.g. reporting to ICES, EDMED, ...).

3.1.2. Function hierarchy

Based on the information gathered during the previous steps, an initial global function hierarchy model has been created. The functions, together with the time estimated to develop them, are listed in Annex 5.

3.1.3. Entity relationship diagram

All existing tables were discussed in detail and most of them were withheld for the new system. All relationships between the different entities were determined in order to make sure that with the new model all questions concerning data retrieval could be answered and that all relations between the different entities were covered.

For several of these tables, n-n relationships existed and this had to be reflected in the diagram. For such a relationship, an entity was used that connects both involved entities.

As the prototype, at the time of the study, was not yet finished, several features were discussed that led to new entities and relationships.

3.1.4. Overview of the different phases and tasks to perform

The different tasks that have to be performed after the high-level analysis is added in Annex 4. Three phases can be distinguished : the analysis phase, the design phase and the build phase.

3.1.5. Budgetary estimate

The budgetary estimate was based on the functionalities as they existed at the end of the high-analysis phase. An overview for the different modules is given in Annex 5.

It should be noted that all modules identified so far have been taken up. On the base of the prior needs, MUMM determined the necessary priorities in the development of the modules.

3.1.6. Conclusions concerning timing

Together with those 160 man-days needed to implement the different modules, still about 40 man-days have to be reserved for database design, writing the analysis report and the design report, concluding the backup/recovery strategy, security implementation, initial upload, ... This makes a total of 200 man-days. By adding 25 % for contingency, an estimate of 250 man-days was concluded.

3.2. Analysis phase

In November 2000, the analysis phase was initialised. This phase uses the results of the strategy phase as input. These results will be verified and worked out to have a correct, executable model that acts as a base for the final development of the system.

This includes the detailed description of the entities, attributes, primary keys, ... , and of all the functionalities of the system.

This phase is implemented as a mixed team approach between MUMM and a full-time consultant in which the consultant provides the necessary project leading and support.

For this phase, a timing of about 2 months is foreseen. Consequently, this phase should be finished by the end of January 2001.

3.3. Next phases

The next phases are the Design phase (planned timing about 2.5 months) and the Build phase (planned timing about 2.5 months).

The design phase takes over the detailed requirements of the analysis phase and finds the best way to satisfy the needs, tries to reach and maintain the agreed service level, given the technical environment.

During the build phase programs are written and tested, using the appropriate tools. These tools are dependant on the technical environment and the type of programs.

3.4. Training courses

Since MUMM did not have the necessary knowledge to work in an ORACLE environment, the MUMM team followed some basic training courses in the ORACLE training center during the last semester of 2000.

- Business Modelling and Data Design with Designer : 5 days (2 persons)
- Develop PL/SQL Program Units : 3 days (3 persons)
- Developer Forms Part I : 5 days (3 persons)
- Application Design and Generation with Designer : 5 days (3 persons)

4. GIS: visualisation and queries

The spatial analyst functions available with the GIS tool can be of great interest to answer basic queries, such as distance mapping, density function, surface functions … It is also possible to determine suitable areas for a particular purpose, from a set of selected criteria. However, in order to allow more specific requests, programming trough the specific ArcView language (Avenue) is necessary. This is another main objective for the GIS applications.

A list of requests that should be answered by the GIS tool is given hereafter.

4.1.Visualisation of the Database query results :

The database query should have a possibility to click on a button to have access to a spatial analysis tool :

- to select a station, an area (draw rectangle, circle, polygon), or a « theme » (natural reserves, dredging areas, …);
- to visualise the result of a query e.g. selected parameter, period, … (: basic map with ponctual, graduated symbol or color representation);
- to make a simple query : show the results of the database query within 2 km from the coast or at a distance of 50 m from the pipelines.

4.2.Examples of queries to be considered :

Specific spatial analysis queries have been considered. Some of them are simple queries that should be developed as an automatised task, while other are more complex requests and will necessitate more developed functions.

- Determination of a buffer-zone
- Calculation of surface or distance
- Show the areas where the salinity is higher than X in winter
- Interpolation: some automatisation should be envisaged, for the main parameters, and the maps produced should be well-documented with the methods used, the parameters chosen… A minimum of point measurements is probably necessary to produce a realistic map : so if this need is not answered, the tool should not make it !
- Where is the maximum concentration of this parameter ?
- Show the metals concentrations (max. or mean for a ten years period) measured at a distance of 50m of the pipelines and/or wrecks positions.
- Give the stations where metal concentration is higher than X and which are at a distance of X meters from a pipeline.
- Find the areas where fish concentrations is higher than x and where bird concentrations is higher than x in winter.
- Determine the area where distance from pipelines and communications cables is smaller than x km.
- Determine the surface of natural reserves (or % of the Belgian continental

shelf)

- Find the wrecks where depth is smaller than x meters.
- Create a buffer zone of 50 meters around all natural reserves
- From a map representing the ponctual measurement of one parameter, retrieve other parameter information (depth, salinity…) by clicking on the station.
- Giving the geographical co-ordinates of a point, retrieve information of a chosen layer.
- Show the results of two maps : Determine the area where there are birds colonies and where temperature is higher than X in winter.
- Determine the shortest way to go from A to B (geographical co-ordinates or click on the map), avoiding depth smaller than X meters during the low tides (and/or avoiding zone where there are pipelines and cables at X meters).
- Interpolation of ponctual measurements
- More complex queries
- Decision making : find the better area for a particular purpose : the « user » should determine which parameters are to be considered (eventually in order of importance)

This list of requests has been discussed with the ULg team in order to determine the best way to resolve these queries. The detailed analysis of the spatial and temporal requests has been made by ULg and is presented in their report.

5. Products

5.1. International reporting obligations

MUMM has fulfilled the Belgian international obligations in the frame of the "Joint Assessment and Monitoring Programme" (JAMP) and "Nutrient Monitoring Programme" of the Oslo and Paris Commission (OSPAR) by reporting the monitoring data for 1998 and 1999 to ICES. Due to a delay in the data submission for 1998, a first draft of the National Comments had been sent last year. In July 2000, the full report 1998 was sent, together with the 1999 results.

5.2. Newsletter 4

Due to the heavy workload to prepare and then initiate the system implementation, only one Newsletter has been issued in 2000. A copy of it is given in Annex 6.

COLOPHON

This report was issued by MUMM in January 2001.

Its reference code is DMG/0/KDC,MD,SJ,SS/200101/EN/IR.

Status
☐ draft
☒ final version
☐ revised version of document
☐ confidential

Available in
☒ English
☐ Dutch
☐ French

If you have any questions or wish to receive additional copies of this document, please send an e-mail to *idod@mumm.ac.be*, quoting the reference, or write to:

MUMM
100 Gulledelle
B–1200 Brussels
Belgium
Phone: +32 2 773 2111
Fax: +32 2 770 6972
http://www.mumm.ac.be/

MANAGEMENT UNIT OF THE
NORTH SEA MATHEMATICAL MODELS

MODELLING GROUP

The typefaces used in this document are Gudrun Zapf-von Hesse’s *Carmina Medium* at 10/14 for body text, and Frederic Goudy’s *Goudy Sans Medium* for headings and captions.

		COORDINATOR	Vincx			Van Grieken				Lancelot			Dubois			Bouquegneau			
		PARAMETER	UG-Vincx	IN-Kuijken	KUL-Ollevier	VUB-Baeyens	UG-V.Langenhove	UIA-Van Grieken	ULB-Wollast	ULB-GMMA	VUB-ECOL	MUMM	ULB-Dubois	UMH-Jangoux	UMH-Flammang	ULg-Bouquegneau	Ulg-Coignoul	VUB-Joiris	IN-Meire
M	General	Date	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
E		Time	x	x	x	x	x	x	x	x	x	x				x			
T		Position	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
A	Methods	Sampling method	x	x	x	?	x	x	x	x	x	x	x	x	x		?	?	?
D		Sample handling	x	x	x	?	x	x	x	x	x	x	x	x	x		?	?	?
A		Analysis method	x	x	x	?	x	x	x	x	x	x	x	x	x	x	x	?	?
T		Statistical analysis	x	x	x								x	x	x				
A		Programmes used											x	x	x				
	Quality control	Quasimeme scores				x	x	?	x			?				x	?	?	?
		Certified references						?				?	x		x		?	?	?
		Internal references						?		x		?	x		x		?	?	?
		Intercalibration				x		?	x			?						?	?
	Meteorology	PAR										x							
		Wind speed	x				x					x							
		Wind direction	x				x					x							
		Solar radiation	x				x												
		Atmospheric pressure	x				x												
		Temperature	x				x												
W	Physical	temperature	x (v)				x			x		x (v,c)							
A		suspended matter							x	*		x							
T		depth	x (v)				x						x	x	x				
E		secchi depth										x							
R		PAR	x (v)																
		profile of the pore							x										
		specific surface							x										
		average pore radius							x										
	Major inorganic	salinity	x (v)				x			x		x (v,c)	x	x	x				
		dissolved oxygen																	
		pH																	
		carbon								*									
	Nutrients	nitrate	x			x				x									
		nitrite	x			x				x									
		phosphate	x			x				x									
		silicate				x				x									
		ammonia	x			x				x									
		urea				x													
	Metals	Cu, Cd, Ni, Zn, Pb				x (s,p)			x (p)										
		Mn, Co, Cr, Al, Ca, Fe, Si							x (p)										
		As				x (s,p)													
		Hg°				x													
		monomethyl Hg				x (s,p)													
		total Hg				x (s,p)													
	Major organic	organic nitrogen				x (s,p)													
		total nitrogen							x (p)										
		organic carbon				x (p)			x (p)	*									
		total carbon							x(p)	*									
	Pigments	Chlorophyl-a	x						x	x	x	x							
		Chlorophyl-c	x																
		fuccoxanthine	x								x								
		19'hexanoxanthine									x								
		diadinoxanthine									x								
		alloxanthine									x								
		peridinine									x								
		phaeopigment							?			x							
	Aromatic hydrocarbons	benzene					x												
		toluene					x												
		m/p/o-xylene					x												
		ethylbenzene					x												
	Organochlorines	chloroform					x												
		tetrachloromethane					x												
		1,1-dichloroethane					x												
		1,2-dichloroethane					x												
		1,1,1-trichloroethane					x												
		trichloroethylene					x												
		tetrachloroethylene					x												
	Optical	downwelling PAR irradiance										x (v)							
		optical backscatter	x(v)									x (v)							
		scalar PAR irradiance										x (v)							
		upwelling fluoresc. radiance										x (v)							
		PAR attenuation coefficient										x							
		upwelling radiance spectra										x							
		downwelling irradiance spectra										x							
		sub-surface irradiance spectra										x							
		phytoplankton absorpt. spectra										x							
		yellow subst. absorpt. spectra										x							

ANNEX 1

		COORDINATOR	Vincx			Van Grieken				Lancelot			Dubois			Bouquegneau			
		PARAMETER	UG-Vincx	IN-Kuijken	KUL-Ollevier	VUB-Baeyens	UG-V.Langenhove	UIA-Van Grieken	ULB-Wollast	ULB-GMMA	VUB-ECOL	MUMM	ULB-Dubois	UMH-Jangoux	UMH-Flammang	ULg-Bouquegneau	Ulg-Coignoul	VUB-Joiris	IN-Meire
A	**nutrients**	NO2-, NO3-, SO4-, PO4-						x											
I	**heavy metals**	Ag, Cu, Zn, Fe						x											
R	**major inorganic**	F-, Cl-, Si, sea salt, gypsum,						x											
		alumino-silicates						x											
S	**Physical**	granulometry	x										x						
E	**Interstitial water nutrients**	nitrate	x (v)																
D		nitrite	x (v)																
I		ammonia	x (v)																
M		phosphate	x (v)																
E	**Interstitial water pigments**	Chlorophyl-a	x (v)																
N		Chlorophyl-c	x (v)																
T		fuccoxanthine	x (v)																
	Metals	Cd,Cu,Pb,Zn											x						
	Chlorobiphenyls	CB(28,52,101,118,138,153,18																	
		0,37 ,77,81,126,169)													x				
B	**Fish - varia**	genetic structure			x														
I		parasites : spp. and incidence			x														
O		stomach analysis	x																
T	**Sea urchin**	embryotoxicity test												*					
A	**biological effect**	metamorphosis test												x					
	Starfish metals	Cd,Cu,Pb,Zn											x						
	chlorobiphenyls	CB(28,52,101,118,138,153,18																	
		0,37, 77,81,126,169)													x				
	biological effect	MFO^{1} activity											*						
		amoebocyte RO^{2} species											*						
		embryotoxicity test												x					
		amoebocyte phagocytic activity											x						
		metallothioneins											*						
	Seabirds - marine mammals																		
	metals	Cd,Cr,Cu,Fe,Ni,Pb,Zn														x			
		selenium														x			
		total Hg																x	
		methyl Hg																x	
	organic	hydrocarbons																x	
		PAHs															*		
		polar lipids																x	
		total lipids														x			
		PCBs																x	
	organochlorines	DDE																x	
		DDT																x	
		aldrin																x	
		lindane																x	
		heptachlor epoxide																x	
	ecology	diversity																	x
		density																	x
		#per sp, devel., plumage stage		x															x
	varia	pathology															x		
		parasites															x		
		metallothioneins														x			
	Plankton: phyto, bacterio,	composition								x									
	nanozoo & microzoo	enumeration								x									
	mesozoo	composition									x								
		abundance									x								
	Benthos **(meio, macro, epi, hyper)**																		
	ecology	diversity index	x																
		# species	x																
		density	x																
		biomass	x																
		dominance index	x																
		length freq. distribution	x																
		weight freq. distribution	x																
	Maps - satellite derived	suspended matter										x							
		chlorophyll-a										x							

x : parameter expected

* : parameter not measured in 1997 and 1998

? : lack of information about this parameter (to be precised by the laboratory)

[1]MFO mixed function oxidases

²RO : reactive oxygen

p : particulate; s : dissolved, v : vertical profile, c : continuous measuremer

ANNEX 1

nts

RESPONSIBLE PERSON	**COMMENTS**
Pr. M. Vincx	Several data are already provided (epibenthos, hyperbenthos, nutrients, meiobenthos), other data are still missing (f.e. macrobenthos, stomach analysis, some epi- and hyperbenthos). Analysis time is long and data were collected by different persons.
Pr. E. Kuijken	Data since 1992 were received in one MS Access database. As all data are consequently added to this file, no problems are expected to receive the data of 1999. Expected receival at time of the visit (not yet done a/o due to PC-problems and then other priority tasks of IDOD).
Pr. J-M. Bouquegneau	The data are always received at time, well documented and nearly complete, except for the results concerning metallothioneins (due to a delay in the analyses).
Pr. Fr. Coignoul	Documentation (meta-information and autopsy reports) is missing concerning pathological analyses.
Pr. Cl. Joiris	An example of the results that IDOD would receive in future was sent to IDOD in September 1999 (2 excel files). In October 1999, IDOD had a meeting with the labo. During this meeting, it was decided that the labo would send IDOD complete data files, including information on methodology, precision and quality control. In March 2000, IDOD sent an email to repeat this question for complete data files and including meta-information. Together with the laboratory, it was decided that the date to receive this information would be 27.04.2000 instead of the end of March 2000 for the years 1997, 1998 and 1999. On 23.05.2000, IDOD asked for another time for this information but no response from the laboratory was received until now.
Pr. P. Meire	Data since 1992 were received in one MS Excel-file. As all data are consequently added to this file, no problems are expected to receive the remaining data of 1999. Expected receival at time of the visit (not yet done a/o due to PC-problems and then other priority tasks of IDOD).
Pr. Ch. Lancelot	On the demand of the PDM who was finishing a doctorate, IDOD gave permission to send the data for the year 1999 later than March 2000 since we had complete datasets for 1997 and 1998 and a very good cooperation with the institute. The institute is finishing the data for 1999 and will send them to IDOD as soon as possible.
Dr. M. Tackx	In the dataset of 1999, the results for pigments are missing.
Pr. W. Wollast	Some analyses are time consuming. Some data are still missing for the years 97-98 and the results for 1999 analyses are not yet available.
Pr. R. Flammang	The PCBs analysis (1999) was not finished in March 2000. The complete file will be sent in March 2001. But the same problem will arise each year, and another deadline should be envisaged

Figure 2 – Data sets inventory: comments concerning the incomplete/not received data sets

Annex 2 - Inventory of GIS data to consider in the frame of the IDOD project

1.Introduction

In the frame of the IDOD project, a spatial analysis tool is developed in order to visualise the oceanographic data and to help their interpretation. Moreover, specific requests can also help decision making for different types of activities.

The spatial analysis functions will be envisaged with the use of a GIS (Geographic Information System, ArcView 3.2). Digital geographical data are necessary to elaborate the maps. The types of information required have been evaluated in relation to the actual and possible needs of this tool.

The aim of this report is to make an inventory of the digital data available and to have a first visualisation of the possibilities of the GIS.

2. Definition of the interest zones

Four zones have been defined, with increasing interest respectively (see map 12):

1. European Continental Plate
   44°N, 12°W – 62°N, 12°E

2. North Sea and Channel
   West limit: line between 6°W, 48°N and 4°W, 59°N
   East limit: 12°E

3. South Bay of the North Sea and Channel East
   50°N, 0°W – 53°N, 4°E

4. Belgian Coastal Zone (=territorial sea + EEZ)
   Delimited by the coast in one part and by a line composed by the following points in the other part.

51°16’09”N	02°23’25”E
51°33’28”N	02°14’18”E
51°36’47”N	02°15’12”E
51°48’18”N	02°28’54”E
51°52’34.012”N	02°32’21.599”E
51°33’06”N	03°04’53”E

3. Types of geographical data required

A list of the geographical features of interest has been set up. This is not an exhaustive list but it includes the main data that should be useful for interpretation of oceanographic processes or for decision making.

These are:

- Coastlines
- Bathymetry
- Estuaries, rivers
- Continental shelf limit
- Terrestrial frontiers
- Navigation routes
- Extraction zones
- Pipelines
- Communication cables
- Anchoring areas
- Military exercises areas
- Wrecks
- Ammunition disposal
- Zones “Bonn Agreement” (1983)
- Zones ICES
- Administrative zones: seaward boundary of provinces/municipalities, 3 and 12 miles boundaries

Due to the difficulty to find these data (and yet more difficult to find them in homogenous files), the choice has been made to focus principally on the Belgian coastal region, while continuing the prospect for data about the third other areas. As other features, activity zones or areas of particular interest will be defined in the future, the catalogue of data layers will be continually kept up to date.

4. Catalogue of the available geographical data

Different data sources have been explored and a list of the actual state of the available data files from the main data sources are described in table 1. A more detailed description for each layer is given in table 2. Maps are also drawn in order to visualise the extent and the content of these data sets.

List of figures of annex 2

Table 1 – Available Geographical datasets

Name	Description	Coverage	Source	Original Name	Copyrights	Format
ESRI-Europe	Terrestrial data: country and administrative boundaries, cities, lakes, rivers, highways, demographic attributes	Europe (Zone 1)	ESRI (Cd-Rom)	ESRI Data and Maps CD	Copyright-esri	*.shp, *.shx, *.dbf
GISCO	Terrestrial data: administrative limits, altimetry, infrastructure, hydrology, land and nature resources	Europe (Zone 1)	Eurostat Data Shop	GISCO Cd-Rom	Copyright-Eurostat	*.e00 (ArcInfo)
Limited Atlas - BCP	Bathymetry, cities, BCP, 3-6-12 nmiles limits, gas pipelines, telephone cablesn dumpsite, shooting areas, paardemarkt, bird areas, sand extraction areas, oil slicks, Natura2000	Belgian part of the North Sea (Zone 3)	Institute of Nature Conservation	Limited Atlas of the Belgian Part of the North Sea	?	*.shp, *.shx, *.dbf
Bathymetry - Vlaamse Banken	Bathymetry of a part of the Vlaamse Banken	Part of the BCP (Part of zone 4)	AWK	d11		dxf
Coastlines - BSEX	European coastlines	Europe (Zone 2 + part of Zone 1)	BSEX Project	gmd420 gmd426	?	ascii
Bathymetry soundings - AWK	Ponctual soundings on a part of the BCP	Part of the BCP (Part of zone 4)	AWK	vlb_xyz.lis vlb_xyz_id.lis	?	ascii
QSR dataset	Europe frontiers and coastlines, continental platforms, ICES areas, catchment area, eurostreams.	Europe (Zone 1)	OSPAR	QSR dataset	?	*.shp, *.shx, *.dbf

Table 2 – Geographical data layers

File name	Source	Description	Zone covered
Admins - ESRI	ESRI- Europe	Administrative informations: province, republic, region or independent town for each country, population, superficie, ,,,	Zone 1 (terrestre)
Cities - ESRI	ESRI- Europe	Cities informations: city name, country name, administrative name, status, population class	Zone 1 (terrestre)
Countries - ESRI	ESRI- Europe	European countries	Zone 1 (terrestre)
Major rivers - ESRI	ESRI- Europe	Major european rivers	Zone 1 (terrestre)
Rivers - ESRI	ESRI- Europe	Other european rivers	Zone 1 (terrestre)
Major urban - ESRI	ESRI- Europe	Major european urbanized areas	Zone 1 (terrestre)
Urban - ESRI	ESRI- Europe	Other european urbanized areas	Zone 1 (terrestre)
Places - ESRI	ESRI- Europe	Populated places (without or within urbanized areas)	Zone 1 (terrestre)
Railroads - ESRI	ESRI- Europe	Railroads: types (added railroad connector, single/multiple track railroad), status (functioning, schematic road/urbanized areas only, under construction)	Zone 1 (terrestre)
Roads - ESRI	ESRI- Europe	Roads: types (primary and secondary roads, dual line highway, track, trail, foothpath, connector within urbanized areas), status (functioning, schematic rail line/urbanized areas only, compiled railroads)	Zone 1 (terrestre)
Water - ESRI	ESRI- Europe	Perennial inland water	Zone 1 (terrestre)
			Zone 1 (terrestre)
ADARNE1MV7/arne1mv7gg	GISCO	Administrative regions (1:1million scale, Version 7, 1999	Zone 1 (terrestre)
ADNUEC1MV6/nuec1mv6gg	GISCO	Admin/NUTS regions 1-95 (1:1million scale, version 6, 1995)	Zone 1 (terrestre)

ADNUEC1MV6/nu291v6lcgg	GISCO	Admin/NUTS regions 1-91 (sub-layer)	Zone 1 (terrestre)
ADNUEC1MV6/nu292v6lcgg	GISCO	Admin/NUTS regions 1-92 (sub-layer)	Zone 1 (terrestre)
ADNUEC1MV6/nu293v6lcgg	GISCO	Admin/NUTS regions 1-93 (sub-layer)	Zone 1 (terrestre)
ADNUEC1MV6/nu294v6lcgg	GISCO	Admin/NUTS regions 1-94 (sub-layer)	Zone 1 (terrestre)
ADNUEC3MV6/nuec3mv6gg	GISCO	Admin/NUTS regions 3-95 (1:3 million scale, version 6, 1995)	Zone 1 (terrestre)
ADNUEC3MV6/nu291v6lcgg	GISCO	Admin/NUTS regions 3-95-91 (sub-layer)	Zone 1 (terrestre)
ADNUEC3MV6/nu292v6lcgg	GISCO	Admin/NUTS regions 3-95-92 (sub-layer)	Zone 1 (terrestre)
ADNUEC3MV6/nu293v6lcgg	GISCO	Admin/NUTS regions 3-95-93 (sub-layer)	Zone 1 (terrestre)
ADNUEC3MV6/nu294v6lcgg	GISCO	Admin/NUTS regions 3-95-94 (sub-layer)	Zone 1 (terrestre)
ADNUEC1MV7/nuec1mv7gg	GISCO	Admin/NUTS regions 1-99 (1:1 million scale, version 7, 1999)	Zone 1 (terrestre)
ADNUEC1MV7/nu291v7lcgg	GISCO	Admin/NUTS regions 1-99-91 (sub-layer)	Zone 1 (terrestre)
ADNUEC1MV7/nu292v7lcgg	GISCO	Admin/NUTS regions 1-99-92 (sub-layer)	Zone 1 (terrestre)
ADNUEC1MV7/nu293v7lcgg	GISCO	Admin/NUTS regions 1-99-93 (sub-layer)	Zone 1 (terrestre)
ADNUEC1MV7/nu294v7lcgg	GISCO	Admin/NUTS regions 1-99-94 (sub-layer)	Zone 1 (terrestre)
ADNUEC3MV7/nuec3mv7gg	GISCO	Admin/NUTS regions 3-99 (1:3 million scale, version 7, 1999)	Zone 1 (terrestre)
ADNUEC3MV7/nu291v7lcgg	GISCO	Admin/NUTS regions 3-99-91 (sub-layer)	Zone 1 (terrestre)
ADNUEC3MV7/nu292v7lcgg	GISCO	Admin/NUTS regions 3-99-92 (sub-layer)	Zone 1 (terrestre)
ADNUEC3MV7/nu293v7lcgg	GISCO	Admin/NUTS regions 3-99-93 (sub-layer)	Zone 1 (terrestre)
ADNUEC3MV7/nu294v7lcgg	GISCO	Admin/NUTS regions 3-99-94 (sub-layer)	Zone 1 (terrestre)
ADNUEC10MV7/nuec10mv7gg	GISCO	Admin/NUTS regions 10-99 (1:10 million scale, version 7, 1999)	Zone 1 (terrestre)
ADNUEC10MV7/nu291v7lcgg	GISCO	Admin/NUTS regions10-99-91 (sub-layer)	Zone 1 (terrestre)
ADNUEC10MV7/nu292v7lcgg	GISCO	Admin/NUTS regions10-99-92 (sub-layer)	Zone 1 (terrestre)
ADNUEC10MV7/nu293v7lcgg	GISCO	Admin/NUTS regions10-99-93 (sub-layer)	Zone 1 (terrestre)
ADNUEC10MV7/nu294v7lcgg	GISCO	Admin/NUTS regions10-99-94 (sub-layer)	Zone 1 (terrestre)
ALDe/deeu20m	GISCO	Altimetry/Digital elevation model (1:20 million scale)	Zone 1 (terrestre)
HYWP/wpeu10mv2gg	GISCO	Hydrography/water patterns (1:10 million scale)	Zone 1 (terrestre)
HYWSEU10M/wseu10mgg	GISCO	hydrography/water sheds (1:10 million scale)	Zone 1 (terrestre)
HYWSEU3M_/wseu3mgg	GISCO	hydrography/water sheds (1:3 million scale)	Zone 1 (terrestre)
INAp/apeugg	GISCO	Infrastructure/airports (precision 1 minute)	Zone 1 (terrestre)
INApTn/apeutnsyv4gg	GISCO	Infrastructure/airportsTEN (Transeuropean network, precision 1 minute)	Zone 1 (terrestre)
INPo/poeugg	GISCO	Infrastructure/ports (precision 1 minute)	Zone 1 (terrestre)

INPO-Tn	GISCO	Infrastructure/ports TEN	Zone 1 (terrestre)
INRd/rdeu1mv4gg	GISCO	Infrastructure/roads (1:1 million scale)	Zone 1 (terrestre)
INRw/rweu1mv4gg	GISCO	Infrastructure/railways (1:1 million scale)	Zone 1 (terrestre)
INSt/steugg	GISCO	Infrastructure/settlements (precision 1 minute)	Zone 1 (terrestre)
LR-FA10m/faeu10mg	GISCO	Land resources/Fishing areas (1:10 million scale)	Zone 1 (terrestre)
LR-FA3m/faeu3mgg	GISCO	Land resources/Fishing areas (1:3 million scale)	Zone 1 (terrestre)
NRLs/lseugg	GISCO	Nature resources/Landscapes (1:25 million scale)	Zone 1 (terrestre)
Nrvg/vgeugg	GISCO	Nature resources/Vegetation (1:3 million scale)	Zone 1 (terrestre)
Nrvg/vgeul2gg	GISCO	Nature resources/Vegetation (1:3 million scale)	
WorldDataBase/altimetry	GISCO	World Data Base/ altimetry	Zone 1 (terrestre)
WDFa/fawd25mg	GISCO	World Data Base/ fishing areas (1:25 million scale)	Zone 1 (terrestre)
Bathymetry – IN	Limited Atlas - IN	Bathymetry: land, - 2, 5, 10, 20, 30, 40, 50, 60, 70, 80 m	Zone 4 + part of zone 3
3nmile - IN	Limited Atlas - IN	3 nmile limit	Zone 4
6nmile - IN	Limited Atlas - IN	6 nmile limit	Zone 4
12nmile - IN	Limited Atlas - IN	12 nmile limit	Zone 4
24nmile - IN	Limited Atlas - IN	24 nmile limit	Zone 4
Belgian Continental Platform - IN	Limited Atlas - IN	Belgian Continental Platforms	Zone 4
Belgium Borders - IN	Limited Atlas - IN	Belgian frontiers	Zone 4
Bird areas - IN	Limited Atlas - IN	Bird areas determined on original results from regular aerial, ship- and landbase surveys (1992-1998)	Zone 4
Cities - IN	Limited Atlas - IN	Main cities	Zone 4 + part of zone 3
Dumpsites - IN	Limited Atlas - IN	Dumping sites of dredged material	Zone 4

Gas InterC - IN	Limited Atlas - IN	Interconnector: operational in October 1998, joins Bacton on the South Coast of England and Zeebrugge.	Zone 4 + part of zone 3
Gas Norfra - IN	Limited Atlas - IN	NorFra: operational in 1998, joins the Daupner E Platfrom on the Norwegian continental shelf and the port of Dunkerque in France.	Zone 4 + part of zone 3
Gas Seapi - IN	Limited Atlas - IN	Seapipe (zeepipe): in operation since 1st Oct. 1993, joins a pipeline from the Sleipner area on the Norwegian continental shelf with the Distrigaz terminal in the port of Zeebrugge.	Zone 4 + part of zone 3
Indwaste - IN	Limited Atlas - IN	Former dumping site for industrial waste	Zone 4
Natura 2000 - IN	Limited Atlas - IN	Proposed Special Conservation Area under the EC Habitats Directive (17000 Ha, comrpising the entire Ramsar 'Vlaamse Banken' site)	Zone 4
Oil slicks 9195 - IN	Limited Atlas - IN	Results of aerial surveiilance (MUMM: 1991-1995): illegal operational oil pollution from ships according to the estimated volume of the discharges)	Zone 4 + part of zone 3
Paardemarkt - IN	Limited Atlas - IN	Former dumping site for war munition	Zone 4
Ramsar - IN	Limited Atlas - IN	Sand bank area from Oostende to the French-Belgian border, extending 3nmiles from the baseline and shaloower than -6m MLLWS: designated as a Belgian Marine Wetland of International Importance under the Ramsar convention (1984)	Zone 4
Sand extraction areas - IN	Limited Atlas - IN	Sand and gravel extraction areas	Zone 4
Shooting areas - IN	Limited Atlas -	Military exrecises area	Zone 4

	IN		
Telef2 - IN	Limited Atlas - IN		Zone 4 + part of zone 3
TSS - IN	Limited Atlas - IN		Zone 4 + part of zone 3
Bathymetry - Vlaamse Banken AWK	AWK	Bathymetry of a part of the Vlaamse Banken	Part of zone 4
Coastlines - BSEX	BSEX Project	European coastlines	Zone 1
Bathymetry - Soundings AWK	AWK	Complete file of the sounding points	Part of zone 1
Akkaert93 - AWK	AWK	Akkaertbank 1993	Part of zone 1
Akkaert93 – AWK	AWK	Akkaertbank 1993	Part of zone 1
Bligh99 – AWK	AWK	Blighbank 1999	Part of zone 1
D11-1998 – AWK	AWK	Depth of the cart D11 edition 1998	Part of zone 1
Dfairy93 – AWK	AWK	Detail Fairybank 1993	Part of zone 1
Fairy94 – AWK	AWK	Fairybank 1994	Part of zone 1
Goote98 – AWK	AWK	Gootebank 1998	Part of zone 1
Hinder95 – AWK	AWK	Belgian Territory SW Fairybank	Part of zone 1
Kwinte95 – AWK	AWK	Kwintebank + Middelkerkebank 1995	Part of zone 1
Ody97 – AWK	AWK	Oostdyck 1997	Part of zone 1
Raan95 – AWK	AWK	Vlakte van de Raan 1995	Part of zone 1
Raan95 – AWK	AWK	Vlakte van de Raan 1995	Part of zone 1
Rat96 – AWK	AWK	Buitenratel 1996	Part of zone 1
Ruyt92 – AWK	AWK	In + out Ruytingen + Bergues Bank 1992	Part of zone 1
Thor97 – AWK	AWK	Thorntonbank 1997	Part of zone 1
Wdh97 – AWK	AWK	Territory Westende - De Haan 1997	Part of zone 1
Wh96 – AWK	AWK	Westhinderbank 1996-1997	Part of zone 1
Whbe94 – AWK	AWK	Gebied tussen Westhinderbank - Gootebank 1994	Part of zone 1
Whroute93 – AWK	AWK	Navigation route Westhinder 1993	Part of zone 1

Wrak_1998 – AWK	AWK	Wreck depth from the cart d11 edition 1998	Part of zone 1
Ws96 – AWK	AWK	Wielingen Scheur 1996	Part of zone 1
Ws98 – AWK	AWK	Wielingen Scheur 1996	Part of zone 1
Zw97 - AWK	AWK	South coast - Westende	Part of zone 1
Names - QSR	QSR	Names of the main features represented (cities, rivers, places…)	Zone 2 + Part of zone 1
Catchmentarea - QSR	QSR	Catchment areas of the Greater North Sea	Zone 2 + Part of zone 1
Eurostreams - QSR	QSR	Drainage area: main river systems considered in the study of the Greater North Sea	Zone 2 + Part of zone 1
Contplat - QSR	QSR	Continental Platforms	Zone 2 + Part of zone 1
Icesgrens - QSR	QSR	ICES regions	Zone 2 + Part of zone 1
Europa - QSR	QSR	European coastlines and frontiers	Zone 2 + Part of zone 1
Geoline - QSR	QSR	Geographic griding	Zone 2 + Part of zone 1

Annex 3 – Onboard registration Programme

1.Introduction

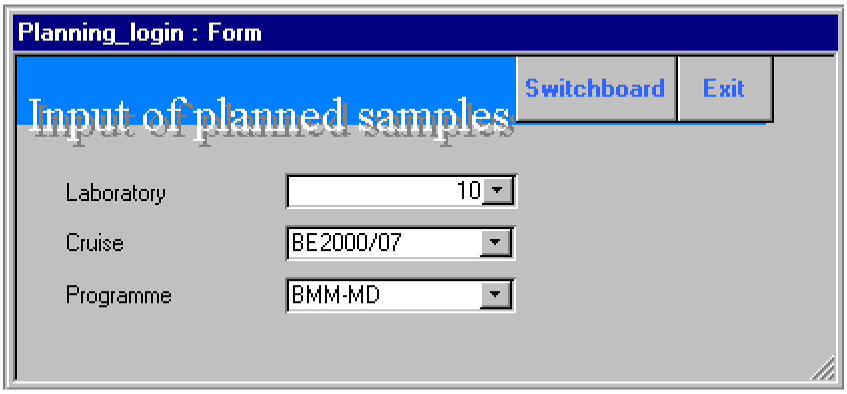
The registration of samples taken onboard of the RVBelgica can be divided in three stages.

- The first phase consists of entering the sample planning. This occurs one month before the campaign by the individual laboratories participating to the campaign. The file should be returned to MUMM-Ostend.
- The second phase is executed by MUMM-Ostend only, the sample plannings of the individual laboratories are merged in one sample planning for the whole cruise.
- The third phase is the registration of sampling time and sampling conditions onboard of the ship.

The three stages are described as three different applications hereafter.

2.Application Sample Planning

To start with this application, the form 'Planning_login' should be opened.



Planning_login : Form

Input of planned samples

Laboratory: 10

Cruise: BE2000/07

Programme: BMM-MD

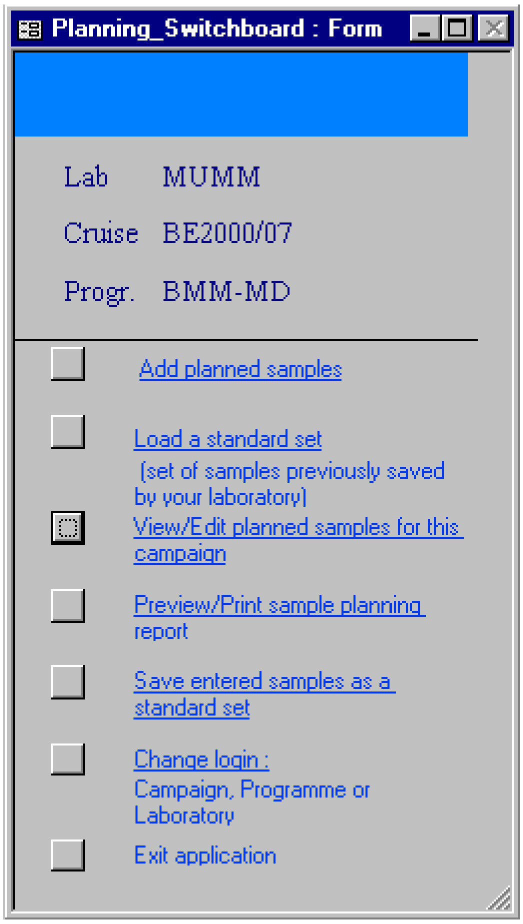
2.1.Login

The laboratory and the programme for which the sample is taken should be entered, together with the cruise during which the samples will be taken. The laboratory that will execute the sampling can however be different from the laboratories that 'orders' the sample.

The login form is hidden when the button 'Switchboard' is clicked. In the different modules the information selected there is used again (e.g. when a form is opened, only the samples for that particular login are showed.)

2.2.Switchboard

Lists all the menu items for the input of the sample planning. After executing a menu item, one should always return to the Switchboard by clicking the appropriate button.



Planning_Switchboard : Form
Lab MUMM
Cruise BE2000/07
Progr. BMM-MD
Add planned samples
Load a standard set
(set of samples previously saved by your laboratory)
View/Edit planned samples for this campaign
Preview/Print sample planning report
Save entered samples as a standard set
Change login : Campaign, Programme or Laboratory
Exit application

2.3.Add planned Samples

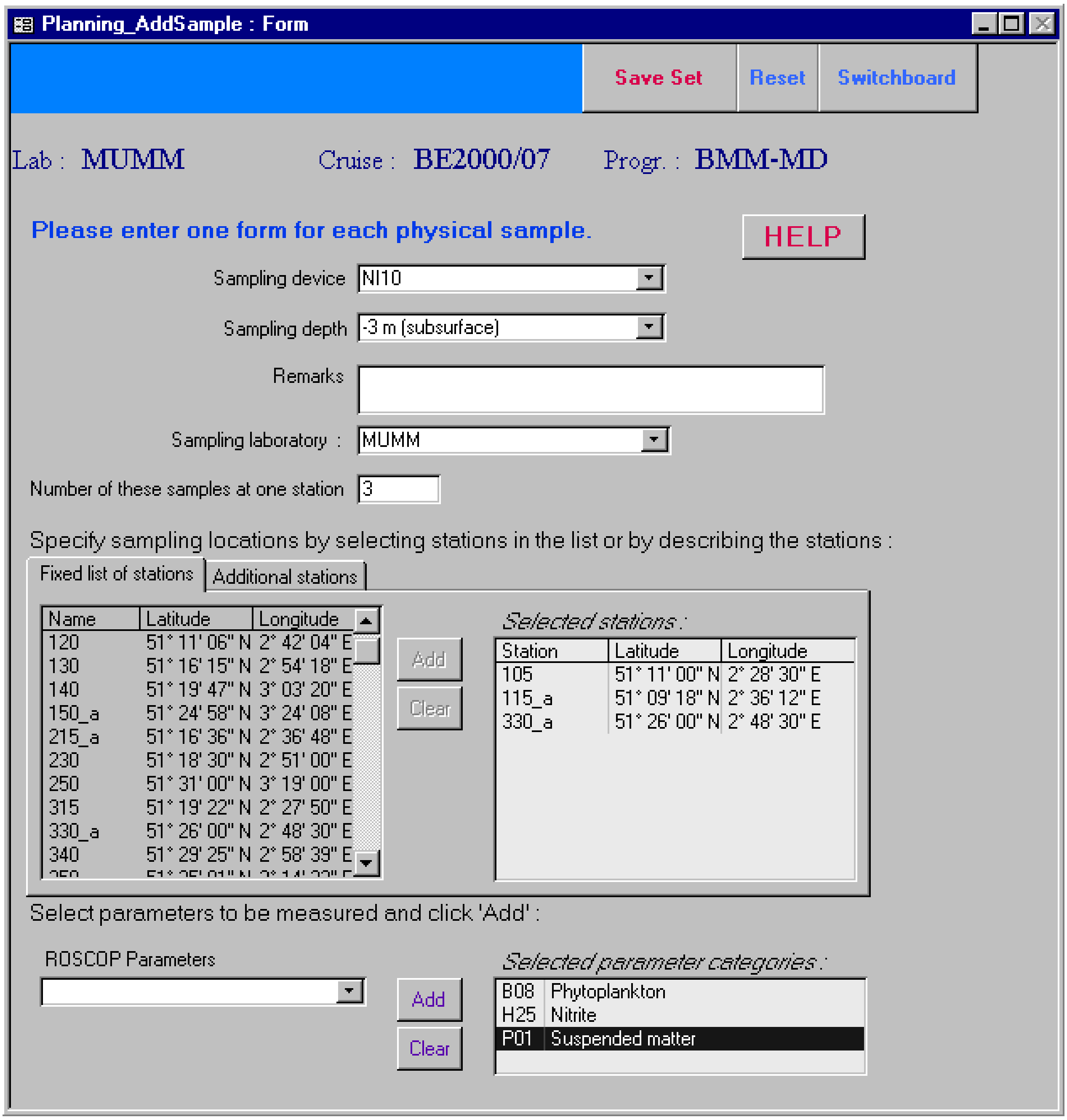
One physical sample can be entered for a list of sampling locations. A physical sample is defined as a single sample taken in the same conditions: same sampling device, sampling depth and sampling laboratory. For each physical sample, many parameters can be measured and all of them should be introduced in the ROSCOP parameters field.

The sampling location can be specified by selecting in the ‘Fixed list of stations’ and pushing ‘Add’, or can be described in ‘Additional stations’ (followed by clicking the button ‘Add’). When a sampling location is not in the list of stations with reference co-ordinates, a reference code and a description should be given for that location. When possible, the geographical co-ordinates should be entered in the description. This description however can also be used when samples are taken in zones that can not be described with co-ord-

ordinates.

ROSCOP parameters can only be selected when all the sampling locations are entered. When a ROSCOP parameter is entered, no sampling locations can be added anymore.

Take care of saving the set before you return to the switchboard !



Lab : MUMM Cruise : BE2000/07 Progr. : BMM-MD

Please enter one form for each physical sample.

Sampling device: NI10

Sampling depth: -3 m (subsurface)

Remarks:

Sampling laboratory : MUMM

Number of these samples at one station: 3

Specify sampling locations by selecting stations in the list or by describing the stations :

Name	Latitude	Longitude
120	51° 11' 06" N	2° 42' 04" E
130	51° 16' 15" N	2° 54' 18" E
140	51° 19' 47" N	3° 03' 20" E
150_a	51° 24' 58" N	3° 24' 08" E
215_a	51° 16' 36" N	2° 36' 48" E
230	51° 18' 30" N	2° 51' 00" E
250	51° 31' 00" N	3° 19' 00" E
315	51° 19' 22" N	2° 27' 50" E
330_a	51° 26' 00" N	2° 48' 30" E
340	51° 29' 25" N	2° 58' 39" E

Selected stations :

Station	Latitude	Longitude
105	51° 11' 00" N	2° 28' 30" E
115_a	51° 09' 18" N	2° 36' 12" E
330_a	51° 26' 00" N	2° 48' 30" E

Select parameters to be measured and click 'Add' :

ROSCOP Parameters

Selected parameter categories :

B08	Phytoplankton
H25	Nitrite
P01	Suspended matter

When a sampling location is specified and 'Add' is pushed, a record is inserted in the temporary table 'Input_planned_samples'. The IDsample, an integer, is calculated. The unique key of the table 'Planned_sample'is the combination of the Service identifier, IDSample and the cruise code. The highest IDSample for the active service is looked for in both tables, 'Input_planned_samples' and 'Planned_samples', to ensure a unique ID.

When a ROSCOP Parameters is selected and 'Add' is pushed, a record is entered in the temporary table 'Input_relation_sample-categories'. The use of temporary tables allows the 'Reset'-button to delete all records in those tables without affecting previously entered sample sets

By clicking the 'Save set' button, the temporary tables are emptied in the final sample planning tables.

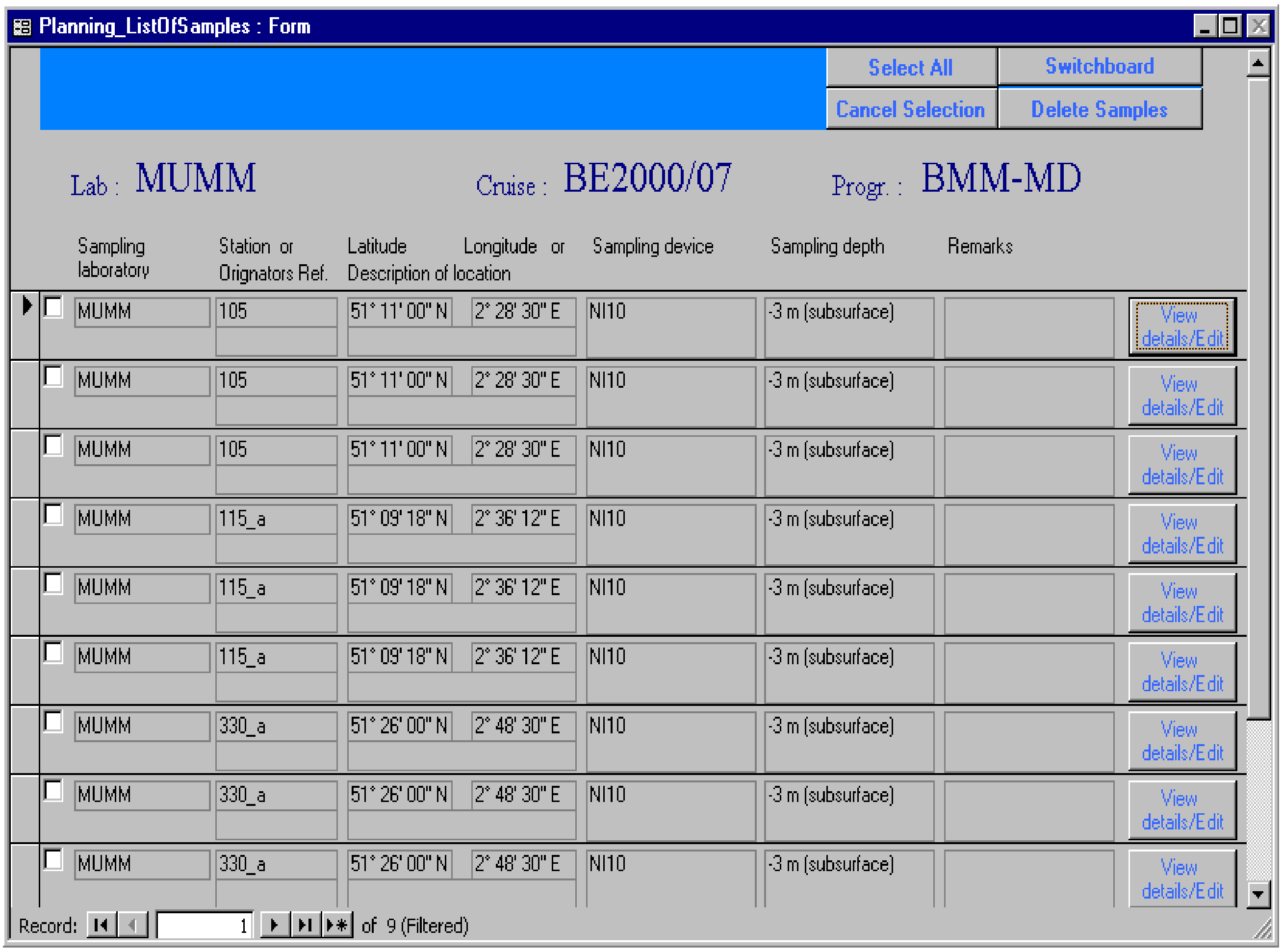
2.4.Load a standard set

Once a set has been entered for a previous campaign, it can be loaded again for a new cruise. The table name containing the set should be selected.

The records are directly added to the tables 'Planned_samples' and 'Relation_sample-categories'.

2.5.View and edit planned samples

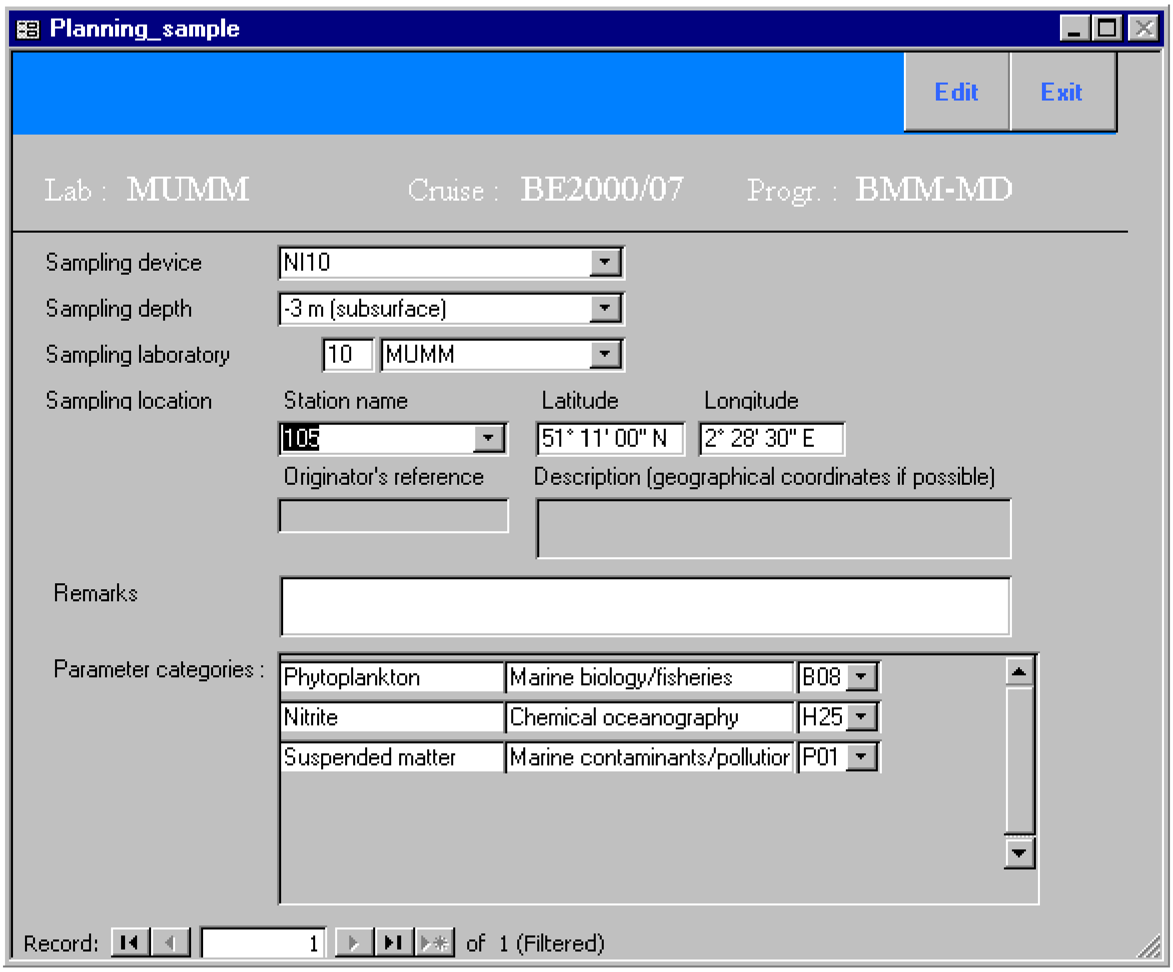
The list of planned samples for the cruise, laboratory and programme as specified in the login, is shown. Here, one has the possibility to delete planned samples, to view details and to edit information about a sample. Edits are only possible when viewing one sample at the time with all details by clicking on 'View details/Edit'. Details are the specified ROSCOP parameters to be measured on the planned samples. Undo is not possible while editing !



Lab : MUMM Cruise : BE2000/07 Progr. : BMM-MD

	Sampling laboratory	Station or Orignators Ref.	Latitude	Longitude or Description of location	Sampling device	Sampling depth	Remarks
□	MUMM	105	51° 11' 00" N	2° 28' 30" E	NI10	-3 m (subsurface)	
□	MUMM	105	51° 11' 00" N	2° 28' 30" E	NI10	-3 m (subsurface)	
□	MUMM	105	51° 11' 00" N	2° 28' 30" E	NI10	-3 m (subsurface)	
□	MUMM	115_a	51° 09' 18" N	2° 36' 12" E	NI10	-3 m (subsurface)	
□	MUMM	115_a	51° 09' 18" N	2° 36' 12" E	NI10	-3 m (subsurface)	
□	MUMM	115_a	51° 09' 18" N	2° 36' 12" E	NI10	-3 m (subsurface)	
□	MUMM	330_a	51° 26' 00" N	2° 48' 30" E	NI10	-3 m (subsurface)	
□	MUMM	330_a	51° 26' 00" N	2° 48' 30" E	NI10	-3 m (subsurface)	
□	MUMM	330_a	51° 26' 00" N	2° 48' 30" E	NI10	-3 m (subsurface)	

Record: 1 of 9 (Filtered)



Lab : MUMM Cruise : BE2000/07 Progr. : BMM-MD

Sampling device: NI10

Sampling depth: -3 m (subsurface)

Sampling laboratory: 10 MUMM

Sampling location: Station name: 105; Latitude: 51° 11' 00" N; Longitude: 2° 28' 30" E

Originator's reference

Description (geographical coordinates if possible)

Remarks

Parameter categories :

Phytoplankton	Marine biology/fisheries	B08
Nitrite	Chemical oceanography	H25
Suspended matter	Marine contaminants/pollutior	P01

Record: 1 of 1 (Filtered)

2.6. Preview and print reports of cruise planning

Two different reports can be previewed and printed. One contains the planned samples with all details (f.e. ROSCOP parameters) listed by sampling location. The other report lists the number of samples for each sampling device at each sampling location. A printed version of this report should be sent to the chief scientist for inclusion in the Belgica campaign programme.

2.7. Save as standard set

The entered samples can be saved as a standard set when these samples at these sampling stations are likely to be taken again in another cruise. The name for this set should start with 'Set' followed by the laboratory code and eventually a sequence number.

The information on samples together with the relations with ROSCOP parameters are stored in one table with the specified name. As such the number of records will be higher than the number of samples. When a new version of the application is sent to the laboratories, the standard sets should be included.

2.8.Change login

When samples for a next cruise or another programme are to be entered, the login should be changed.

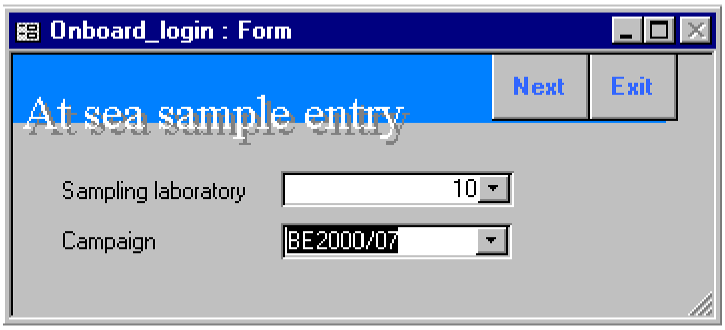
3.Application Sample Merging

To merge the sample lists for use onboard the form 'MergeSampleList' should be opened. This form is only available at MUMM-Ostend. The tables with information on planned samples from the different laboratories participating in one campaign should be merged in one table. The complete list of samples will be installed onboard of the Belgica before departure.

The cruise for which sample lists have to be merged should be specified as well as the Access database from which the samples should be loaded. The Acces database name should contain the laboratory code.
With this information, automatically the tables 'Planned_samples' and 'Relation_sample-categories' will be appended to the tables 'Onboard_samples' and 'Onboard_rel_sample-cat'.

4.Application SampleRegistration

To start this application, the form 'Onboard_login' should be openend.



Sampling laboratory: 10
Campaign: BE2000/07

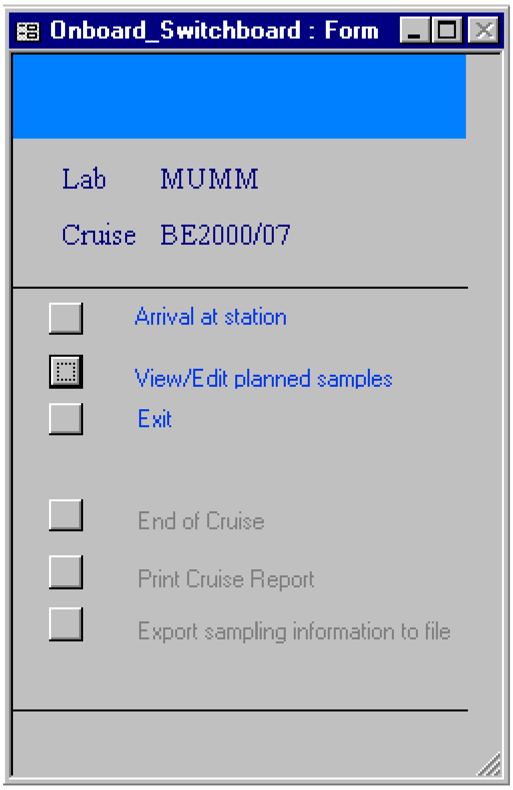
SampleMerging should have taken place before this application can work.
The modules are based on the tables 'Onboard_samples' and 'Onboard_rel_sample-cat'.

4.1.Login

The laboratory that will execute the sampling and the cruise have to be specified. This will result in a list of samples that should be taken by the service logged on.

4.2.Switchboard

Lists all the menu items for the registration of the sampling event. After executing a menu item, one should always return to the Switchboard by clicking the appropriate button.


Onboard_Switchboard : Form
□ Arrival at station
□ View/Edit planned samples
□ Exit
□ End of Cruise
□ Print Cruise Report
□ Export sampling information to file

Lab MUMM

Cruise BE2000/07

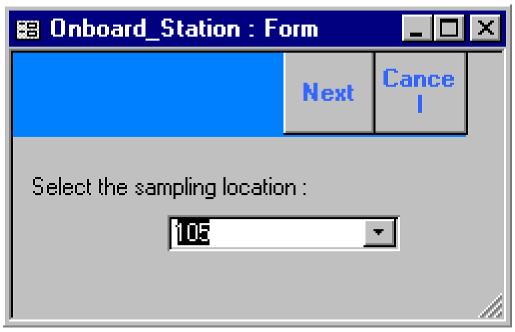
4.3.Arrival at station

When the ship arrives at a sampling location, this menu item should be selected.

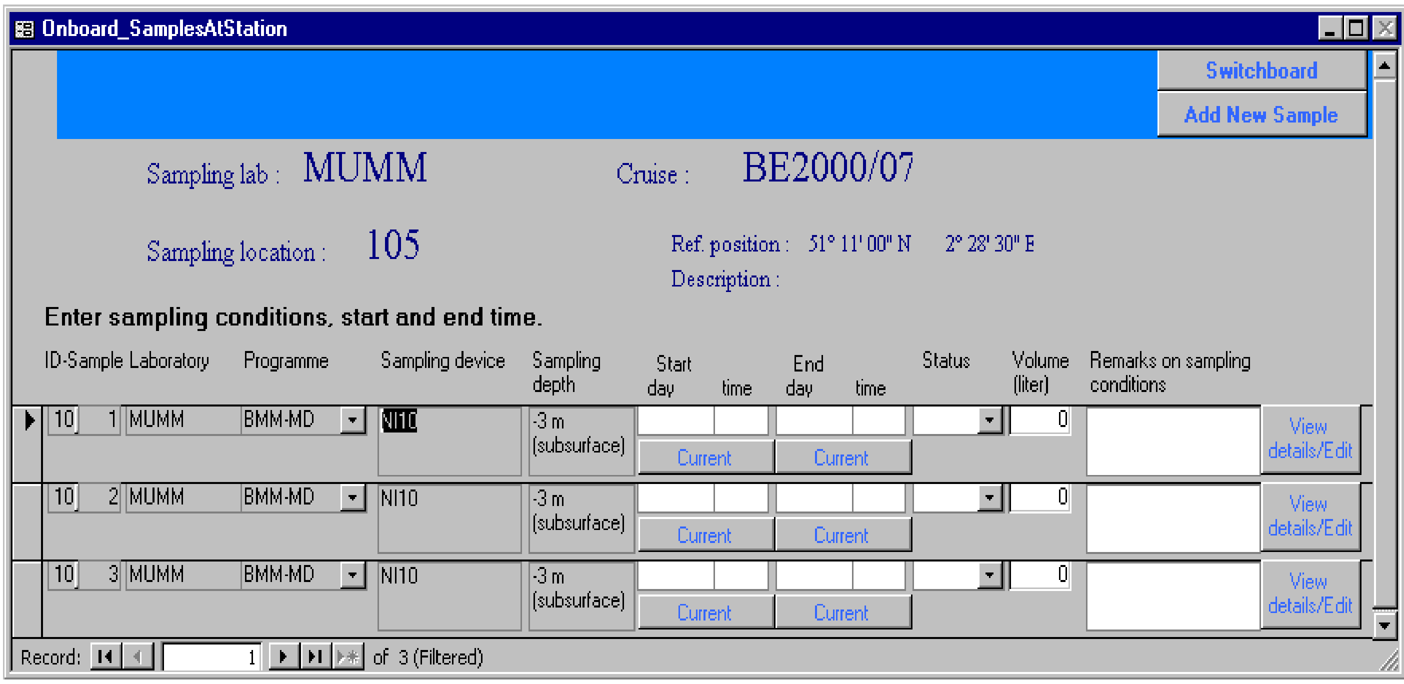
After selecting the sampling location, a list of samples that should be taken there is given. By simply clicking on the buttons ‘current day/time’ the start and end time of sampling can be entered. If however one clicks later, the entered time can be manually edited. If the sampling concerns a point sample e.g. a Niskin bottle or Van Veen grab sample, start time equals end time. The status of the sample should be filled as well as the approximate volume. Remarks on sampling conditions can be entered.

Edits are still possible by clicking ‘View details/Edit’, e.g. when another sampling device has been used than initially planned or when extra parameters will be measured on the sample.

When returning to the switchboard, all samples should be entered as ‘Available’ with start and end time recorded or as ‘Missing’.



Select the sampling location : 105



Sampling lab : MUMM

Cruise : BE2000/07

Sampling location : 105

Ref. position : 51° 11' 00" N 2° 28' 30" E

Description :

Enter sampling conditions, start and end time.

ID-Sample		Laboratory	Programme	Sampling device	Sampling depth	Start day	time	End day	time	Status	Volume (liter)	Remarks on sampling conditions
10	1	MUMM	BMM-MD	NI10	-3 m (subsurface)						0	
10	2	MUMM	BMM-MD	NI10	-3 m (subsurface)						0	
10	3	MUMM	BMM-MD	NI10	-3 m (subsurface)						0	

Record: 1 of 3 (Filtered)

Here tests should be developed when leaving a station (returning to the switchboard

If status = empty : 'Please indicate the status of all samples as 'available' or missing'.

If status available and start day/time missing : 'Please enter start sampling time'

If end day/time missing : 'Please enter end sampling time'.

4.4.View/Edit planned samples

All planned samples for all sampling locations are listed for the specified login: sampling laboratory and cruise.

4.5.End of Cruise

Not yet functional.

Tests should be developed to test whether all samples planned during that cruise (so for all sampling locations) are either missing oreither available with start and end time recorded.

4.6.Print Cruise Report

Not yet functional

A cruise report template still has to be made.

4.7.Export sampling information to file

Not yet functional

Not only information available in these tables but also the information from ODAS should be exported to a file.

5.Objects of the applications

5.1.Objects of the application SamplePlanning

Tables	
Main tables	
Planned_samples	Contains list of all planned samples for that cruise and laboratory.
Relation_sample-categories	Contains the relationship between planned samples and ROSCOP parameters for that cruise and laboratory.
Code tables	
Belgica_programmes	List of programme acronyms as used in the Belgica programme.
Campaign	List of Belgica campaigns
Sampling_depth	List of reference sampling depths
SamplingDevices	List of sampling devices available at MUMM Ostend
Services	List of laboratories
Station	List of stations with reference positions
Temporary tables	
Input_planned_samples	Temporary table to enter a sample set
Input_relation_sample-categories	Temporary table to enter the relationship with parameter categories of a sample set
Userowned tables	
Setxxxxxx e.g. SetMUMM1	Table saved by service to be able to load this set of samples again for a next campaign.

Forms	
Planning_login	To enter laboratory, cruise and programme for which samples are planned
Planning_Switchboard	Menu list
Planning_AddSample	To add one type of sample for a list of sampling locations
Planning_LoadSet	To load a set of previously saved samples for a new cruise
Planning_ListOfSamples	An overview of the planned samples with possibility to edit and delete, print the sample planning and save a set of samples as standard set.

Planning _Sample	To edit information or view details of a planned sample
Categories_subform	To view different parameter categories to be measured on one sample.

Reports	
PlannedSamplesCross	Report with the number of samples for each sampling device at each sampling location
SamplingProgramme	Full report of the planned samples ordened by sampling location together with indication of parameter categories

Queries	
StationDMS	Query to convert real coordinates into degrees, minuts, and seconds

5.2.Objects of the application SampleMerging

Extra Forms
MergeSampleList
AddSamplingDevice

5.3.Objects of the application SampleRegistration

Tables	
Main Tables	
Onboard_samples	Contains list of all planned samples for that cruise.
Onboard_rel_sample-cat	Contains the relationship between planned samples and ROSCOP parameters for that cruise.
Code tables	
Belgica_programmes	List of programme acronyms as used in the Belgica programme.
Campaign	List of Belgica campaigns
Sampling_depth	List of reference sampling depths
SamplingDevices	List of sampling devices available at MUMM Ostend
Services	List of laboratories
Station	List of stations with reference positions

Reports

Forms	
Onboard_login	To enter sampling laboratory and current cruise
Onboard_Switchboard	Menu list
Onboard_station	To select sampling location
Onboard_SamplesAtStation	To view and enter information of all samples to be taken at that station.
Onboard_Sample	To edit information or view details of a sample
Onboard_CategoriesSubform	
Onboard_ListOfSamples	To view and enter information of all samples for a certain login

Queries	
StationDMS	Query to convert real coordinates into degrees, minuts, and seconds

COLOPHON

This report was issued by MUMM in February 2003.

Its reference code is MOD code.

Status □ draft
□ final version
□ revised version of document
□ confidential

Available in □ English
□ Dutch
□ French

If you have any questions or wish to receive additional copies of this document, please send an e-mail to *GroupName@domain*, quoting the reference, or write to:

MUMM
100 Gulledelle
B–1200 Brussels
Belgium
Phone: +32 2 773 2111
Fax: +32 2 770 6972
http://www.mumm.ac.be/

MANAGEMENT UNIT OF THE
NORTH SEA MATHEMATICAL MODELS

MODELLING GROUP

The typefaces used in this document are Gudrun Zapf-von Hesse's *Carmina Medium* at 10/14 for body text, and Frederic Goudy's *Goudy Sans Medium* for headings and captions.

Annex 4 : Phases, Objectives, Tasks and Deliverables

1.Analysis Phase

1.1.Objectives

The analysis phase will use the results of the strategy phase as input. These results will be verified and worked out to have a correct, executable model that acts as a base for further development of the project.

1.2.Activities and Tasks

- Orientation of the Analysis Phase
  - Agree on the scope of the analysis
  - Organize a team and setup of a project team

- Detailed Analysis
  - Develop a function model based on the function hierarchy
  - Develop the entity model based on the entity/relationship model
  - Check between entities and functions
  - Identify all functions which require special attention because of their complexity or use.

- Create a Development Plan
  - Divide the application in work units
  - Estimate the resources necessary for the activities
  - Produce a phased plan

- Create an initial transition strategy
  - Deliverance and Acceptance plan
  - Training plan
  - Data conversion
  - Installation plan
  - Project Management and Quality Assurance

1.3.Deliverables

Analysis document containing the following important components :

- Function hierarchie
- Data model containing detailed entities and attributes
- Detailed function definition and ‘processing’ logic
- Project plan

2.Design Phase

2.1.Objectives

The design phase takes over the detailed requirements of the analysis phase and finds the best way to satisfy the needs, tries to reach and maintain the agreed service level, given the technical environment.

2.2.Activities and Tasks

- Orientation of the Design Phase
- Design of the database
- Definition of Audit Control and Security
- Prepare tests, system tests and acceptance criteria
- Design program modules and applications
- Review the project plan for the remainder of the project
- Project Management and Quality Assurance

2.3.Deliverables

Technical specifications containing :

- Overview of the system architecture
- Module design and specifications
- Documented design decisions
- Initial module/table matrix
- Lay-out of the most important screens, reports and menus
- Detailed database size and assumptions
- Additional specifications for control modules
- Procedures for control and integrity
- Procedures for backup/recovery/security
- Fall back options if needed
- Draft transition strategy
- Reviewed list of all program modules, their complexity and the necessary implementation tasks
- Reviewed plan for the remainder of the project
- Draft training plan
- Acceptation document

3.Build Phase

3.1.Objectives

During the build phase programs are written and tested, using the appropriate tools. These tools are dependant on the technical environment and the type of programs.

3.2.Activities and Tasks

- Orientation of Build Phase
- Program all modules
- Design and develop test data for linked modules
- Acceptation of the tested programs
- Review planning for the remainder of the project
- Project Management and Quality Control

3.3.Deliverables

- Reviewed program specifications and detailed design
- Test plan for all programs
- Expected results
- Tested and in CASE documented program modules, containing source and object code
- Unit and link test results
- System test results
- The tested system
- Error/bug list and their corrections
- Reviewed plan for the remainder of the project
- Acceptation document

Annex 5 : Overview of the timing estimates per type of activity

Module	Timing	Description	Remarks
CAM0000		Exploit Campaign Data	
CAM1000	5	Campaigns	
CAM1050	5	Sampling Occasions	
CAM1100	4	Water Value	
CNS0000		Consult Information	
CNS1000	0.5	Consult Data Themes	
CNS1050	0.5	Consult Quasimeme Group Codes	
CNS1100	0.5	Consult Quasimeme Method Codes	
CNS1150	1	Consult Sea Areas	
CNS1160	3.5	Retrieve information from ODAS	Informix link / BMM Ostend
EXP0000		Feed Other Systems	
EXP1000	10	Feed ICES	sir + selection screen + export
EXP2000	6	Feed EDMED	sir + selection screen + export
EXP3000	5	Feed EDMERP	sir + selection screen + export
EXP4000	6	Feed ROSCOP (may be not necessary)	sir + selection screen + export
EXP5000		Export to excel-format	
EXP5100	2	Station/Date/Salinity per given period	Export 1 to excel-format (csv)
EXP5200	2	Week overview	Export 2 to excel-format (csv)
EXP5300	1.5	Overview ICES codes + parameter	Export 3 to excel-format (csv)
EXP5400	1.5	Overview Stations	Export 4 to excel-format (csv)
EXP5500	1.5	Overview Parameters	Export 5 to excel-format (csv)
EXP5600	1.5	Overview Parameters II	Export 6 to excel-format (csv)
EXP5700	1.5	Parameter/Laboratory/Analytical method	Export 7 to excel-format (csv)
QRY0000		Exploit Sample Data	
QRY1000	25	Query for data	Complex (see Billarry/Discoverer)
IMP0000		Load Information	
IMP1000	3	Import Data Belgica	common layout format
IMP1100	4	Import Sample Data Belgica	
IMP1200	4	Import Result Data Belgica	
REF0000		Maintain Reference Information	
REF1000		General Information	
REF1010	0.5	Cities	
REF1020	0.5	Countries	
REF1030	0.5	Ecosystem	
REF1040	0.5	Position	
REF1050	1	Institute	
REF1060	1.5	Services/service coordinates	
REF1070	1.5	Persons/Person Groups	
REF1080	3	Datasets	
REF1090	4	Projects	
REF1100	2	Publications	
REF1110	0.5	Stations	

REF1120	0.5	Research Programmes	
REF1130	1.5	Mailing/Mailing Items	
REF2000		Methodology related information	
REF2010	0.5	Detection Limit	
REF2020	1	Platform	
REF2030	0.5	Preservation	
REF2040	0.5	Pretreatment	
REF2050	0.5	Separation	
REF2060	1	Sampling Gear	
REF2070	1	Sample Handling	
REF2080	3.5	Analysis Method	
REF3000		Parameter information	
REF3010	0.5	Matrix	
REF3020	1	Parameters	
REF3030	0.5	Category	
REF3040	0.5	Unit	
REF4000		Quality information	
REF4010	0.5	Quality Sampling Handling	
REF4020	1	Reference Material	
REF4030	0.5	Control Chart Reference	
REF4040	0.5	Control Chart	
REF4050	1	IC Exercise	
REF4060	0.5	IC Ex Result	
REP0000		Reporting	
REP1000	3	Parameter dictionary	
REP1050	3	Parameter dictionary with QA information	
REP1100	3	Parameters measured for service/project combination	
REP1150	3	Parameters measured : number of values reported to ICES	
REP1200	3	Parameter versus Lao for every year	
REP1250	3	Sampling frequency of a station	
REP1300	3	Cross matrix year/parameter	
REP1350	3	Position of stations	
REP1400	3	Publications	
REP1450	3	Inventory of data per project/labo/year	
REP1500	3	Onboard registrated samples versus results	
REP1550		Cruise report	Has not been estimated as not yet clear if needed
Total	160		

UNIVERSITE DE LIEGE
Faculté des Sciences
Département de Géomatique

Bathymetric digital elevation model with water volume preservation

Spatial and temporal requests

Spatial data distribution, trend lines and surfaces, interpolation methods and validation

Research contract MN/DD/61

SCIENTIFIC REPORT

YVES CORNET FABRICE MULLER

Leadership : JEAN-PAUL DONNAY

Scientific and technical report

N° 25

Liège (B), 01-2001

S.U.R.F.A.C.E.S.
Service Universitaire de Recherches Fondamentales et Appliquées
en Cartographie et en Études Spatiales

PART I

Bathymetric digital elevation model with water volume preservation

1.1 The data and their implications

The available data consist in 507785 bathymetric sounding points stored in several ASCII files. Each record of these files includes 4 fields that give the position of the points in a UTM referential grid, the depth and an alphanumeric filed that gives the campaign identification. To the total, 23 areas have been surveyed to cover the Belgian continental platform between 1992 and 1999. Each ASCII file corresponds to one surveyed area. For some of these areas, several surveys have been achieved at different times. A little overlay zone exists between the neighboring areas, but sometimes, there is no overlay and a little empty zone exists.

The different survey times for the different areas impose two previous statements to all discussion of the interpolation and integration techniques to adopt.

The first concerns the reduction of the tidal and surge effects. Insofar as these effects have been corrected and that all the measures are comparable from that point of view, we can evoke the second statement that relates to the sedimentary dynamics that, during a so long time period, can produce relatively important modifications of the sea bottom topography. This statement applies more especially to the possible doorstep effect between two neighboring areas. It can also apply to a same geographical area surveyed at different moments (RAAN95 and RAAN99, for instance).

In the first case, one can achieve two separated interpolations on each of the files and compute the mean of the two results or adopt a interpolation method that takes in consideration all the points together and that integrates the two data sets to restore the surface. Kriging is than foreseeable. A nugget effect could be noted because of the temporal depth variation in the common zones.

In the second case, one must consider the last country solely or must take in account the two countries and to restore the surface while also achieving separated two interpolations or while adopting a technique that permits to consider the two countries at the same time.

In the following discussions and tests, we consider that all measures are comparable (data corrected of the tidal and surge effects). For the areas surveyed at two different moments, we have exploited the last one. Finally, we consider that the interpolation method must be applied independently on each area and that the surface of the common area must be computed averaging the results of the two interpolations in the common area (weighted average according to the distance to the limits of the common zone of the two neighboring areas)..

1.2 The problem to solve

The request of the UGMM consists in a production of a low spatial resolution bathymetric DEM (~250 m) of the Belgian continental shelf area from the soundings data set presented above. However the measures have generally been achieved with a larger density that the one of 1 point all 250^2 m^2. Furthermore, this density is higher along the ship track than according to the perpendicular direction. The production of this DEM will then consist in an integration/aggregation of the data. This integration/aggregation process should be achieved while respecting the constraint of water volume preservation.

In the following discussions and tests, this constraint has been verified as considering a reference volume calculated from the TIN (Delaunay triangulation) constructed on the sounding data set. However, if this type of representation can be considered as the most faithful to the collected data, other solutions could be analyzed because it could not be in relation to the reality of the seafloor topography. Nevertheless, the weak variability of the slopes of this seafloor should allow the adoption of such a reference with regard to the relatively high density of the data.

1.3 Which solution verifying the volume preservation constraint can be considered ?

From a theoretical point of view, we can mention and several solutions. We explain briefly five among them.

The first method is based on geoprocessing techniques applied to compute of volumes from the TIN in a square polygonal net with the same dimension than the resolution of the wanted DEM, either 250 m on 250 m (figure 1). This calculation can be achieved while integrating in the creation of the TIN a softline vector layer defining the net and while calculating the position and depth of the intersections between this network and the TIN using the depth value of the TIN. It is then possible to calculate a mean depth value in each of these polygons from the division of the volume by their area. This solution keeps the volume perfectly in every polygonal cell. However, it is achieved in vector mode and needs the processing of the attributes tables that are relatively heavy considering the great number of triangles of the TIN in the whole area and the large number of polygonal cells of the network to consider. Otherwise, this solution is applicable insofar as only one TIN can be created on the full extend of the continental shelf. However, it is not recommended because of the temporal scattering of the measures on the common zones of several surveyed areas.

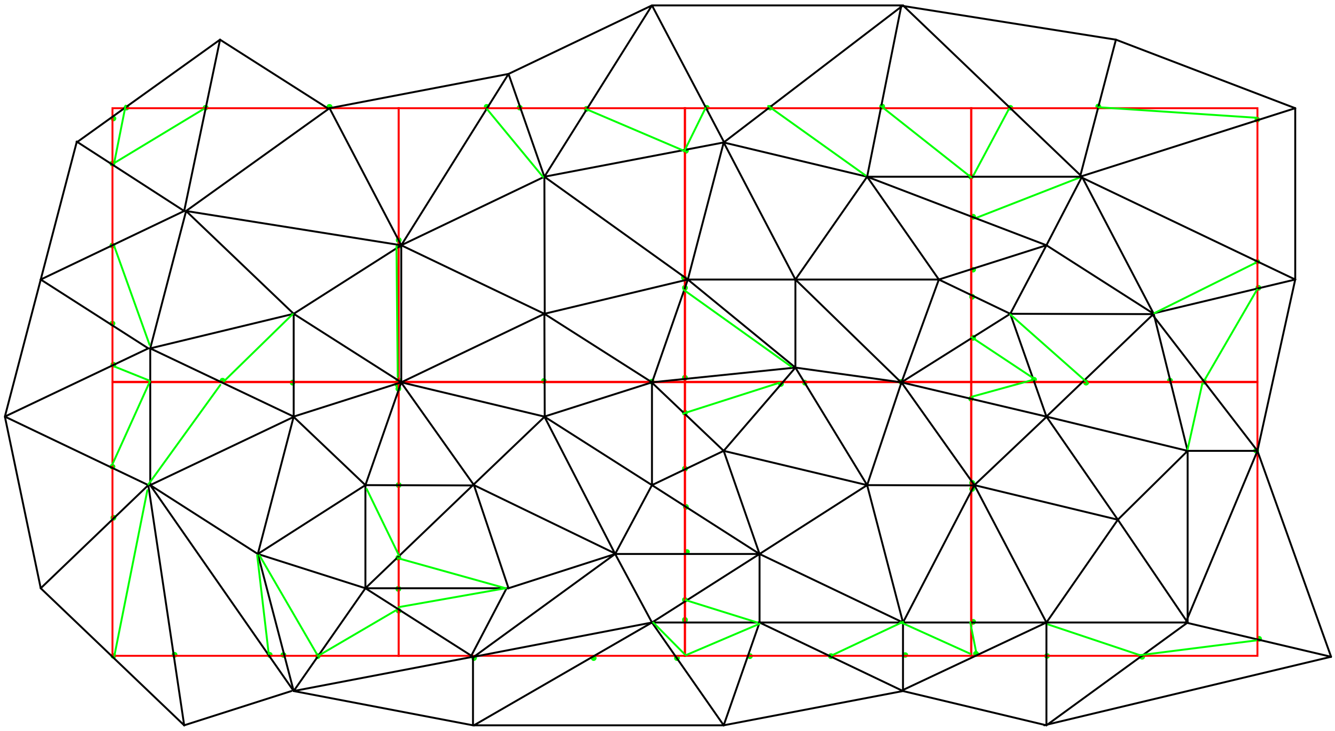


Figure 1 : TIN and softline square polygons cover (water volume and mean depth computation in each square cell).

The <u>second foreseeable method</u> consists in the transformation the punctual entities of the data set in zonal entities (polygons) to which one assigns the attribute(s) (depth, for instance) of the points of which they ensue. This transformation can be achieved while constructing the Thiessen tessellation that is the dual graph of the TIN (Delaunay triangulation). Then, the volume preservation can be verified while doing the same type of geoprocessing that the one proposed above for the first method suggested on the basis of the calculation of the intersections between the same net of polygonal square cells (250 m on 250 m). This method is less heavy than the first one. It is however hardly applicable considering the excessive elongation of the Thiessen polygons along the cross track direction. It could provide satisfactory results however because these tracks are oriented according to the direction of maximum slopes variability. Furthermore, such a method cannot be applied to the totality of the Belgian continental shelf because the Thiessen tessellation on basis of two sets of points not having the same significance in the common zones to several sounding sets cannot be constructed. This method can be illustrated by the figure 1 but in spite of considering the TIN one must consider its dual graph.

A <u>third method</u> consists in the rasterization of the points cover according to the low resolution corresponding to the one of the computed DEM. This rasterization must be performed computing the mean depth of all the soundings that fall in the considered grid cell. This method is however very sensitive to the possible variations of points density and to the orientation of the tracks. Therefore, the volume preservation won't be respected with the same level of precision for all cells.

Finally, the fourth method is based on a TIN to LATTICE conversion. The resolution of the output LATTICE is sufficiently small in relation to the surface of the triangles of the TIN. The second step of the method consists in the computation of a mean depth in a window whose measurements correspond to the one of a new low spatial resolution grid. Finally, the last step is a resampling in this grid at the low resolution of the desired DEM (1 pixel each n pixels, with n = ratio between the very high the low resolutions). This method is foreseeable considering the weak slopes variations of the seafloor. It is otherwise relatively simple and fast to put in work. We have opted this method that, however, induced a light variation of volume in relation to the TIN. As we demonstrate it below, this variation is nevertheless very weak if one exploits a first grid to very high resolution before the calculation of the focal average and the degradation. In order to assess this volume variation according to the adopted very high resolution, we did some tests on the data of 1995 relating to the zone of the *Vlakte van de Raan*. The figure 2 shows the principles of this method tested here after.

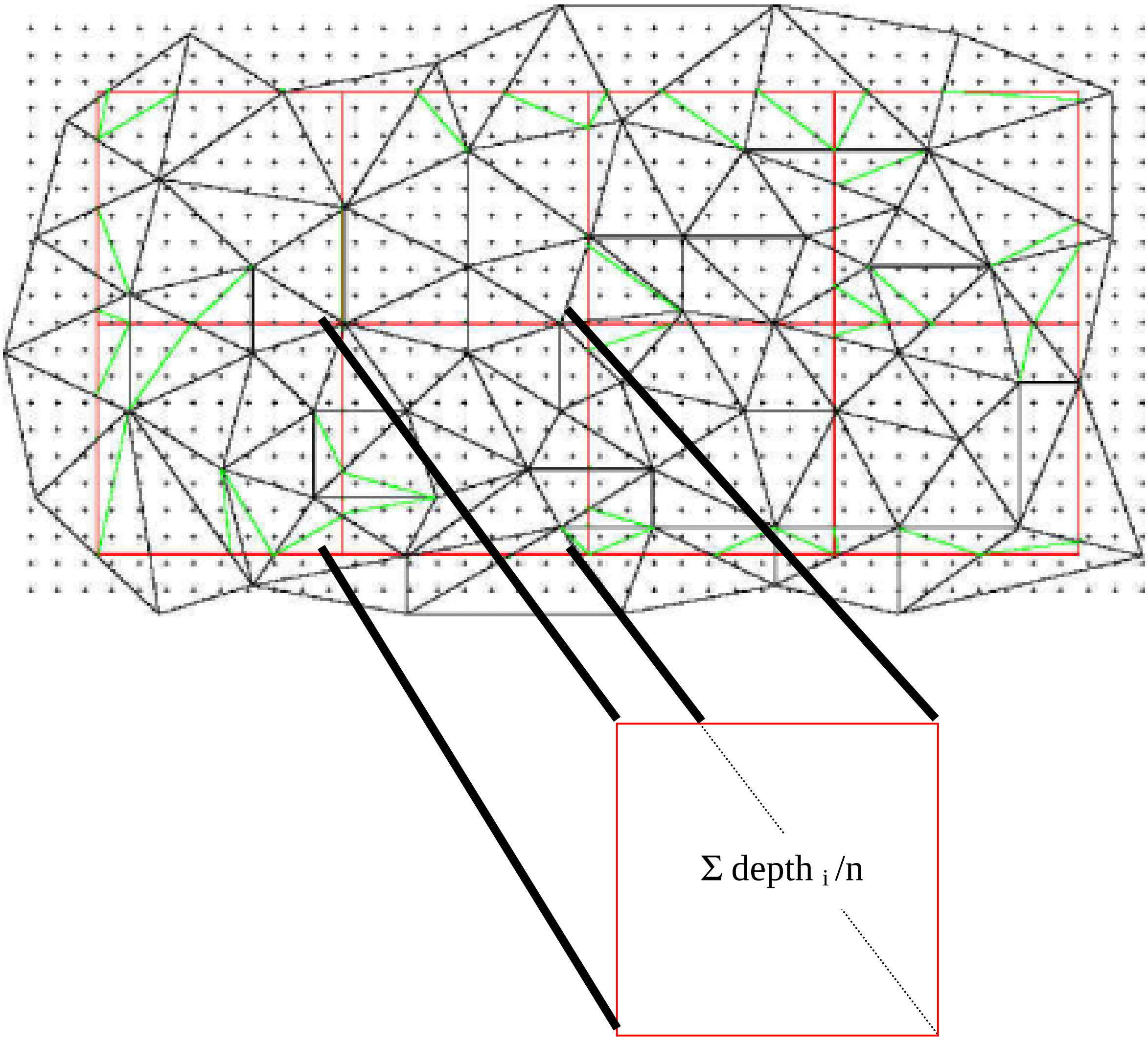


Figure 2: principles of the fourth method suggested to compute the bathymetric DEM

1.4 Tests done on the data set RAAN95

A TIN has been created from the sounding performed in 1995 on the zone of *Vlakte va de Raan* while exploiting the CREATETIN command of the ARC module of ArcInfo. The total number of points taken in account is 19757. Nevertheless, only a part of the zone has been exploited. It is delimited by a rectangular HARDCLIP polygon cover defined to perform an accurate estimation of the volume by computing the intersections between the TIN and this polygon and by interpolation of the depth at these new nodes.

The volume between the reference plane (depth = 0) and the TIN in this polygon has been calculated while exploiting the VOLUME command of the same ARC module. This volume has been chosen as a reference to assess the level of verification of the volume preservation constraint.

Then, we have applied a linear and a quintic interpolation process to compute the depth at each node of a high resolution LATTICE at four different resolution levels (~5, ~10, ~15 and ~20 m) while exploiting the TINLATTICE command of the ARC module of ArcInfo. The volume between the surface of the seafloor represented by these LATTICES and the water surface (depth = 0) has been then calculated exploiting the operator " + = (ADDITIVE SUMMATION) of the ArcInfo GRID module and multiplying the result by the surface corresponding to an element of the LATTICE. The focal mean has then been computed in a convolution window. Its dimensions are the same than the one of the desired DEM after geometric degradation performed by resampling of the pixels according to a constant step that corresponds to this resolution. The degradation after the focal mean doesn’t affect the volume preservation constraint.

```
Arcinfo processing : creation of the TIN
CREATETIN t-raan95-3 # # # #
COVER c-raan95 point depth 1
COVER z-zonecalcvol poly depth 7
End
&return
Arcinfo processing : TIN - LATTICE conversion
TINLATTICE t-raan95-3 g-raan95tll20 linear 1 float
&return
Arcinfo processing : volume computation from the TIN
VOLUMEt-raan95-3 0
Arcinfo processing : volume computaion from the LATTICE
GRID
DOCELL
Var += g-raan95tll20
End
Show var
```

The results presented on table I and figure 2 show the negligible magnitude order of the volume variation between the TIN and the LATTICES created for the tested resolutions of ~5, ~10, ~15 and ~20 m. The magnitude order of the volume variation (overestimate) obtained by the quintic interpolation process is more than 5 times bigger than the one obtained by the linear interpolation. The volume variation of the quintic interpolation process is smaller than

0,5 % of the TIN reference volume. Nevertheless, the use of the linear interpolation is suggested.

With regard to the resolution to choose, a compromise between the constraints imposed by the total area of the zone to analyze, the computational time, the memory management and the minimal dimension of the triangles (density of the data) must be found. Considering the negligible magnitude order of volume variation obtained by the linear interpolation for the 4 tested resolutions, the resolution of 10 m is probably ideal. This one should also be exploited for the other zones if the data densities are comparable to the one of the *Vlakte van de Raan* 1995.

Figure 3 : Relative volume variations computed by the different types of interpolation process and to different resolutions.

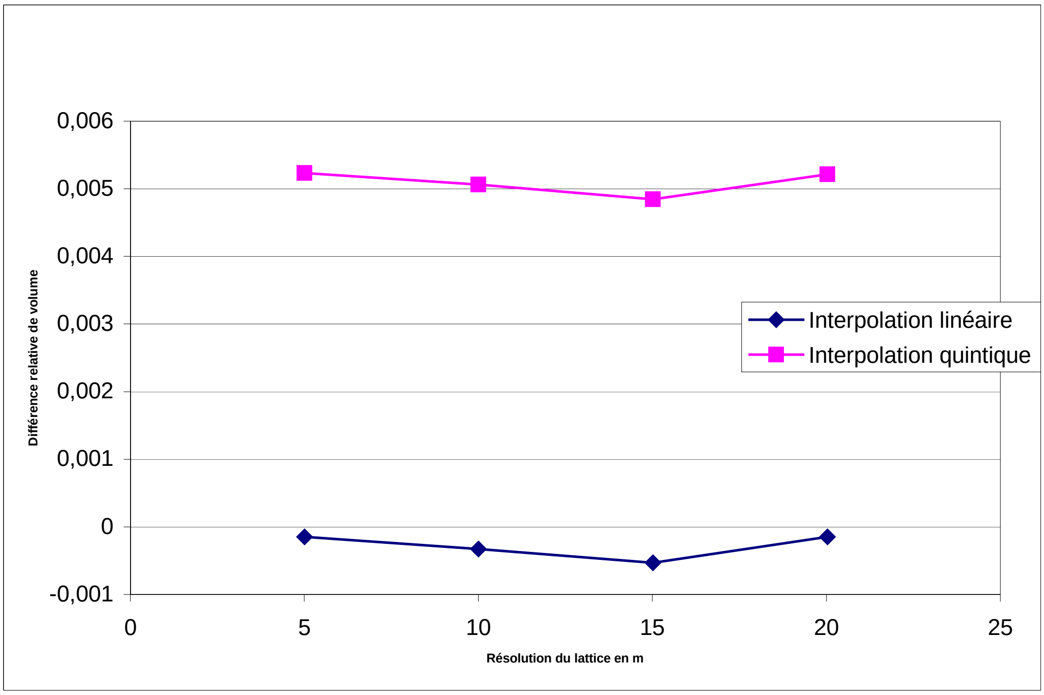


Table I : Results the fourth method of bathymetric soundings integration / aggregation on the area of Vlakte van de Raan 1995.

Bathymetric sounding data integration/aggregation with volumes preservation contraint		
Characteristics of the *TIN* t-raan95-3 :		
Xmin : 500000 m	Ymin: 5698500 m	depth min: 0 m
Xmax : 517500 m	Ymax : 5704500 m	depth max: 27,3 m
Number of nodes : 10742	Number of triangles : 21061	
Covered area : 1035 10^5 m^3		
Volume computed by the VOLUME command on the *TIN* t-raan95-3 :	11859,88864 10^5 m^3	

Volumes computed using the ADDITIVE SUMMATION operator on the *LATTICE* created from the TIN				
Nominal resolution (m)	5	10	15	20
ArcInfo resolution (m)	5,001	10,006	15,013	20,035
Colums number	3450	1725	1150	862
X extension	17250	17250	17250	17250
Computed resolution (m)	5,0014497	10,0058005	15,0130548	20,0348432
TINLATTICE LINEAR INTERPOLATION (cumulated depth in m)	47404984,2	11842256,3	5259095,66	2954232,46
TINLATTICE LINEAR INTERPOLATION (volumes in m^3)	1185811932	1185599838	1185357118	1185813968
Volume preservation	-0,00014918	-0,00032802	-0,00053267	-0,00014747
TINLATTICE QUINTIC INTERPOLATION (cumulated depth in m)	47660076,7	11906133,7	5287391,71	2970077,75
TINLATTICE QUINTIC INTERPOLATION (volumes in m^3)	1192192943	1191994991	1191734816	1192174187
Volume preservation (ratio)	0,00523114	0,00506424	0,00484486	0,00521533

1.5 Geometric degradation of the *LATTICE*

To perform the geometric degradation of the very high resolution LATTICE without altering the volumes, the following procedure must be exploited. The first stage consists in a computation of a focal sum of the depths using the FOCAL SUM function of the ArcInfo GRID module. Then, the computed mean values must be divided by the surface of the window chosen to achieve the focal addition operation. Finally, a third step consists in the geometric degradation by sampling of one element each *N* columns and each *N* lines. *N* is equal to the dimension in pixels of the window chosen to compute the focal addition.

This operation must be achieved after the mosaicing of the results of the linear interpolation achieved individually on the data of every area. To avoid the possible doorstep effect between two neighboring areas, the average of the interpolations done from both dataset can be calculated in the common zone. The depth in the common areas can also be computed using a weighted average method adopting for each of both measure a weight that is inversely proportional to the distance that separates the considered pixel of the boundary of the two areas to which it belongs.

The border of each file that must be mosaiced must be ignored to avoid the border effects of on the mean calculation. The different steps of the suggested method are illustrate by the test area example of RAAN95 on figures 4 to 9.

The full DEM computation is documented by the metadata reported in attachment 1.

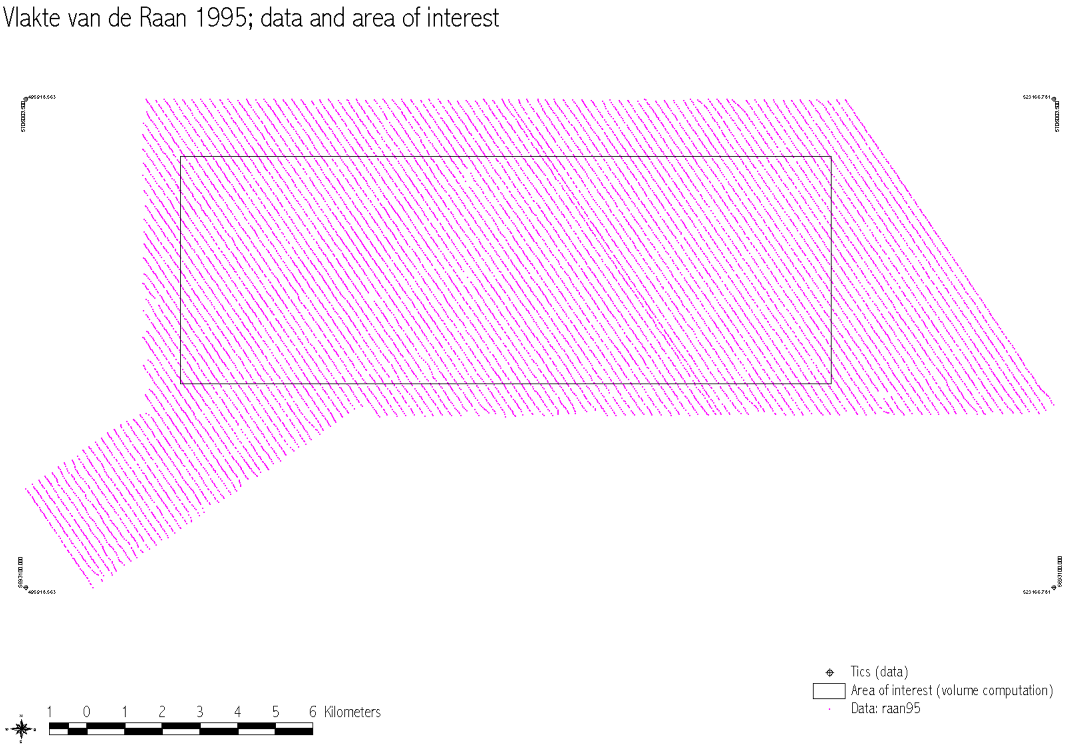

Vlakte van de Raan 1995; data and area of interest
495918.563
523166.781
495918.563
5693100.000
523166.781
5693100.000
1 0 1 2 3 4 5 6 Kilometers
Tics (data)
Area of interest (volume computation)
Data: raan95

Figure 4 : Data at the boundary test area polygon.

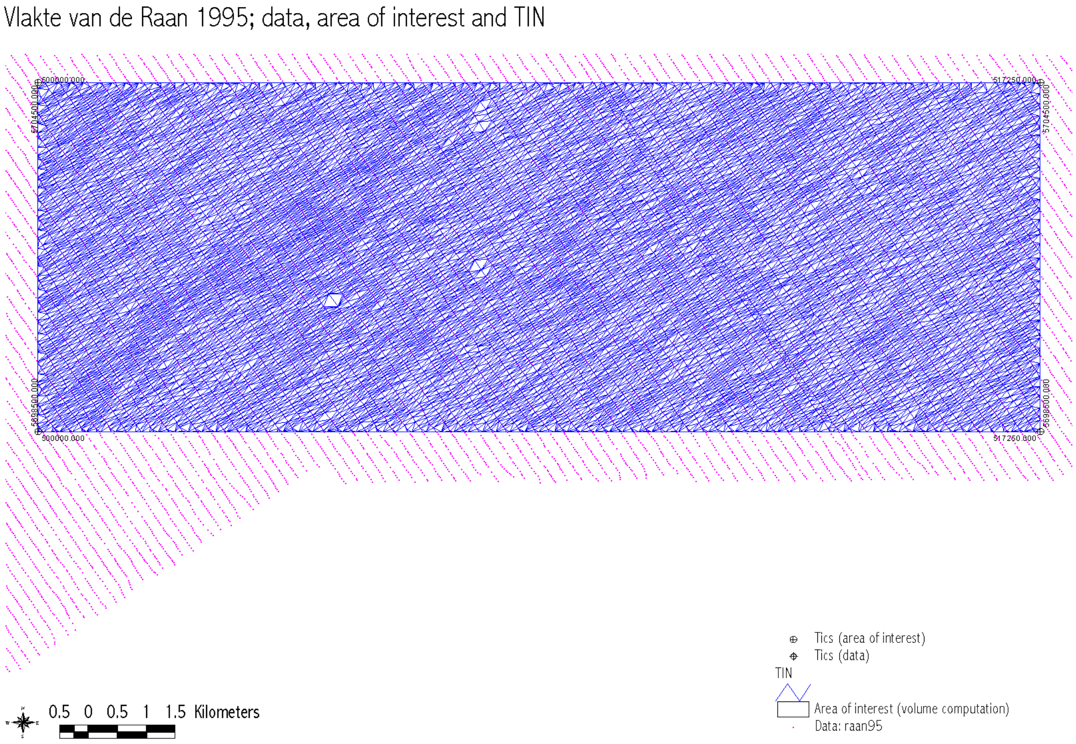


Figure 5 : Data and TIN created in the boundary test area polygon.

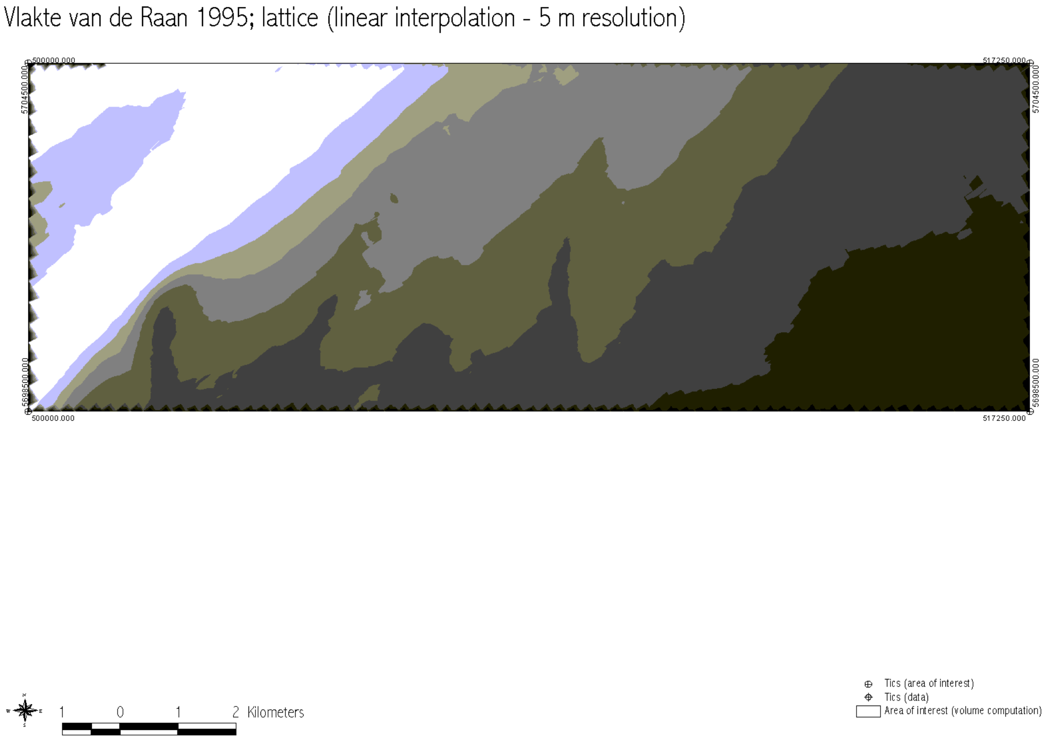

Vlakte van de Raan 1995; lattice (linear interpolation - 5 m resolution)
500000.000
517250.000
5704500.000
5698500.000
1 0 1 2 Kilometers
Tics (area of interest)
Tics (data)
Area of interest (volume computation)

Figure 6 : LATTICE computed using a linear interpolation method with a 5 m resolution.

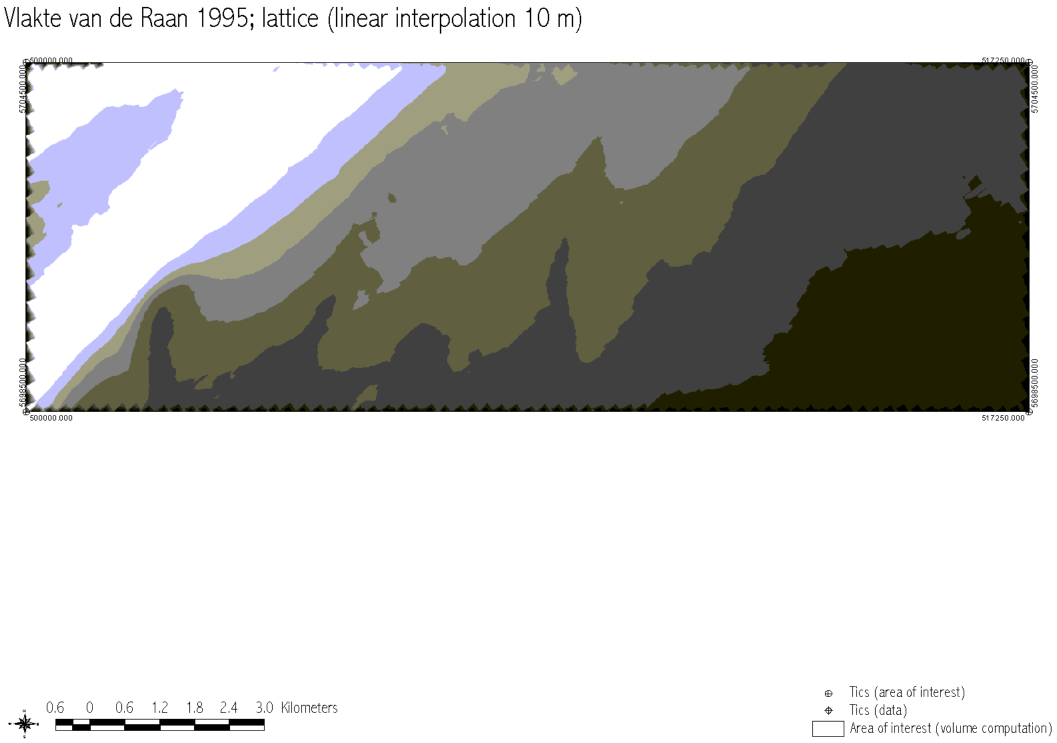

Vlakte van de Raan 1995; lattice (linear interpolation 10 m)
500000.000
517250.000
5704500.000
5698500.000
0.6 0 0.6 1.2 1.8 2.4 3.0 Kilometers
Tics (area of interest)
Tics (data)
Area of interest (volume computation)

Figure 7 : LATTICE computed using a linear interpolation method with a 10 m resolution.

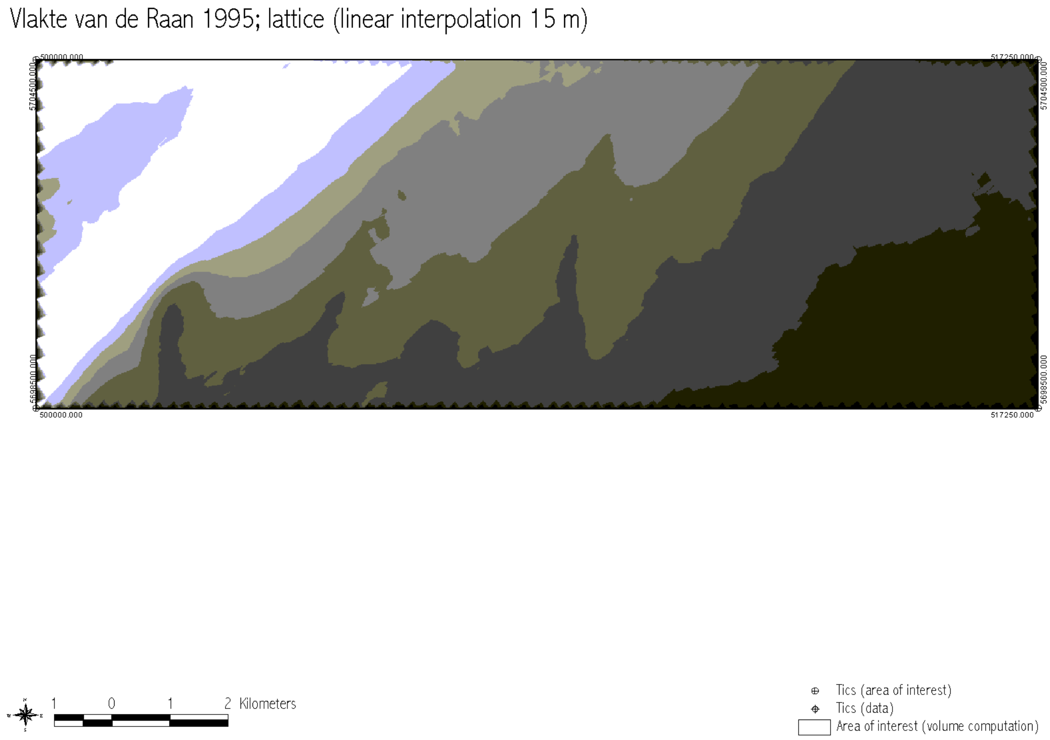

Vlakte van de Raan 1995; lattice (linear interpolation 15 m)
500000.000
517250.000
5704500.000
5698500.000
1 0 1 2 Kilometers
Tics (area of interest)
Tics (data)
Area of interest (volume computation)

Figure 8 : LATTICE computed using a linear interpolation method with a 15 m resolution.

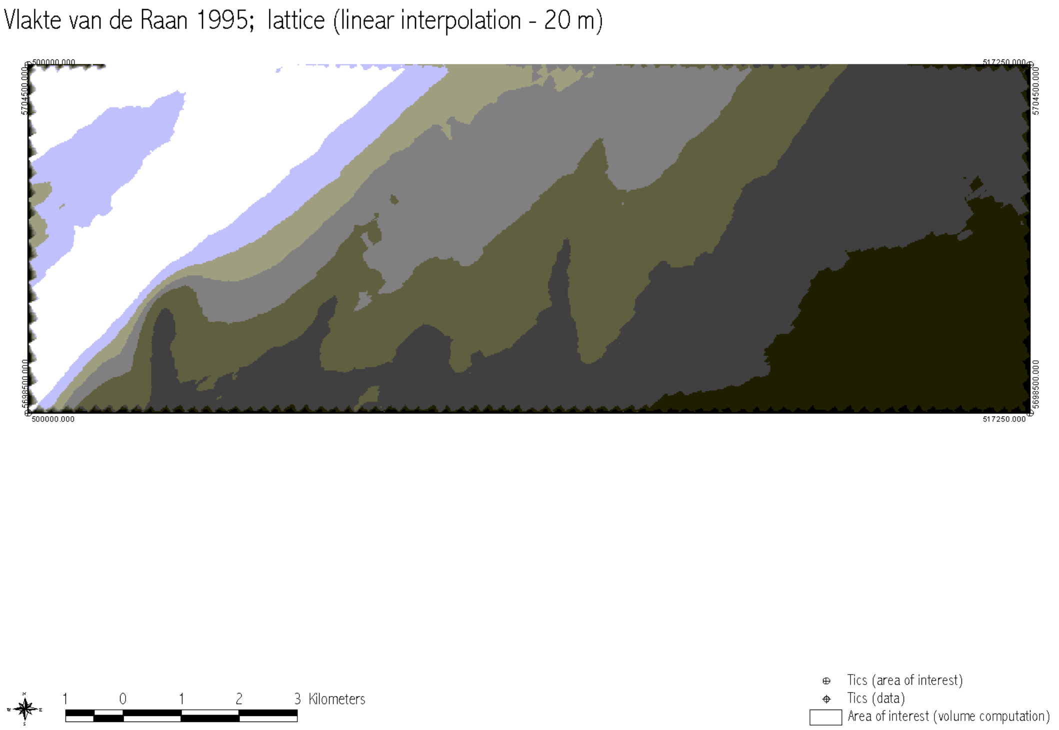

Vlakte van de Raan 1995; lattice (linear interpolation - 20 m)
500000.000
517250.000
5704500.000
5698500.000
1 0 1 2 3 Kilometers
Tics (area of interest)
Tics (data)
Area of interest (volume computation)

Figure 9 : LATTICE computed using a linear interpolation method with a 20 m resolution.

PART II

Spatial and temporal requests

1 Theoretical discussion

1.1 Introduction

The spatio-temporal requests in vector geographical databases, in general, and in the IDOD data base, in particular, can be regrouped in three main types. From these three types of requests, a solution to the main questioning that can be considered via the Web, for example, should be found. It is this first aspect of the requests that will be discussed in the § 1.2.

Only, the requests on a raster geographical information layer (grid or lattice) or on a TIN require other methods. We will present some particular examples in the § 1.3 of the following discussion.

Finally, in the § 1.4, we will approach some more complex request examples that are as with difficulty foreseeable via the Web, because they call on treatments that combine in input vector and also raster layers. Some necessary treatment examples to provide an answer to these complex requests must often be achieved in raster mode. These types of treatments are generally a lot heavier to undertake.

The different discussed cases are inspired of the note that has been transmitted to us by S. Scory and S. Jans the 4/10/2000.

1.2 Requests in a vector geographical data base

1.2.1 Simple request on the thematic attributes

The first kind of request, the simple request on the thematic attributes, is the simple request that is not constrained by spatial or topological relationships between objects. The answer to this kind of request consists in a research in the attributes tables of the interrogated database. It is based on the verification of a set of conditions that is formulated on basis of the values of the specific fields from this database. One can give an answer to this kind of request regardless of all geographical treatment and it can be achieved therefore without calling on the available GIS tools under ArcInfo or ArcView, for instance.

Thus, as we will develop it in a synthesis paragraph on the strategy to adopt to answer the requests, this statement will probably allow to simplify the geographical treatment insofar as the tables associated to the layers of geographical information could be simplified by a previous selection thanks to a questioning and a selection using the of the traditional management tools of databases as Access or Oracle, for instance.

Insofar as the time is defines by one or several fields of the database, the temporal requests can be considered as a simple request carrying on the thematic attributes.

1.2.2 Simple request on the position

The second type of request is the simple request on the position. The selection of the elements of the database is constrained by one or several spatial conditions. An answer to these requests can also be given while doing a set of tests. These test have a topological character: they analyse the relative position of the objects of the database (points in the present state of the development of the data base) that must be selected in relation to target objects (points, lines or polygons) whose position can be defined in a vector mode in the considered geographical space. The vector layers containing these target objects will generally have an exogenous origin and it will be necessary to foresee a method permitting to introduce these target objects in this type of request.

Generally, to solve such a request, one can call on two types of treatments.

The first consists in an analysis of the position without requiring a particular geographical treatment. This treatment is discussed below as simple request on the position without requiring any geoprocessing to be solved.

The second type of treatment consists in a recombination of the objects and in recreation of new objects while achieving particular topological operations (intersection, union, etc.). This type of treatment is discussed below as simple request the position requiring geoprocessing to be solved.

1.2.2.1 Simple request on the position without geoprocessing

The first type of treatment is applied to solve a simple request on the position without using any geographical treatment or geoprocessing.

The geoprocessing can be defined as a spatial processing of the target objects cover and the source objects cover. This source objects cover is the layer in which some elements must be selected in function of the topological (spatial relationships between source and target objects) condition (criteria of relative position). Generally, a new cover combining the target and source objects layers must be created.

The criteria on the relative position of objects are different depending on whether the cover to which the request applies is a points or lines/polygons cover. In our case and in the present state of the IDOD database development, it is a points cover (temperature, salinity, heavy metals concentration, measurement points). Nevertheless, we report in table I the whole of the possible combinations and operations that is permitted according to the considered types of geographical objects.

Table I

A> **B∨**	**points**	**lines**	**polygons**
points	5, 6	5, 6	2, 4, 5, 6
lines	5, 6	5, 6	2, 4, 5, 6
polygons	1, 3, 5, 6	1, 3, 5, 6	1, 2, 3, 4, 5, 6

Legend of the table

- A cover (points/lines/polygons) containing the objects that must be selected
- B cover containing the targets objects (points/lines/polygons)
- Authorized operations:
  - 1. Selection of the elements completely in…
  - 2. Selection of the elements that completely contain …
  - 3. Selection of the elements whose center is in…
  - 4. Selection of the elements that contain the center of…
  - 5. Selection of the elements that intersect…
  - 6. Selection of the elements that are to a distance of…

1.2.2.2 Simple request on the position with geoprocessing

The second type of treatment permits to solve a simple request on the relative position of the objects using geoprocessing. This type of treatment is more complex than the former one. It takes in consideration the different topological operations reported here after.

- Dissolution (DISSOLVE): regrouping of neighbouring (having a common border) spatial entities (polygons) having the same code for a given thematic attribute.
- Fusion (MERGE - APPEND): fusion of several vector covers by which, only, the common attributes are kept in the attributes table of the new created cover.
- Clipping (CLIP): creation of a new points, lines or polygons layer resulting from the intersection between two geographical information layers. One of the two layers must be a polygons cover. The attributes of the layer cut by the polygons contained in the second layer are not altered. If the cut layer is a points cover, this operation can be assimilated to the selection corresponding to some treatments discussed in the § 1.2.2.1.
- Intersection (INTERSECT): intersecting two vector layers of which one is a polygons layer, a new cover of can be created. It can be a points, lines or polygons cover depending on whether the other theme is a cover of points, lines or polygons. The attributes table of the cover resulting from this topological operation contains the attributes of the two provided covers. If the intersected cover is a points cover, then this operation can also be assimilated to one of the operations discussed in the § 1.2.2.1.
- Union (UNION): union of two polygons covers whose result is a polygons cover whose attributes come from the two input covers and the covered area corresponds to the one covered by the whole of the two geographical input covers.
- Assignment by location (ASSIGN DATA BY LOCATION -): this is a spatial join in which there is an assignment to one the two combined covers (points, lines or polygons covers) of the attributes of the second cover that must be a polygons cover according to the belonging of the entities of the first layer to the entities of the second one. If the first cover to which one assigns the code of second one according to the relative position of their objects is a points cover, this operation can also be assimilated to some of the operations reported in the table I of the § 1.2.2.1.

1.2.3 Complex request on position and thematic attributes

While combining the 3 types of request discussed higher (simple request on the thematic attributes, simple request on the position without geoprocessing and simple request on the position with geoprocessing), it is possible to consider a type of request that we qualify “complex” and that allows to answer the majority of the request in vector geographical databases that are indifferently based on criteria of position and thematic attributes.

Thus, among the examples of requests discussed during the meeting of the 4/10/2000, the following cases can be solved thanks to the methods presented before.

- **Salinity greater** than a **given value** during **the December - June periods** and that are **less than 10 km far from the coasts**.
- **Salinity greater** than a **given value** during the **wintry periods** and that are **less than 10 km far from the coasts** between **January 1st, 1990 and November 30, 1996** (under the condition to define some what consists a **wintry period** - cfr. § 1.2.4).
- On a map that gives the distribution of the measurement points of a physicochemical parameter of water or sediments (concentration of heavy metals, for instance), to put this data in relation with another spatial information (position of the wreckages, for instance), is it possible, for example, to recover in the parameter measurement stations the **measures of heavy metals concentrations** that are **greater** than a **given value** and localized not farther than **a given distance** to **some defined wreckages** (whose position can either to be defined by coordinates given by requiring people - **exogenous data** - or by a geographical information cover containing the wreckages present in the IDOD database - **endogenous data**).
- Possibility to choose (“to click on”) **randomly a location** on the whole surface of the zone that doesn't necessarily correspond to an element of **a point cover from the IDOD database** and, from this location, to **recover the geographical information present in the IDOD database bound to this place** (is the chosen location above a wreckage, or a pipeline or ... ?). This question has been formulated in this way during the meeting of the 4/10/2000. In fact, it is badly formulated. To answer, the geographical information about wreckage or pipeline must be present in the database. If it exists, it will be possible to answer provided that requiring people can introduce in the IDOD GIS coordinates system the desired location point - **exogenous data** (cfr. § 1.2.4). Otherwise, it agrees to define with what **accuracy** one defines the position " to the vertical of " (**point or buffer zone**), especially if one analyses the position in relation to a punctual or linear element. Finally, what is the information that must be selected if it is not about a punctual geographical data (surface of a continuous variable computed by interpolation or analysis of the neighborhood if the variable is not continuous).
- Possibility to superimpose the results of a double request: to determine (calculation and visualization) the area (polygon) that corresponds at the intersection between the **colonies of birds** and to a **temperature greater** than a **given value** provided by requiring people. If a spatial information layer on the temperatures exists (interpolation) for the desired period (that must be defined also), then it is possible to answer the questioning. Nevertheless, a notion is not clearly defined: what are the borders of the zone occupied by the colony of birds needs. To answer without ambiguousness this request, an answer must be given. Is it the polygon that circumscribes **the whole of the observed nests**? Is it a set of the 1 squared kilometer

area (in the referential UTM generally exploited for the cartography of zoogeographical data) **where at least one nest has been recognized**? Is it the polygon delimiting the zone **where the density of nests is greater than a given value**? Is it another zone?

1.2.4 Remarks

Remark 1: For both simple request on the position requiring geoprocessing or not, the achieved operations of selection can be considered on the totality of the input geographical dataset or on a part of these (previous selection in the tables, for instance).

Remark 2: How is it possible to take select some elements of the database by relative localization to other exogenous geographical elements? The offered possibilities are of two types depending on whether the layer that contains the objects in relation to which the objects are localized is present in the IDOD geographical database or whether requiring people must define their position. If the layer exists in the IDOD database (coastlines, for instance), and to let to requiring people the possibility to access to that in the formulation of the request is sufficient. If the layer doesn't exist (track of a measurements campaign achieved by requiring people, for instance), it is necessary to let to requiring people the possibility to define the elements of this layer in a simple format exploitable under ArcView (ASCII vector file, DXF file, Generate file or events table in ASCII format, for instance). The way whose elements can be encoded could be an interactive one, mouse click on basis of a map to define points, vertices and nodes, or could consist in the constitution of a list of co-ordinates encoded using the keyboard. The referential system in which the data must be encoded can be either a geographical referential system (longitude and latitude in decimal degrees, for instance), either a cartographic referential (Postel azimuthal equidistant). This referential can be let by choice of requiring people. The geographical referential system must be encouraged however because of its more universal character for a little experimented user. An operation of re-projection is then necessary to give answer to the request in the cartographic referential system chosen to present the cartographic documents of the project. In the case of the IDOD project, it is the Postel azimuthal equidistant cartographic projection system discussed during the meeting of the 13/9/2000. Let's recall that, contrary to the Lambert azimuthal equivalent projection system chosen in first approach, the Postel projection keeps the distances along the radial strikes from the projection center, whereas the Lambert azimuthal projection keeps the surfaces. For the calculations of density carrying on big surfaces, the azimuthal projection of Lambert presented a light advantage therefore. For all spatial on a big scale analysis (in a near neighborhood) close to the center of projection, the two projections are globally similar. For complementary information about these projections systems see appendix 2.

Remark 3: For all requests in the database according to seasonal fluctuations, summer or winter, it is advisable to define the criteria’s to recognize a summery situations or a wintry situations, for instance. This definition must be specified independently by temporal analysis of the data on several years. It could be the object of an ulterior specific study.

1.3 Request on vector coverages, TIN’s, grids and lattices

This fourth type of request should be envisaged to answer the more complex cases the following one of which, presented during our meeting of 4/10/2000, constitutes a beautiful example:

- From the bathymetric map, is it possible to define a location clicking on a point (encoding its geographic co-ordinates using the keyboard) and to have its depth?

As far as the bathymetry is defined in a mode raster (grid or lattice) in the IDOD database and in a resolution enough precise (*cfr.* report concerning the integration of bathymetric data), such an interrogation is possible by looking for to the nearest neighbour or by bilinear interpolation to get the depth at this location. The second solution is proposed under ArcInfo, for example, and is operated thanks to the LATTICESPOT command of the ARC module. Besides, the same type of interpolation is also exploited to realize profiles along rectilinear arcs by sampling points according to a constant step along these arcs thanks to the SURFACEPROFILE command of the ARCPLOT module of Arcinfo. The results of this last command can be stored in an “info” table and are so exploitable besides. It is possible to generalize the request to other types of objects than points.

With ArcView, such a procedure is more possible by harder from a raster file such grid or lattice. Nevertheless, there is a possibility to use the interactive “identify” tool of elements to obtain punctual information. This tool exploits also the bilinear interpolation method. Another solution consists in the search for a Z (depth or another continuous variable) value from a TIN. It is calculated from the plane equation characterizing the triangular facet to which belongs the introduced point.

It exists however one library of additional extensions for ArcView available on ESRI's Web site that allows to envisage solutions comparable to those that can be operated under ArcInfo. The extension " Profile Extractor 6.0 for 3D Analyst " developed by Ianko Tchoukanski (http://www.ian-ko.com / - 09 04 2000) allows to extract profiles from TIN’s or grids. An exploitable version under " Spatial Analyst " is also available (*cfr.* appendix 1).

1.4 Search for shorter path on base of a cost surface

Among the requests examples discussed during the meeting of 4/10/2000, one of them is not directly resolved by the methods proposed above. It is about a request which requires, to be resolved, the definition of a treatment allowing the search for the shortest path, that is the route which minimizes a cost function calculated from a cost surface (grid) which can be based, for example, on a bathymetric map or another grid resulting from the combination of several constraints. This request can be formulated in the following way.

- - Determine the shortest path to go from A to B (located by their geographic or cartographic Postel azimuthal equidistant co-ordinates) by avoiding all the depths lower than a given value at ebb tide as well as any zones crossed by pipelines or submarine cables.

Generally the search for the shortest path is made in spatial analysis at a spatial scale which allows to neglect the earth curvature. The calculation of the path is then made step by step in the cartographic referential by accumulation of distances corresponding to the resolution of a raster file representing the surface defining the movement cost (cost of pixels crossing). This distance is then weighted according to the direction of the movement - the movements in diagonal are √2 times as long as the movements according to medians - and according to the cost. To determine the shortest path, it is necessary to define so first of all the cost surface on which one moves. In the example presented above, it is enough to prevent the passage upright zones the depth of which in ebb tide is lower than a given depth, on one hand, and to prevent the motion upright zones where cables or pipelines are present. We explain below the way of creating the surface cost.

1.4.1 The creation of the cost surface

The cost surface establishes the raster datum necessary besides the co-ordinates of the origin and destination. These last ones can be supplied in the same way as that presented higher (keyboard encoding). The first can be generated by creating two masks (binary images) from two geographic information layers of which should exist. For the higher presented example, these layers are the DEM and a vector layer. The DEM represents the bathymetry corrected according to the height of the ebb tide of considered moment with regard to the reference plane adopted to create this DEM. **Then, a table supplying the heights of tides must also be available in the database, but this table is not sufficient if the covered area is wide and tidal state is variable in this area**. The vector layer is the arcs coverage representing cables and pipelines networks. This one should undergo a rasterization after definition of a buffer zone of width to be determined according to the criterion of security to adopt.

So, the first mask allows the definition of a zone with a depth lower than the considered value and the second, the restricted zones due to the presence of cables and pipelines. A Boolean operation must then be realized between both masks to create a synthetic mask representing the restricted zones. To this last one, one can then assign new numeric codes which represent the cost of crossing of the various zones. By assigning a null value to the pixels of the authorized zones with low cost value and a value trending towards the infinity at the pixels of zones the where the motion is not allowed. One obtains then the wished cost surface. Let us notice that the creation of the mask can be also envisaged by defining zones in vector mode (isobaths and polygons coverages resulting from the definition of buffer zones on the arcs coverage representing networks) by realizing an geoprocessing operation, recoding the result so as to define two types of zones according to their characteristics of depth and presence of an arc of the network, by computing a the motion cost values in a new field - crossing cost - as a result of the previous stage and rasterizing the result.

1.4.2 Determination of the shortest path

Generally, the determination of the shortest path knowing the surface that defines the cost of the displacement is achieved in raster mode. Therefore, the whole of the input data must be provided as image files (grids or lattices).

Under ArcView, these operations are foreseeable while calling on the Costdist extension (avx file) developed by ESRI (see appendix 1). **Nevertheless, in the setting of a request in the database via the Web, such operations risk to be very heavy to manage.**

The processing consists in two stages : the computation of the displacement cost and the direction toward a destination (CostDistance) and computation of the path to follow from a given origin to this destination (CostPath).

1.4.2.1 Computation of the displacement cost distance and direction toward a destination

The first stage, performed under ArcInfo thanks to the COSTDISTANCE command of the GRID module, is the creation of an cost distance surface to the target (or to the targets). This (or these) target(s) is (are) defined by a (some) point(s), arc(s) or polygon(s). A surface (raster file) that gives the direction of displacements can also be generated. Let's notice that if the space is isotropic, that means that the cost of the displacements is invariable, then the calculation of the surface of the distances (accumulated costs) is merely a construction of Thiessen or Voronï polygons.

The second stage is the determination of the shortest path (or paths) thanks to the raster layers thus created from different possible origins that can be the points, arcs or polygons. Under ArcInfo, it is performed thanks to the COSTPATH command of the GRID module. The result is a travel path drawn in raster mode. Therefore, it must be vectorised to be integrated in a cartographic document as a view and a layout of an ArcView project.

2 Practical examples with ArcView

2.1 Example of simple request on the attributes (tables)

To illustrate this kind of request performed under AcView, we have used data that consist in a table extracted from the IDOD database using the standard research form defined in Access and exported toward ODBC under the name of *Search_Results*.

At this level of the request process, let's notice that the simple request in the tables could be executed under Access and not under ArcView. Thus, a simplified table containing the wanted fields should be created so and submitted to the spatial request under ArcView. **At this level of the request process, the different strategies to adopt to facilitate and to accelerate the access to the database via the Web should be debated.**

Under ArcView, a SQL connection toward ODBC can be established so :

- Project/SQL connect
- IDOD - *Search_Results*
- All Columns
- Output Table: *Search_Results*.

Provided that this table contains geographical localization fields, this table constitutes an punctual events table. The geographical co-ordinates of the version 2 of the database - *Seawater2.mdb* – are not coherent. We have corrected the co-ordinates of the incorrect records and the following new version has been created : *Search_Results_corrigés_géogr*. An events theme can then be created using the View menu tools of ArcView as reported above:

- View/Add Event Theme
- Table: *Search_Results_corrigés_géogr*
- Xlongitude: *Longitude*
- Ylatitude: *Latitude*.

The created theme has been converted in shapefile and added to the Project View. This theme has been exploited thereafter to illustrate the different types of requests considered. The View containing the themes necessary for the cartography of the results has been completed by the Theme that consists in a polygons shapefile representing coastlines called *Bcoast-geog-poly.shp*.

The Themes used in the View are geometrically defined in a geographical reference system (geographical latitude and longitude in decimal degree). The View has been projected in the azimuthal equidistant cartographic reference system of Postel as while modifying the properties of the View (View/Properties/Projection). The unit used is the “meter” and none offset has not been defined for the fundamental point (projection centre) situated in 55°N and 0°E.

At this level of the illustration, we must underline the necessity to modify the way whose **temporal information** is stocked in the database. The " Start_dates " and " Start_Times " fields are defined like **alphanumeric** fields or characters chains - *String*. It is then complex to formulate a request clearly on the temporal attributes. We created therefore under ArcView three new fields that come from a recodification of these alphanumeric fields " Start_date ". These three new fields are the following numeric fields : " Day ", Month " and " Year ". The procedure used to perform the recoding is the following:

- Table/Start Editing
- Edit/AddField
- Field : *Jour (4,0)* ...

They have been computed using the Field/Calculate function so:

- ([Day] = ([Start_date].Left(10)).Right(2).AsNumber).

Nevertheless, we recommend to add these fields directly in the Access database however because the ODBC connection probably don't allow the date and time formats of the Access database. The same operation should be undertaken also on the " Start_Time " field, if the requests on the hour of the measures are to be considered.

The simple request on the thematic attributes of the tables is illustrated by the example of a conditional formula on the salinity and the season :

- ([Psalat31_Value] > 30) and (([Mois]>11) however [Mois]<5)).

She/it has been formulated using the Theme/Query function of ArcView. The geographical objects answering this condition positively have then been converted in shapefile (*Result - Requête_simple_tables.shp*) and added to the View of the Project. This conversion is not necessary if one wants to achieve a request based on a condition on the relative position of the geographical objects chosen in the analysed database or on the position of these same objects in relation to other exogenous geographical objects (see higher). In the case of a simple request, the addition of this theme is necessary for the creation of a map that contains the results of the request presented as a Layout.

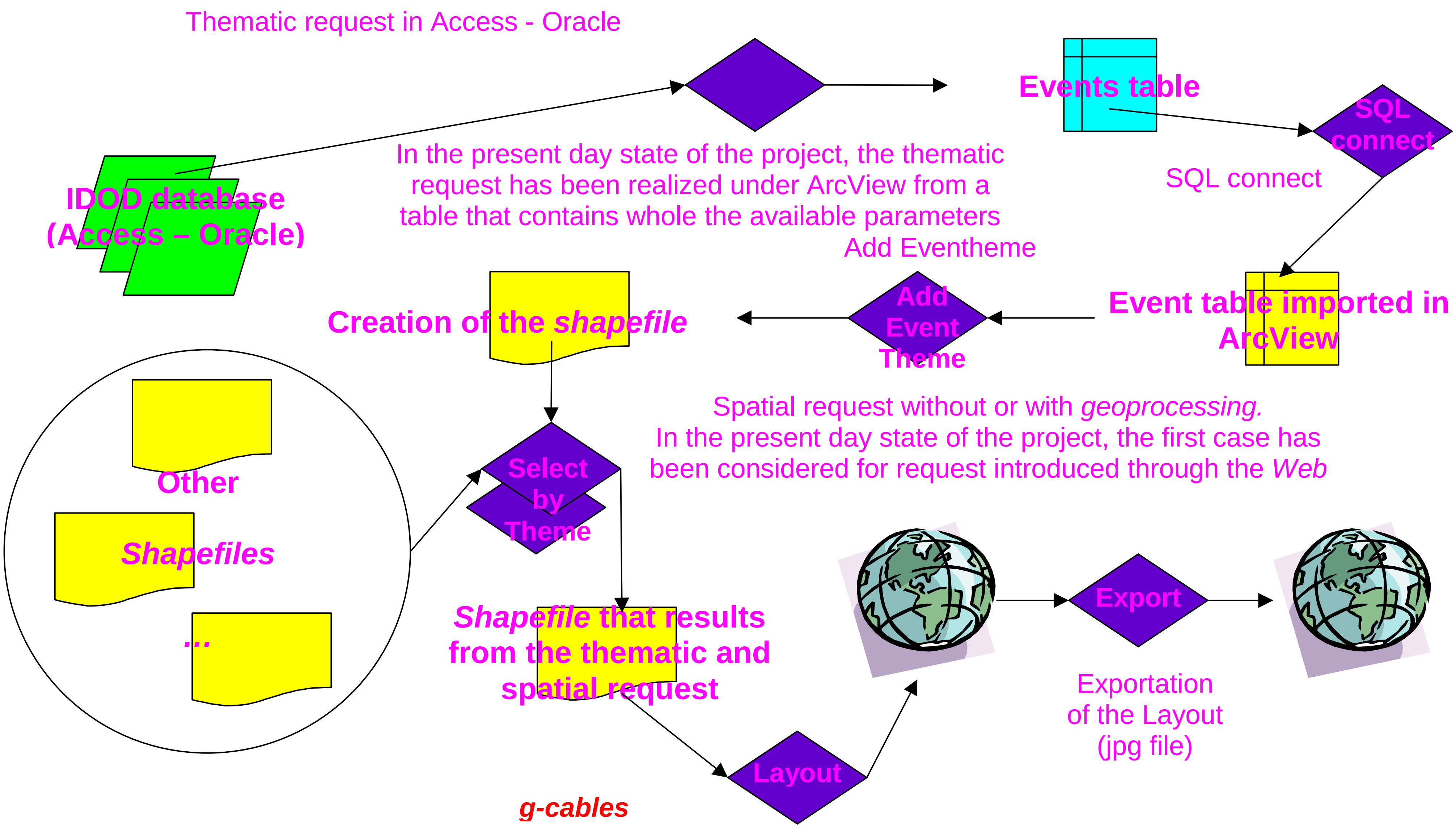
2.2 Example of simple request on the thematic attributes and the position resolved without geoprocessing

To illustrate this type of request, simple request on the thematic attributes and the position resolved without geoprocessing, we start the process from the previous example that constitutes the first phase of the treatment. This phase can be disregarded if no particular criteria concerning these thematic attributes must be considered.

To show the way to proceed to take in consideration the spatial aspect, we add a criteria concerning the position of the measurement points in relation to the coastline, for instance. Among the measures that answer the thematic condition recalled higher, we illustrate here under a selection of a new set of measurement points. The spatial criteria that must be verified is the following one : the distance between the coastline and the measurement points that must be selected must be to less of 15000 m. This type of research in the database can be performed while exploiting the Selection by Theme function of the View menu of ArcView. The active theme is the one that contains the geographical information that must be selected. It is about *Result - Requête_simple_tables.shp*. The condition is about a maximum distance of 15000 m from the theme that define the coastline, *Bcoast-geog-poly.shp*. The result of the selection can then be converted in shapefile and added to the View to be presented in a Layout.

The following figure shows the processing flow chart of this kind of request.

Figure 1 : processing flow chart of a simple request on the thematic attributes and the position resolved without geoprocessing



In the present day state of the project, the thematic request has been realized under ArcView from a table that contains whole the available parameters

Spatial request without or with *geoprocessing*.
In the present day state of the project, the first case has been considered for request introduced through the *Web*

2.3 Simple request example on the thematic attributes and the position resolved by geoprocessing

This example illustrates the most general case of treatment requiring geoprocessing. It consists in the creation of a polygons cover from two other polygons covers by the computation of the intersections and creation of an attributes table taking the attributes of the two input polygons covers.

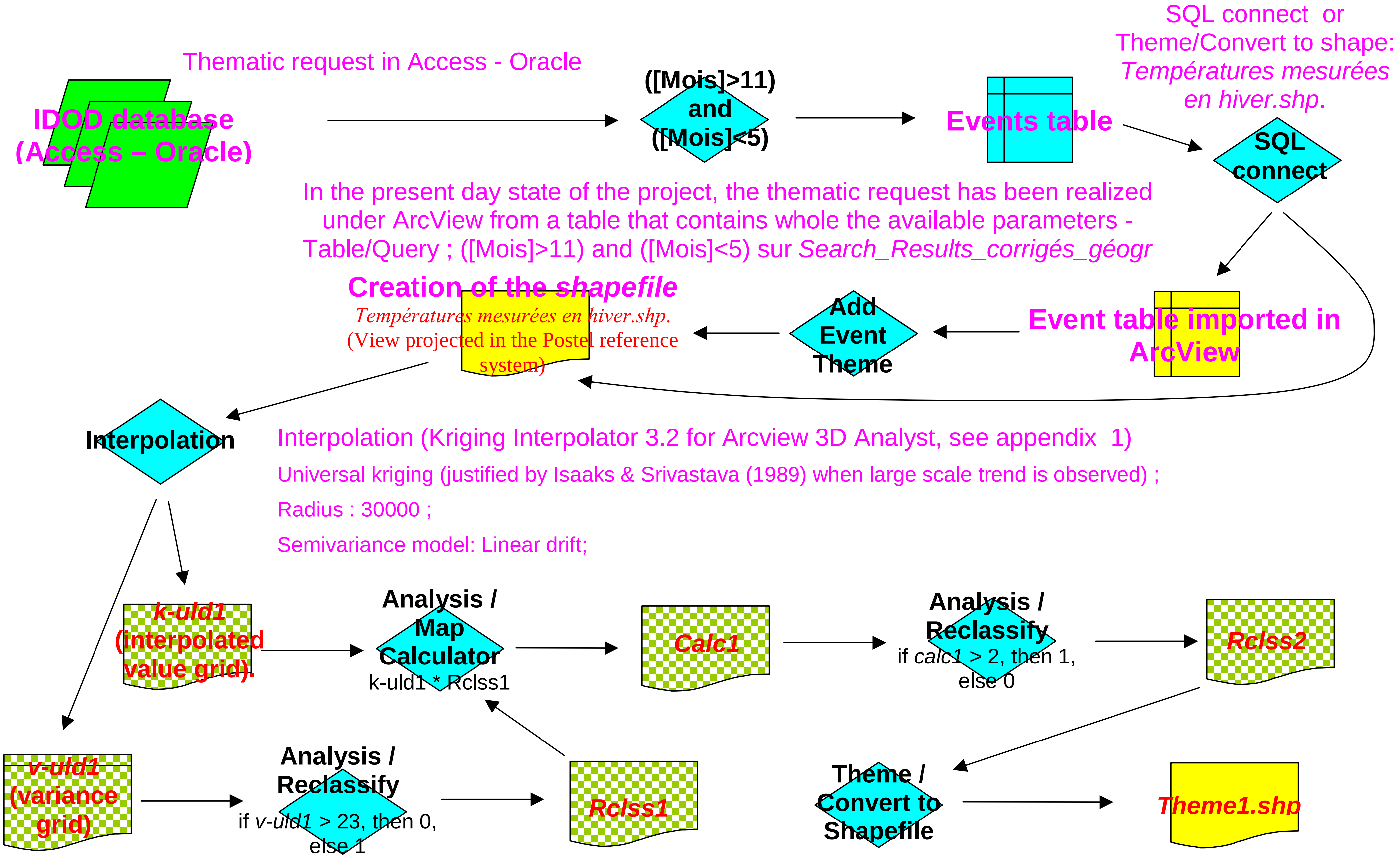
The example of request is the next one: one wishes to select the zones where the wintry temperature (from the beginning of December to the end of April) is greater than 2° and localized nearer than 10000 m from the coasts.

The data are the following ones :

- the IDOD database as an Access *Search-Results* event table (see higher);
- a polygons shapefile representing the coastlines, *Bcoast-geog-poly.shp*.

The operations to achieve are shown in the flow chart of figure 2..

Figure 2 : processing flow chart of a simple request on the thematic attributes and the position resolved by geoprocessing



Thematic request in Access - Oracle

SQL connect or Theme/Convert to shape: *Températures mesurées en hiver.shp*.

In the present day state of the project, the thematic request has been realized under ArcView from a table that contains whole the available parameters - Table/Query ; ([Mois]>11) and ([Mois]<5) sur *Search_Results_corrigés_géogr*

Creation of the *shapefile*

Températures mesurées en hiver.shp. (View projected in the Postel reference system)

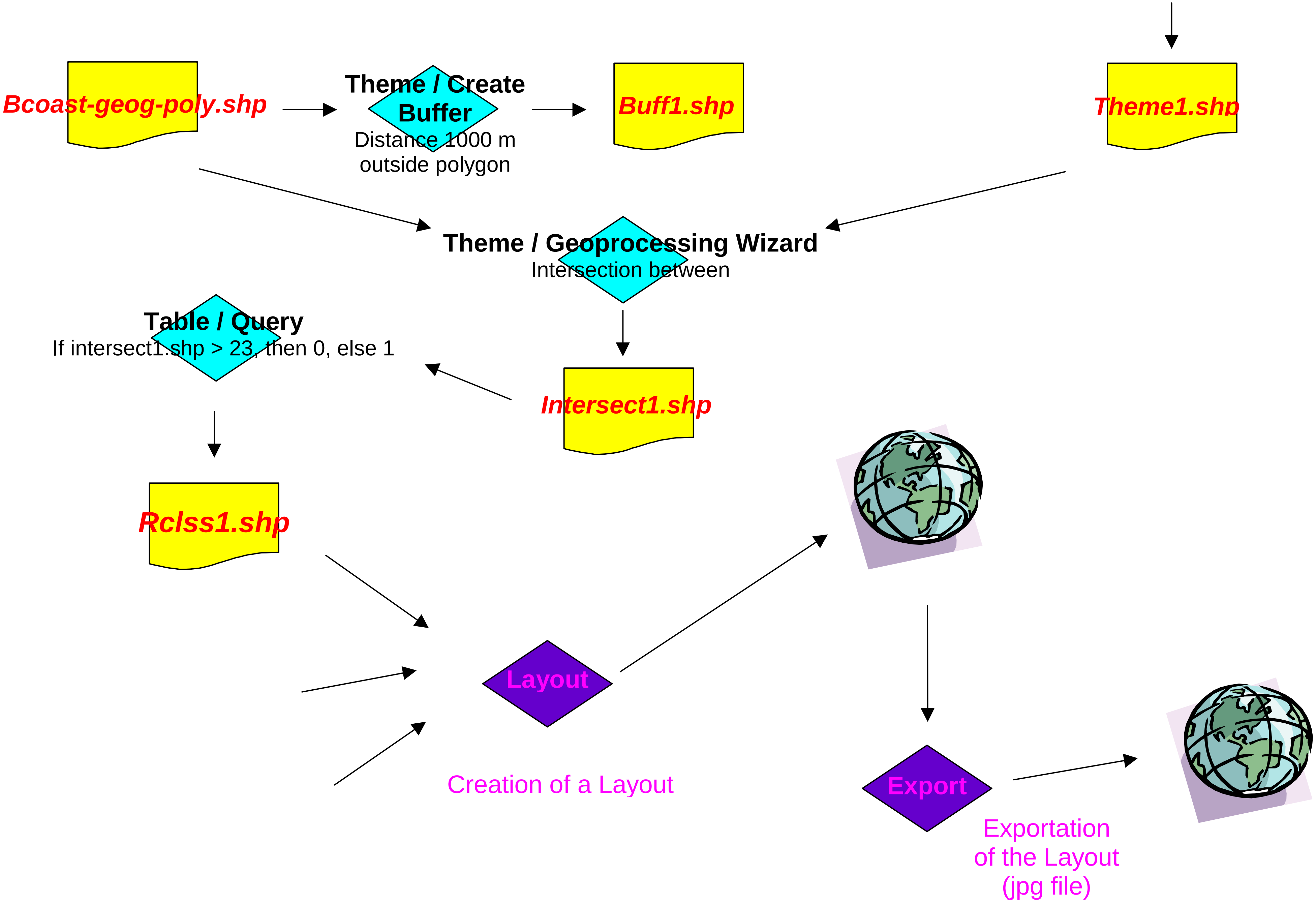
Interpolation (Kriging Interpolator 3.2 for Arcview 3D Analyst, see appendix 1)

Universal kriging (justified by Isaaks & Srivastava (1989) when large scale trend is observed) ;

Radius : 30000 ;

Semivariance model: Linear drift;

IDOD database (Access – Oracle)
([Mois]>11) and ([Mois]<5)
Events table
SQL connect
Event table imported in ArcView
Add Event Theme
Interpolation
k-uld1 (interpolated value grid).
Analysis / Map Calculator
k-uld1 * Rclss1
Calc1
Analysis / Reclassify
if calc1 > 2, then 1, else 0
Rclss2
v-uld1 (variance grid)
Analysis / Reclassify
if v-uld1 > 23, then 0, else 1
Rclss1
Theme / Convert to Shapefile
Theme1.shp



Bcoast-geog-poly.shp

Theme / Create Buffer

Distance 1000 m outside polygon

Buff1.shp

Theme1.shp

Theme / Geoprocessing Wizard

Intersection between

Table / Query

If intersect1.shp > 23, then 0, else 1

Intersect1.shp

Rclss1.shp

Layout

Creation of a Layout

Export

Exportation of the Layout (jpg file)

2.4 Requests that combine vector covers, TIN, grids and lattices

An example of request whose solution must be found while editing a vector layer, a TIN, a grid or a lattice has been formulated at the meeting of the 4/10/2000. Here is its formulation: is it possible to get the depth at a given punctual location on basis of a DEM of the bathymetry.

In stand alone, a solution to this request can be found while exploiting the cover (theme) inquiry tool. The operations to put in work are to following ones :

- creation of a View, adding to it the theme (grid) on which the request must be applied, that constitutes a data existing in the IDOD database, *g-bathy100aze* and activating this theme;
- activation of the identification tool (Identify) and edition of the given point.

An alternative to these operations consists in exploiting the avx extension that allows profiles extraction (cfr. appendix 1). In fact, it can be also used in more complex cases as those evoked below. Indeed, one can generalize the problematic to more complex requests that combine different kinds of vector covers and raster. **An example consists in extracting a profiles, for example, according to a straight or more complex arc that can correspond to the track of the ship at the time of a measurement campaign achieved by the requiring people and to extract the grid values according to a constant step along this arc in different raster covers that corresponds to the depth, the temperature and the salinity, for instance.**

In the different relative steps to this kind of request introduced on the Web, the interactivity of the tools used is very limited and the definition of the points, lines where polygons that will orient the spatial research constitute a unfavourable factor to their use on the Web. The Arc IMS software will probably be an alternative.

2.5 Determination of the shorter path on basis of a surface cost

As mentioned higher in this report, the determination of the shorter path between on origin and a destination point will probably be difficult to be put in work on the Web considering the heaviness of the processing that it imposes. Indeed, the largest part of the processing is achieved in a raster mode. The processing time and the dimension of the created files are the main inconveniences of the method.

An example of request requiring this kind of processing has been evoked during the meeting of the 4/10/2000. Here is its formulation: **how can we search for the shortest path between two points, an origin and a destination, without travelling through the zones where water depth is lower than 19 m and through the area that are less distant from cables, pipelines… than 1000 m.**

The data are the following ones :

- a 100 m resolution bathymetric DEM, *g-bathy100aze*; this DEM is referenced in the Postel azimuthal equidistant cartographic system;
- a shapefile giving the position of cables and pipelines, *cables-gas-oil.shp*;
- an ASCII event table, *orig-dest.txt,* corresponding to the Postel cartographic co-ordinates of the origin and destination computed from the geographic co-ordinates encoded by requiring people.

The processing that must be performed is presented on the flow chart of the figure 3.

A. Création de la surface coût :

   - Surface/Find Distance sur *g-cables : dist1* (cette opération pourrait être effectuée directement sur le shapefile en évitatn, donc l'étape précédente) ;
   - Surface/Reclassify si *dist1* < 1000, alors 1, sinon 0 : *rclss1* (remarquons que les limites de classes peuvent être lues dans un fichier ASCII de type .avc qui peut être écrit sur base de valeurs fournies par le requérant);
   - Analyis/Reclassify si *g-bathy100aze* < 19, alors 1, sinon 0 : *rclss2* (même remarque) ;
   - Combinaison des deux images binaires : Surface/Map Calculator : *rclss1* = 1 où *rclss2* = 1, alors 1, sinon 0 : *calc1* ;
   - Calcul de la surface coût : Surface/Map Calculator : *calc2* = (1/[*g-bathy100aze])+[calc1]* 500*

B. Calcul des distances-coût et directions de déplacement

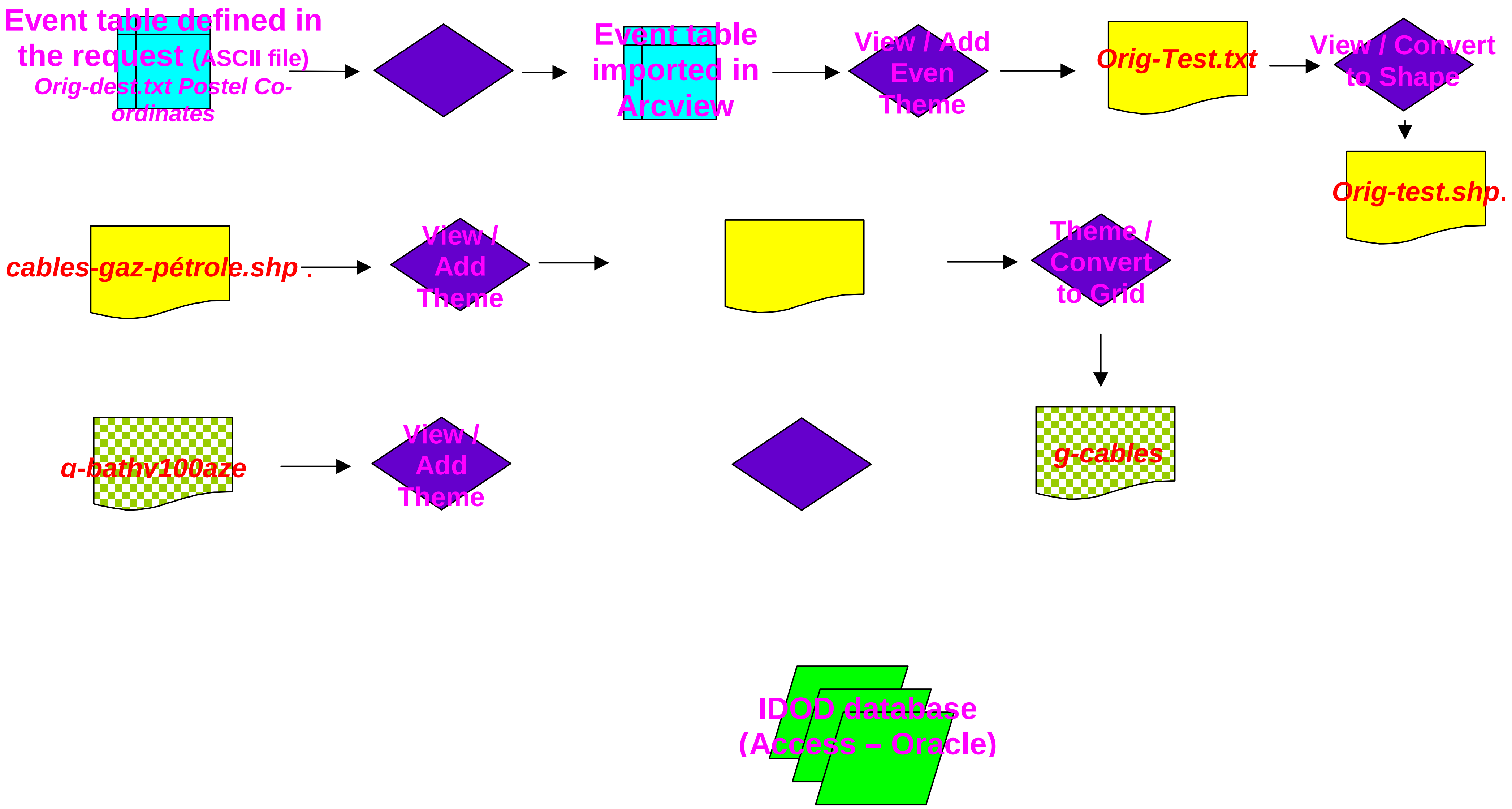
   - Costdist/CostDistance sur *calc2* à partir de la destination qui doit être sélectionnées dans le thème d'évènements correspondant à la tables *orig-dest.txt*. Le résultat consiste en deux *grids* : *Distance from origine* et *Direction ;*

C. Extraction du chemin à suivre en mode *raster* (*grid*) ou verteur (*shapefile*) :

   - Costdist/CostPath sur *Distance from origine* et *Direction* à partir de l'origine qui doit être sélectionnées dans le thème d'évènements correspondant à la tables *orig-dest.txt*. Le résultat consiste soit en un *grids*, soit en un *shapefile*.

D. Habillage et création du *Layout* exporté ensuite en fichier .bmp.

Figure 3 : processing flow chart of the least cost path determination process



ANNEX 1 : THE LIST OF THE AVX EXTENSIONS AND AVE SCRIPTS FROM THE ESRI WEB SERVER (HTTP://WWW.ESRI.COM)

PE60_SA et PE60_3D

Profile Extractor 6.0 for 3D Analyst

Extracts cross section profile from a DTM. The user can draw cross section line (polyline, rectangle, circle or polygon), select existing one, move selected line or it's ends or rotate a line around it's middle point. The profile is drawn into a Chart.

Many users asked for a PE 6.0 version for 3D Analyst and it is available now

Extracts profiles from TINs as well as from GRIDS. The user can set GRID (or TIN) as a primary surface and TINs or GRIDs as secondary ones and create profiles comparing them.

Currently volume calculations and Cut & Fill analysis not available for the version for 3D Analyst

Author: Ianko Tchoukanski

Date: 09-Apr-00

Primary Software: ArcView 3.x

Language: Avenue

Category: ArcView 3D Analyst

User Extension: No

Keywords: cross section, profile, surface

WEBVIEW11

WebView 1.1

The Internet Extension for ArcView GIS

Present your maps on the Web or on CD ROM!

FREE Evaluation copy of WebView 1.1, the Internet Extension for ArcView GIS.

WebView is for you if you want

to present your geodata on the Web - including a detail and an overview map, legend and scale bar, and important GIS functionality like Zoom in and Zoom out, Pan, Identify and Hotlink want to present your GIS-projects on CD ROM - without the need to deliver your original geodata (shapefiles, images).

but don't

want to spend a lot of money want to bother with programming in Avenue, HTML or JavaScript want to take the troubles of maintaining your own internet server.

The created map pages provide the following functionality: Detail map and overview map Zoom in and zoom out (up to 3 zoom levels, free selection of the scale for the 1st zoom level) Pan in the detail map, using buttons or by clicking into the overview map Different map contents at different zoom levels (scales) Legend for map contents Scale bar Interactive layers:

* up to 5 layers

* different layers for different zoom levels

* attribute information on mouse move (map tip) or mouse click (identify)

* hotlink to images, web pages or email addresses

To test WebView 1.1, download the FREE Evaluation copy from the ArcScripts page or from our homepage: www.zebris.com or www.3w-scope.de

To install WebView, unzip the files to an installation directory (make sure that checkbox 'Use folder names' is selected when unzipping the files) and execute install.apr.

Author: Thomas Zerweck

Date 21-Aug-00

Primary Software: ArcView 3.x

Language: Avenue

Category: ArcView Project,ArcView Scripts,ArcView Views: General,Miscellaneous,Projects

User Extension: Yes

Keywords: Internet, Web, Intranet, JPG, Map, ArcView, Avenue, HTML, Presentation

BE_PROFILESWITHFEATUREPTS.AVE

Profiles with feature points

Use to create elevation profiles of selected lines and selected point features.

Requires Spatial Analyst Extension.

Requires three active themes, The first a point theme (the feature points) and the second a line theme and the third a grid theme of elevations. Points do not need to be exactly on the line. Uses request PointPosition to find point on line nearest to feature point.

Requires feature points to be selected.

Requires lines to be selected. In general, selected lines should be connected for results to be realistic. There is no programmatic check to ensure selected lines are connected. In general, selected lines should not contain multiple branching for results to be realistic. That is, each node should connect no more than two _selected_ line segments.

Merges selected lines then finds interval points along the merged line at equal intervals.

User is queried for how many divisions of merged line.

Queries active grid theme cells for elevation values at interval points and selected feature points.

Outputs a dbf file of distances starting at the lowest interval point and includes all interval points, feature points and corresponding z values. Dbf table contains a field "Type" indicating if a distance is an interval point or a feature point.

Outputs a scatter-diagram graph (the line profile) of distance from origin vs. z values if no more than about 50 points are produced. Use a more robust graphing package to graph larger outputs.

Use a more robust graphic package to construct scatter-diagrams which differentiate interval and feature points by color.

Author: Bill Eichenlaub

Date 19-May-98

Primary Software: ArcView 3.x

Language: Avenue

Category: ArcView Spatial Analyst,ArcView Views: Analysis

User Extension: No

Keywords: profiles, grids, streams

N_KRIGINGINTERPOLATOR3D.AVX et N_KRIGINGINTERPOLATORSA.AVX

Kriging Interpolator 3.2 for Arcview 3D Analyst

This is the extension version of the kriging1.1 script. It needs 3D Analyst and Dialog Designer installed. NOTE: The functionality of the Spatial Analyst and 3D version are exactly the same, they are only depended on different extensions. It has been tested under NT4 and Unix, and should run fine on both of these platforms. Enhancements made over 1.1:*Dialog designer setup, "Geoprocessing Wizard - like". *All functionality accessible from 1 main dialog. *Enhanced help functionality using Arcviews online help and list of FAQ's. *Goodness of Fit calculation between actual and fitted semivariance *Non-modal dialogs allows you to zoom in on layout containing svg's. *UNIX bugfix. Other functionality already available in 1.1: This is a FULL implementation of the kriging commands in Spatial Analyst and includes features like: -1) All methods implemented: For ordinary kriging the spherical, circular, exponential, gaussian and linear methods for modelling the semivariance are available. For universal kriging both options (linear and quadratic drift) have been implemented. -2) Ability to process several methods in a single run. -3) Abilitiy to use selected records for either the calculation of a semivariogram and / or the interpolation -4) Pre-examination of semivariograms on-screen as a layout. -5) Ability to print this layout during execution of script. -6) Ability to export semivariogram data to a dBASE, INFO or TEXT file during execution of script for examination of data in another program like excell. -7) Ability to set a barrier line theme -8) Ability to set either a fixed or variable radius -9) Ability to create an optional variance grid for each method. -10)Goodness of Fit calculation

Author: Marco Boeringa

Date 27-Feb-99

Primary Software: ArcView 3.x

Language: Avenue

Category: ArcView 3D Analyst,ArcView 3D Scenes,ArcView Views: Analysis,ArcView Views: Grid Themes

User Extension: Yes

Keywords: kriging, interpolation, analyst, 3D, surface, grid, DEM, DTM

INTERP10.AVX

Surface Interpolator Extension

Provides a dialog from which all four Spatial Analyst surface interpolation methods can be used. For Kriging interpolation, a chart representing the semi-variogram is produced before a surface is interpolated. Note: to use with 3D Analyst, the extension file can be edited to remove the dependency on Spatial Analyst. README file included.

This version of the "Surface Interpolator" Extension is a free sample extension for use withArcView GIS. This ArcView extension was designed to run on ArcView 3.x with Spatial Analyst v1.1. ESRI's Spatial Analyst and the Dialog Designer must be installed.

This extension allows the interpolation of surfaces using the IDW, Spline, Trend, or Kriging technique.

For Kriging, the user can choose one of the two Universal methods (linear or quadratic drift), or one of the five Ordinary Kriging methods (linear, exponential, circular, spherical, or gaussian). When using Ordinary Kriging, the user is able to view the Semi-Variogram before performing the surface interpolation. The Semi-Variogram shows the relationship of semi-variance to distance in the sample points, and the quality of fit between the actual and fitted values for the method chosen. Beyond a certain distance (referred to as the 'Range'), variation in the sample point Z-Value is no longer spatially correlated. Within the range, variation in z-values is smaller when sample points are closer. Consult the ArcView on-line help for more information on the Kriging interpolation technique: search the index with the keyword 'MakeFromVariogram', and then click on the 'Discussion' hypertext link at the top of the page.

Semi-Variograms are produced as layout documents, although the user also has the option of producing a table.

Author: Thad Tilton

Date 22-Aug-98

Primary Software: ArcView 3.x

Language: Avenue

Category: ArcView Spatial Analyst

User Extension: Yes

Keywords: Kriging, interpolation, dialog, semi-variogram

XKRIGING.AVX

Kriging Interpolation Extension

Applying kriging method to interpolate points coverage to grid. Runs under Windows NT/95/98 for ArcView 3.x with Spatial Analyst. This extension will add a button to the tool bar and the whole process is just one-click-away. Click here for a detailed instruction on installation and other information.

Author: Ningchuan Xiao

Date 25-Oct-99

Primary Software: ArcView 3.x

Language: Avenue

Category: ArcView Menus/Buttons/Tools,ArcView Scripts,ArcView Spatial Analyst,ArcView Views: Analysis

User Extension: Yes

Keywords: kriging

GEN2SHP3.AVX et G2SV40.AVX ET G2SV50.AVX

Add theme from GENERATE format ASCII file.

Import points, lines, polygons and graphic shapes, all either 2D or 3D

The extension Gen2shp reads one or more ASCII text file(s), converts the data to the appropriate feature type and saves it as one or more shapefile(s).

This latest (and probably last) version supports multi-part Line and Polygon shapes as well as Point, Line, Polygon, Rectangle, Circle and Donut shapes. All types can be 2D (X,Y) or 3D (X,Y,Z).

For more information, examples and sample data visit: http://warden.www.cistron.nl/geo/

Available below are the extension (file g2sv50.zip) and scripts (file g2sv5src.zip). Note that the scripts are only of value to Avenue developers.

Gen2shp is licensed under LGPL.

Author: Ron S.W. Wardenier

Date 27-Aug-98

Primary Software: ArcView 3.x

Language: Avenue

Category: ArcView 3D Analyst,ArcView Spatial Analyst,ArcView Views: Data Conversion/Alteration,Data Conversion

Keywords: import, conversion, GENERATE, ASCII, 3D

TXTTOOL.AVX

ASCII <-> Shapefile Tool

Converts asii file containing coodrinates to shapefile and exports the coordinates of shapefiles to ascii file. Works for import space delimited ascii to point, polygon and polyline. The format for point ascii file is id, x, y (no comma for real data, space delimited). For polygon & polyline ascii files, the format is code (1 for start point, 2 for middle points, 3 for end point), x, y (no comma for real data, space delimited). Export shapefile to ascii file works for point, polyline and polygon shapefiles. The output format file is id, x, y. For polygon and polyline, the id is the sequence id of vertice.

Author: Wei Sun

Date 15-Feb-00

Primary Software: ArcView 3.x

Language: Avenue

Category: ArcView Views: Data Conversion/Alteration

User Extension: Yes

Keywords: ascii, import, export, data conversion

AGFSHP.AVX

To convert Atlas Gis files to shapefiles.

SC_EDITOR.AVX

Single Cell Editor

This extension will allow the user to change values of individual Grid cells. A temporary polygon theme is used to store the user selected cells and values. The user has the option of saving this temporary file for reference or versioning purposes. The values of the neighboring cells, and statistics for these, are also displayed and the user can select any of these values as the new value for the currently selected cell. The changes are stored in a new Grid and a Theme is added to the View.

Author: John Grayson

Date 11-Aug-99

Primary Software: ArcView 3.1

Language: Avenue

Category: ArcView Spatial Analyst,ArcView Views: Grid Themes

User Extension: Yes

Keywords: single cell editor grid change value spatial analyst

COSTDIST.AVX

Cost Distance Grid Tools

This extension presents the Cost Distance and Cost Path tools in the Spatial Analyst as a menu of choices added to the view GUI. The extension also adds a tool to the view GUI allowing the user to enter points from which to return a least cost path. The specific Avenue requests for each menu choice are described in the accompanying text file. Specific information about the use of the request is available in the Arcview online help.

Using the Cost Distance Tools extension for the Spatial Analyst version 1.1.

This extension presents the Cost Distance and Cost Path tools in the Spatial Analyst as a menu of choices added to the view GUI. The extension also adds a tool to the view GUI allowing the user to enter points from which to return a least cost path. The specific Avenue requests for each menu choice are listed in brackets following the menu choice name below. his extension presents the Cost Distance and Cost Path tools in the Spatial Analyst as a menu of choices added to the view GUI. The extension also adds a tool to the view GUI allowing the user to enter points from which to return a least cost path. The specific

Avenue requests for each menu choice are described in the accompanying text file.

Specific information about the use of the request is available in the Arcview online help.

Author: ESRI Date 27-Jul-98

Primary Software: ArcView 3.x

Language: Avenue

Category: ArcView Spatial Analyst

User Extension: No

Keywords: Spatial Analyst, CostDistance, CostPath

GRDMAKER.AVE

View.GridMaker

Creates arbitrary grid(vector) shapefile.

This script is attached to a tool. When you drag a rectangle with the tool, you are then prompted for how many rows and columns you want in the grid. Then you are asked for the name of a shapefile to create, and the grid is created. One attribute is added to this grid shape theme, a label, that is "A1 A2 B1 B2...." as you might see on an index map. The labeling starts at the upper left.

Author: ESRI Date 25-Mar-98

Primary Software: ArcView 3.x

Language: Avenue

Category: ArcView Views: General

User Extension: No

Keywords: grid, vector

NEAREST.AVE

View.SpatialNearestNeighbor

Perform Spatial Nearest Neighbor Analysis.

Create a new tool in a ViewDocGUI with this script as the Apply script. To use this you must have an active theme. Just drag a rectangle around the features you wish to conduct a spatial nearest neighbor analysis of. The wait cursor will appear and then a messagebox will tell you the R value and how many features were accounted for in the analyis. R values relate how clustered or dispersed points (or centroids of polygons and polylines) are within the rectangle you specified. An R value of 0 (zero), indicates an intensely clustered pattern, while an R value of 1 indicates a random distribution, and an R value of 2 (or higher) indicates strongly dispersed or organized pattern. Requires: The active document must be a view, with an active feature theme.

Author: ESRI

Date 25-Mar-98

Primary Software: ArcView 3.x

Language: Avenue

Category: ArcView Views: Analysis

User Extension: No

Keywords: spatial, nearest, neighbor, ViewDocGUI

COORDINATEZ_E.AVX

Put the Coordinate Z in the table

If you have a theme 3D and you dont´t have the coordinate Z in the table, try this script download the English version - coordinatez_E.

Author: Pedro Guimaraes

Date 19-Sep-00

Primary Software: ArcView 3.x

Language: Avenue

Category: ArcView Tables,ArcView Views: General,Tables

Keywords: Coordinates, coordenadas, Z, Tables

DD2DMS.AVE

DD2DMS

DD2DMS is a conversion calculator for converting a single coordinate between decimal degrees and degrees-minutes-seconds.

Author: Ilyas Goja

Date 08-Sep-99

Primary Software: ArcView 3.1

Language: Avenue

Category: ArcView Views: Data Conversion/Alteration

User Extension: Yes

Keywords: decimal degrees DD2DMS.AVE

DMS2DD.AVE et DMSTODD2.AVE

DMS to Decimal Degrees

This Avenue script converts a table with two fields (longitude and latitude) in the form of DDMMSS and creates two new fields for decimal degrees. (Multiplies the longitude by –1 will need to edit for other locations.) (This is a modification of the script that comes with ArcView.). Converts DegreesMinutesSeconds values in a table to DecimalDegrees. A new field for the converted values is added to the table. Values greater than 180 or less than -180 may not be converted correctly. Requires: The table must be editable (dBASE or INFO) and you must have write access to it. The table must have a field in which the DMS values are stored in the format DDD MM SS.

Author: Kiana Zimmerman

Date 13-Apr-00

Primary Software:ArcView 3.x

Language: Avenue

Category: ArcView Tables,ArcView Views: Data Conversion/Alteration

User Extension: No

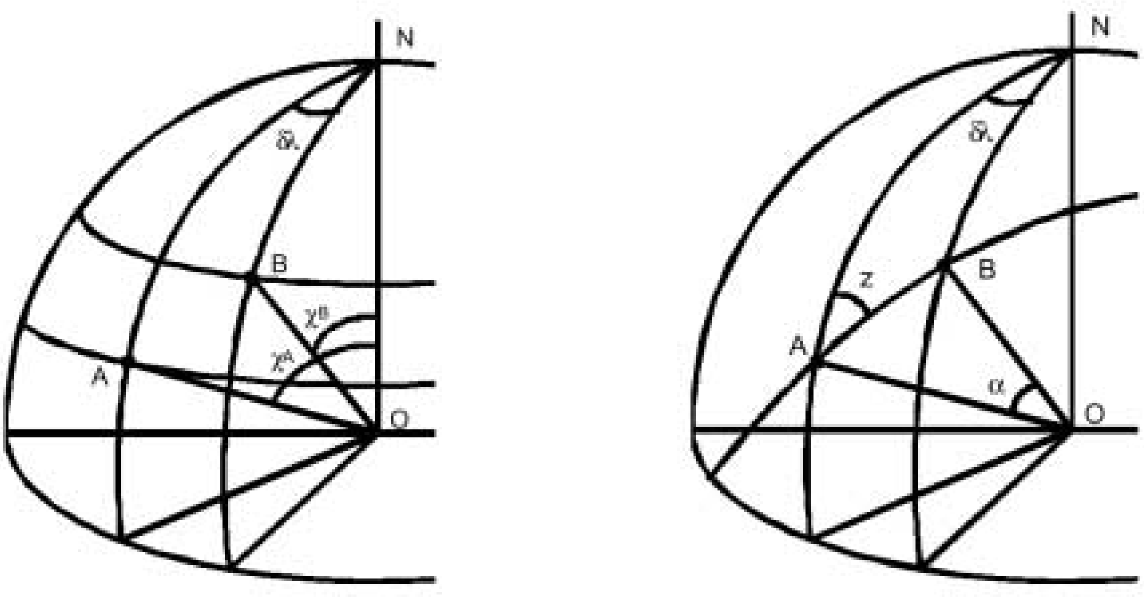
Keywords: latitude,longitude,degrees minutes seconds, decimal degrees

ANNEX 2 : THE CARTOGRAPHIC PROJECTION SYSTEM

The choice of the cartographic projection has been discussed during several meetings. The final choice has been performed during the meting of the 13/09/2000. The cartographic projection system used is the azimuthal equidistant projection system of Postel. This system has been used in spite of the azimuthal equalarea projection of Lambert because of its possible use in ArcView 3.2a (Projection/Wizard tool).

The oblique aspect has been used with a tangential point (projection centre point) located at 0°E, 55°N and a sphere with a radius of 6371 km. Nevertheless, the Postel projecion conserves the distances along the radius coming from the projection centre and the Lambert one conserves the area. For both reference systems, the azimuths from the projection centre are the same. In the first case, the projected distance is the length of the great circle bow joining the projection centre and the considered point and in the second case, it is the cord of this bow. The distance between the projection centre and a considered point is than greater in the case of the Postel projection than in the case of the Lambert projection. No cartographic offset has been defined (*X* and *Y* values can be greater, equal or less than 0 in function of the azimuth angle)

The following figure shows the geometrical concept of these projection systems.



The formula that give the cartographic co-ordinates for the Postel projection system are:

$Y_{Postel} = R\,k'\left[\cos\varphi_0 \sin\varphi - \sin\varphi_0 \cos\varphi \cos(\lambda-\lambda_0)\right]$

$X_{Postel} = R\,k' \cos\varphi \sin(\lambda-\lambda_0)$

$k' = \alpha / \sin(\alpha)$ and $\cos\alpha = \left[\sin\varphi_0 \sin\varphi - cis\,\varphi_0 \cos\varphi \cos(\lambda-\lambda_0)\right]$

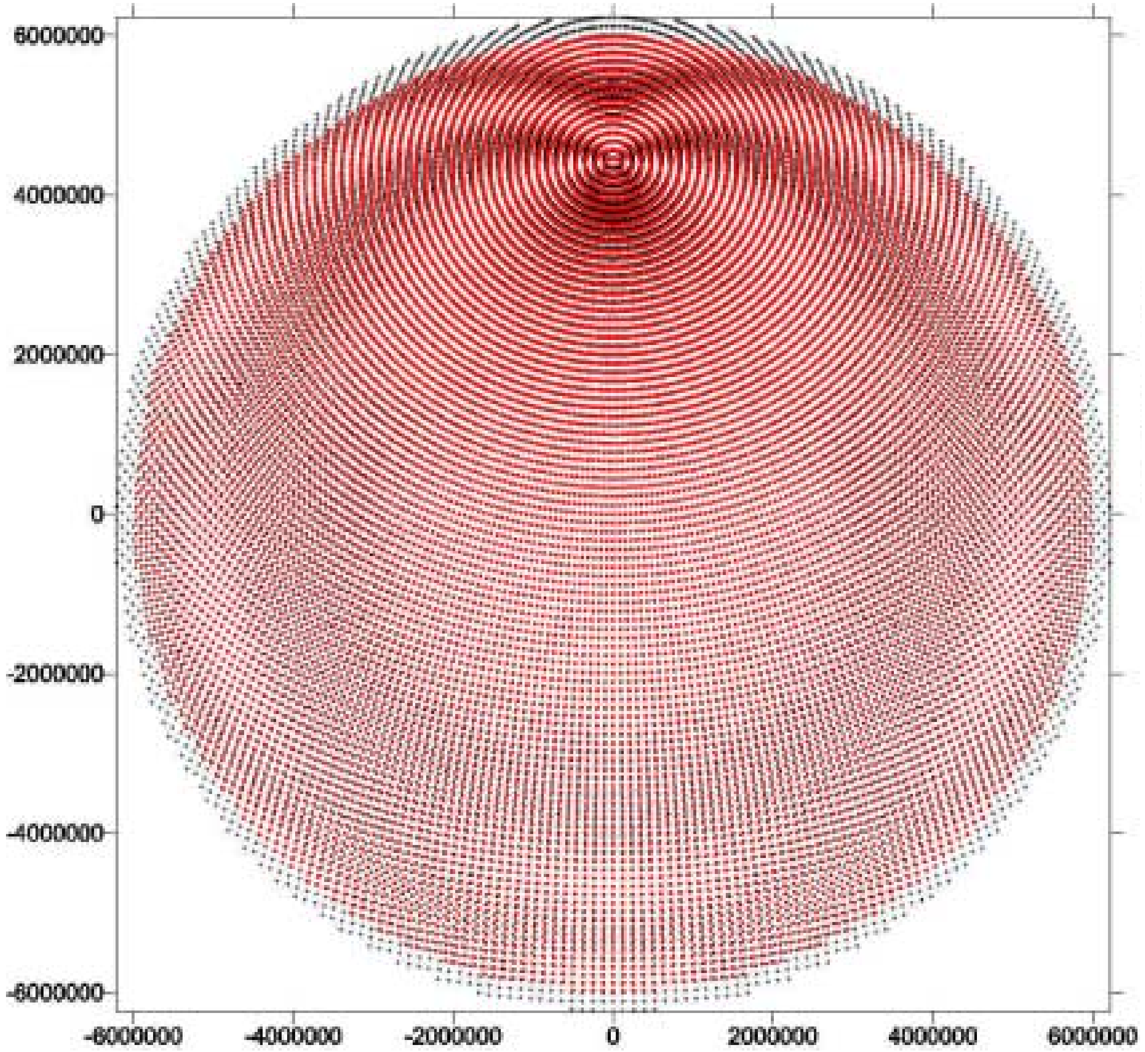
$\lambda_0 = 0°$ and $\varphi_0 = 55°$

Azimuthal equidistant projection of Postel: projection of the arc AB - Equalarea azimuthal projection of Lambert: projection of the cord of the arc AB

A is the projection centre (tangential point)

The following figures show the distribution of a set of projected nodes of a 1 degree latitude and longitude net in both Postel and Lambert projection systems.

Azimuthal equalarea projection of Lambert (red)
Azimuthal equidistant projection of Postel (black)
Cartographic co–ordinates in meters



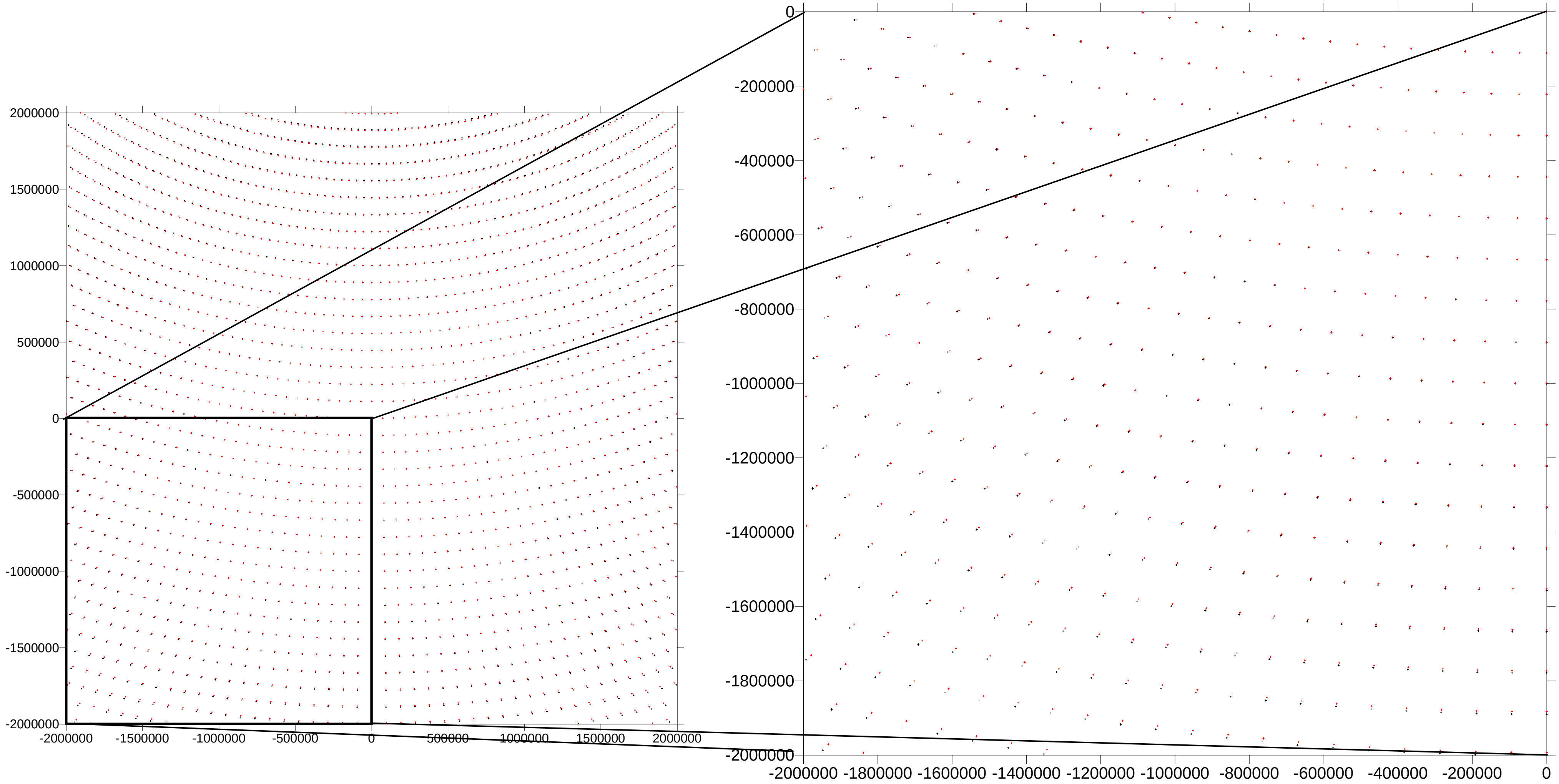
Projection center: (50°N , 0°E) in geographic co-ordinates = (0 , 0) in cartographic co-ordinates.

Selected points : α < 0.98 rad.

Distance along the cord from the projection center to the little circle corresponding to this angle : about 6000 km.

Radius of the sphere : 6371 km.

Zoom in the SW quadrant from 0 to – 2 000 000 m along the *X* and *Y* axes



PART III

Spatial data distribution, trend lines and surfaces, interpolation methods and validation

I. General introductive considerations

Water temperature and salinity are quantitative variables. They are generally measured and known from oceanographic campaigns at some stations or points along the track of the ship that are localized. Since these variables are known for some punctual entities and their spatial variation is continuous on the mathematical point of view, the surface that represent the value of the variable in function of the location (S(x,y)) can be restituted using interpolation processes.

This interpolation processes can be exact or approximate. In the first case, the restituted surface passes through the observed values (inverse distance weighted interpolation process, TIN-linear and quintic interpolation process, kriging interpolation process without nugget effect). In the second case the surface doesn’t pass through the observed values (spline, first, second and third order trend surfaces). For all kinds of interpolation processes, the restitution of this surface allows the cartography of this continuous variable in image (at a defined resolution level) and vector (after contouring the computed image) modes. The interpolation process must be chosen, defined and performed in function of the spatial distribution of the sampled data that must be representative of the spatial variability of these studied phenomena’s.

So, the measurement must only be influenced by the phenomena’s the scale of which is in relation of the density of sample. Moreover, in oceanography all the point measurements are not performed at the same moment of the tidal cycle for each location. This observation induces a bias in the data set because all the measurements are not temporally comparable. Moreover, the small number of points in a relatively large area of interest will only allow the restitution of large spatial scale variability. But this can be considered if the individual observations are not influenced by small scale phenomena. The following example can illustrate this statement. It has been shown in the *Vlaamse Banken* area that the temperature measured across the mega-dunes of the sea bottom is systematically different on both sides of the dunes. The magnitude of this difference is generally several 1/10 of °C and depends on the orientation of tidal currents relatively to the bank orientation. If a part of the data set is affected by this phenomenon, then the interpolation process will be influenced by this spatial influence. If its magnitude is high with respect to the variance of the data set, then it must be eliminated before performing the interpolation by applying a correction model. This could eventually be defined using a DTM of the sea bottom topography and tidal state estimated at each station at the acquisition time. The variability of the measured parameter can be estimated even in function of the direction using specific geostatistical methods based on the construction of omnidirectional and directional semivariograms.

But, independently of this variability and before analysing it, the spatial distribution must be qualified and quantified by several geostatistical parameters like the position, the length and the orientation of principal axes of an ellipse, the dispersion ellipse. The position of this ellipse center is the gravity center of the points. The length of these axes corresponds generally to a real multiple of the standard deviation along these axes. The orientations of these axes give the quantitative information about the orientation of the spatial dispersion of the data set. The precision of computed gradients using exact or approximate interpolation methods generally depends on this spatial dispersion of the data. The statistical information given by the principal axes of this ellipse is then a useful tool to compare the spatial distribution of several dataset collected during different oceanographic campaigns. The differences in dimension and orientation of this ellipse can be an important cause of divergence between the results of spatial interpolation processes independently of the methodological approach chosen to perform it.

Any interpolation process is allow, after the assessment of the spatial distribution regarding the global density of the sample distribution. This can be achieved at different topological neighborhood order comparing the mean distance between each point and its k^{th} order neighbor.

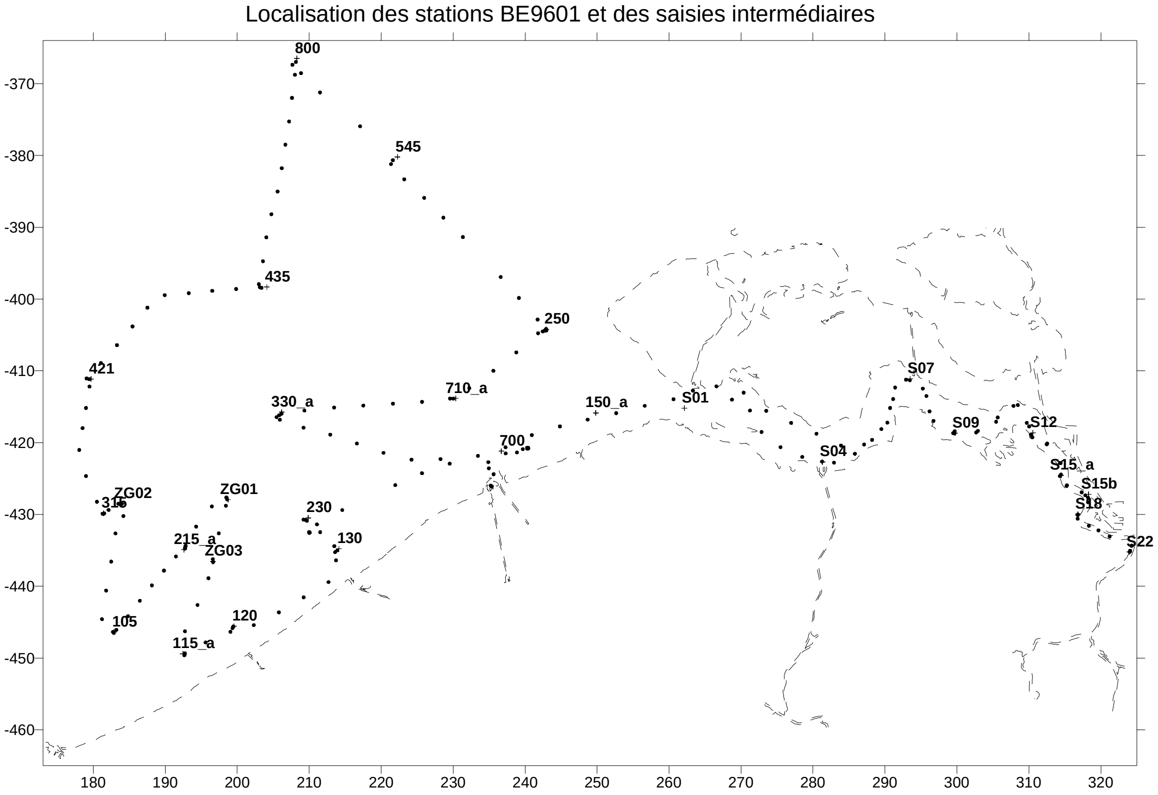
So, it is obvious that the evaluation of the precision of an interpolation process will always be biased because of the different scale of the phenomena influencing the measured parameter. All the following geostatistical computations including interpolation have been performed on the 9601 and 9626 campaign datasets collected for the main stations of the Scheldt Estuary and of the North Sea areas. The validation of the interpolation and other spatial analysis processes have been performed using reference and more dense dataset coming from the same campaigns at a sampling time interval of 10 minutes along the ship track. The results have discussed keeping in mind the above observation on the spatial and time scale of the influencing phenomena.

These spatial analysis methods are presented, applied and discussed here after. We firstly discuss the spatial distribution of the data, mainly water temperature and salinity measured during two campaigns, the campaigns 9601 and 9629. This discussion shows the necessity to consider independently the analysis of the continental shelf shallow water area and the Scheldt estuary area. We then describe some methods suggested to analyse the Scheldt estuary area and discuss the first purchased results. Thereafter, we explain and discuss some other methods for assessing spatial variability of salinity and temperature in the continental shelf shallow water area. Directional semivariograms have been computed and analysed and some interpolation processes applied there. For both areas, we have discussed the possibility to validate the methods using the data collected continuously along the ship track and stored at a 10 minutes time interval.

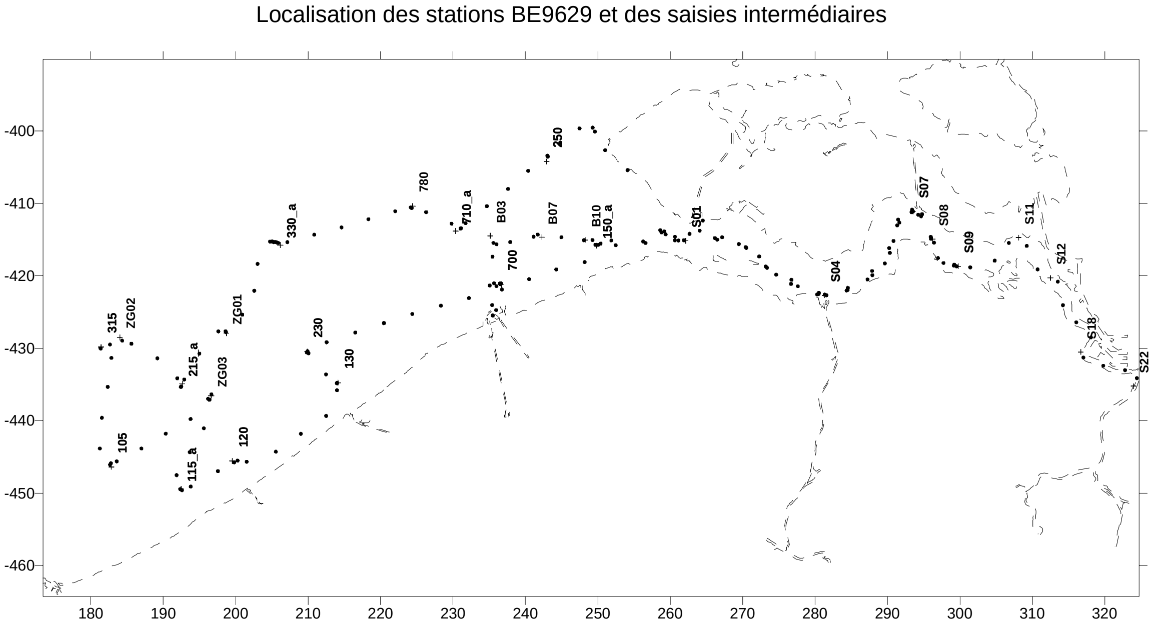
II. Spatial data distribution

On maps 1 and 2, we can observe that the spatial distribution of the data (measurements performed at the stations) is very different in the continental shelf shallow water area and in the Scheldt estuary. Nevertheless, it is quite difficult to define the position of the borderline between the two areas.

Map 1: Campaign 9601 – localization of the stations and of the intermediate acquisition points at a 10 minutes time interval along the ship track.



Map 2: Campaign 9629 - localization of the stations and of the intermediate acquisition points at a 10 minutes time interval along the ship track.



A. The continental shelf shallow water area

In the first area, a 2D distribution is perceptible. This kind of distribution is quantified for two campaigns computing some geostatistical parameters describing the mean location (gravity center) and the pattern of data dispersion (length and orientation of the main axes of the dispersion ellipse around this gravity center).

Moreover, the randomness of the spatial distribution is tested computing, for different defined neighborhood orders or "contiguity" levels, the ratio between the mean distance between the two points of each couple of points of the dataset at each level and the corresponding distance for randomly distributed points in the same area with the same observations density.

1. Gravity center and spatial dispersion of the data

The gravity center is defined by the mean position of the punctual observations of the spatial data set. The cartographic coordinates of the gravity center (X_g ,Y_g) are the weighted means of the x and y values of all the observations points (stations). In our case, it is not necessary to weight the different stations. The following simple arithmetic mean formula is then used to compute the gravity center coordinates:

$$X_g = \frac{\sum X_i}{n} \qquad Y_g = \frac{\sum Y_i}{n}$$

The dispersion around the gravity center can be quantified using the Bacchi standard distance. Nevertheless, this distance doesn't take in consideration the anisotropy of the spatial distribution. To assess this anisotropy, another method is commonly used. It is based on the computation of the orientations and dimensions of the mains axes of the dispersion ellipse.

The length of the main axes is proportional to the standard deviation along these axes. For a 95% probability ellipse, the length is equal to 2 times this standard deviation. The orientation of these axes is determined by the maximum dispersion.

To compute these parameters, the coordinates are normalized. The process is based on the signed distances along the *x*- and *y*-axes of the cartographic reference system from the gravity center to the different points of the data set. This computation depends then on the characteristics of the chosen cartographic projection and the dimensions of the area. In our case, the dimensions are small enough to neglect scale variations along both reference axes. This hypothesis is also acceptable for the construction of semivariograms and interpolation process.

The orientations of the main axes are computed in two steps. The first step consists in the determination of the dispersions (variances) perpendicular to the x (s^2_x) and the y (s^2_y) axes and of the covariance C.

$$s_x^2 = \frac{1}{n}\sum y_i^2 \qquad s_y^2 = \frac{1}{n}\sum x_i^2 \qquad C = \frac{1}{n}\sum x_i y_i$$

The dispersion perpendicular to an axe that makes an angle α with the *x*-orientation can be computed using the next equation:

$$s_\alpha^2 = \frac{1}{n}\sum (y_i \cos\alpha - x_i \sin\alpha)^2$$

$$s_\alpha = s_x^2 \cos^2\alpha + s_y^2 \sin^2\alpha - 2C \sin\alpha \cos\alpha$$

The equation of the axe s_α is given by:

$s_\alpha \equiv y = tg\alpha x$

The dispersion perpendicular to this axe is expressed as a sum of distances. The formulation of the distance in this perpendicular direction is the following one:

$d(x,y) \text{–>} s_\alpha = \cos\alpha(y - tg\alpha x)$

$d(x,y) \text{–>} s_\alpha = \cos\alpha y - \sin\alpha x$

To find the orientations of the axes of maximum and minimum dispersion, we have to find the values of α that maximize or minimize the dispersion $s^2{}_\alpha$. This is performed annulling the first derivative of the function. This is mathematically expressed by:

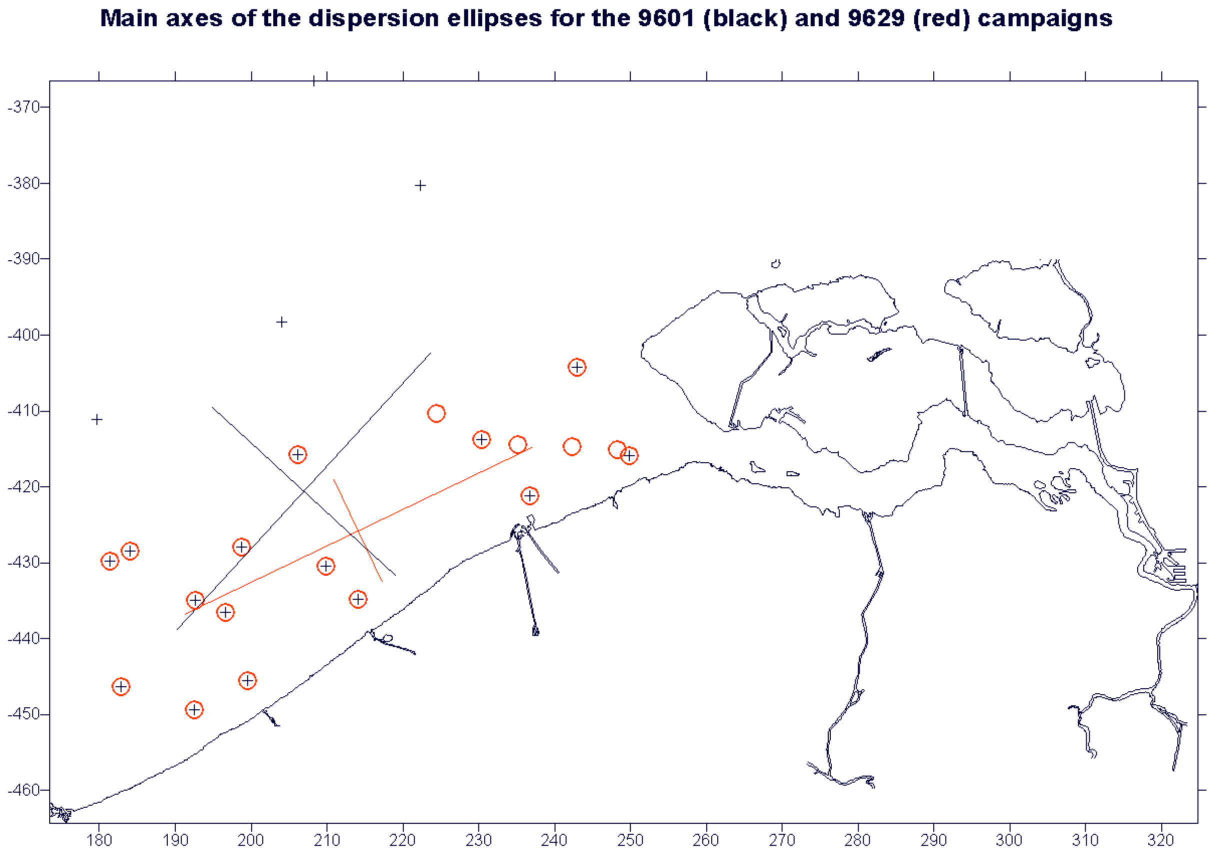
$(s_y^2 - s_x^2)\sin 2\alpha - 2C\cos 2\alpha = 0$

We can then find the 2 values of the angle α that must verify the following equation:

$tg2\alpha = \frac{2C}{\left(s_y^2 - s_x^2\right)}$

The method described here above allows the computation of the position, the orientation and the dimensions of the dispersion ellipse. It has been applied to the sea stations positions of the 9601 and 9629 campaigns (Map 3). This dispersion gives quantitative information about the reliability of the estimated gradients in function of the direction if an exact interpolation process is applied or a trend surface is adjusted on the observation points

Map 3: Spatial distribution of the data for the 9601 and 9629 campaigns – gravity center and main axes of the dispersion ellipses (1 σ, probability ≈ 68%)



The map 3 shows a clear difference between the two dispersion patterns. For the campaign 9601, the distribution is nearly isotropic and the gradients in all directions should then be equally reliable. For the other campaign, the dispersion is clearly anisotropic with a great dispersion along the coast and a rather small perpendicular dispersion.

Nevertheless the precision of interpolation processes is also a function of the density of the data points and of its variability in space with eventual concentration zones. These aspects are discussed hereafter, but the vicinity of points expresses the density. This is studied using neighborhood statistical parameters.

2. *Neighborhood statistics*

Neighborhood statistics are generally based on distance measurements between couple of points in function of a neighborhood order *k*. The concept of neighborhood order is similar to the topological concept of contiguity level for polygons entities. The typical index based on this mean distance estimation and used to assess the neighborhood is the *R*-statistic.

As told before, the neighborhood can also be put in relation with the density variability in space. This aspect is generally estimated using the quadrats method. In our case, we have used the *R*-statistic index to describe the spatial arrangement of the observation points.

This index corresponds to the ratio between two mean distances computed for a defined contiguity order level *k*. *R* is then a function of *k* (*R(k)*). For *k* = 1, for instance, the distances between each point and its nearer neighbor are computed. The mean of all these values is then calculated. This mean distance value is the numerator of the ratio *R(k),* with *k* = 1. The denominator of this ratio is the mean distance for the same value of *k* that would correspond to a set of points randomly distributed in the same area with a same density.

The mean distance expected for a random distribution is computed using the Poisson law:

$$\rho(k)=\frac{(2k)!k}{\left(2^k k!\right)^2\sqrt{d}}$$

In this equation, *d* is the density. It is computed dividing the number of points by the area of the analyzed territory. This area is the one of the polygons that contains all the data. It is constructed on the external data points.

If the *R(k)* ratio is 1, the points can be considered as randomly distributed in space. If it is greater than 1, then the points are dispersed. If it is lower, than the points are concentrated in space. The best condition to perform an interpolation is the random or the dispersed distribution.

To compare the value of *R(k)* to 1, it is necessary to perform a test at a defined confidence level that determines the confidence interval around 1. At a α probability level, the upper and lower limits of this interval are given by the next formula for *n* > 20 (*n* = number of points):

$$1{,}0 \pm z_\alpha \sigma\left[R(k)\right]$$

An empirical estimation of the standard error for a multiple set of simulated random distributions give the following values for *k* = 1 and 2:

$$\sigma\left[R(1)\right]=0{,}5228/\sqrt{n} \qquad \sigma\left[R(2)\right]=0{,}3630/\sqrt{n}$$

We must observe that the number of data points is generally smaller than 20 (n = 18, 19).

Furthermore, it is possible to correct the border effect, using a border buffer, without computing the distances from the points that fall in this area, but allowing the computation of the distances toward these points. This correction has not been applied. The results of this process are graphically reported and discussed here above.

Figure 1: R-statistic (campaign 9601) – for all the k values, R is significantly higher than 1 at a 95% statistical level. The points are then dispersed.

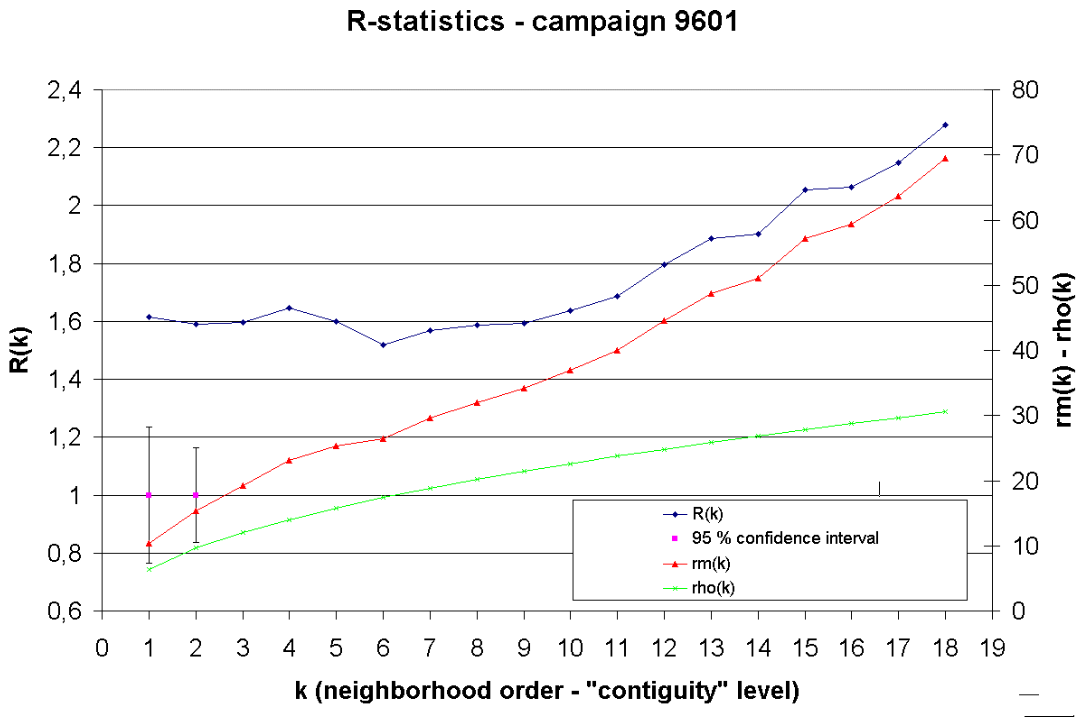
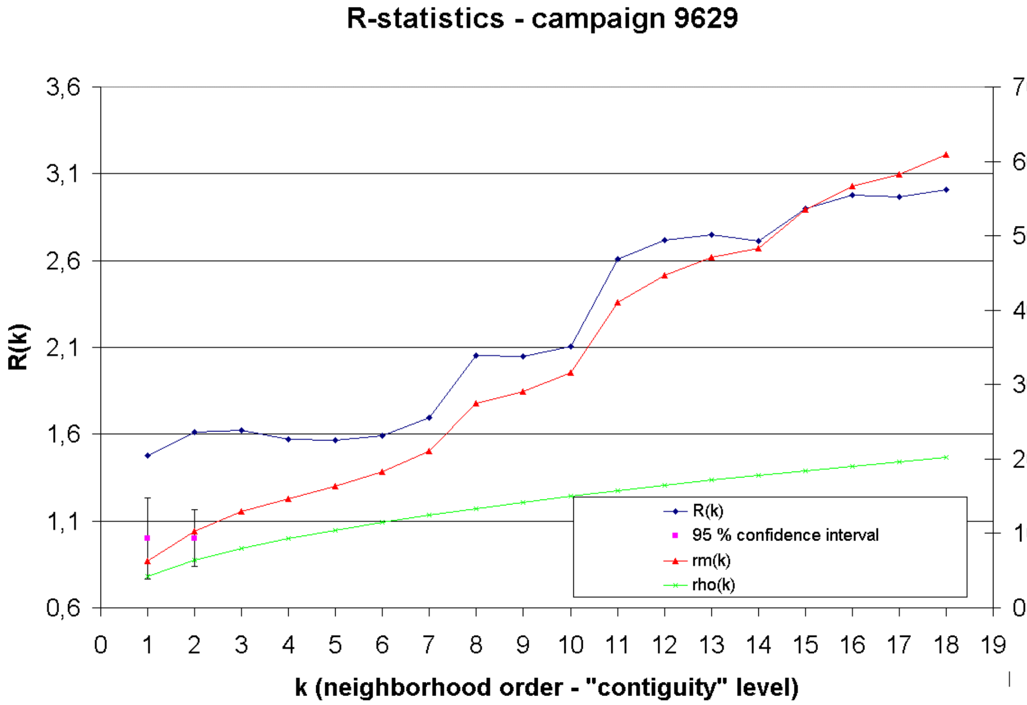


Figure 2: R-statistics (campaign 9629) – for all the k values, R is significantly higher than 1 at a 95% statistical level. The points are then dispersed.



For both campaigns, the spatial distribution of the points is dispersed. This means that, without considering the spatial variability of the considered parameter, the distances between points are very large to allow an accurate interpolation process.

In the case of the campaign 9601, there is no important break visible along the *R(k)* curve. This means that there are no significant clusters of points in space. In the case of the campaign 9629, two breaks are visible after the 7th and 10th *k*-levels. This means that the points are clustered in space. Two clusters are clearly identified on the map 3. The first cluster of 8 points is localized in front of the Zeebrugge harbour and 11 points that are clustered in front of Nieuwpoort and Oostend compose a second one. Such clusters are generally considered as a limitation factor for applying interpolation processes.

If we have a look at the general shape of the *R(k)* curve, the ideal number of points on which the interpolation process must be performed is smaller than 10, in the first case, and 7, in the second case. In the interpolation processes applied after, the number of 10 points has been chosen for both campaigns.

B. The Scheldt estuary area

As mentioned her above, in the Scheldt estuary, the spatial distribution of the observation points is organized along a line (maps 1 and 2). The spatial variability of the measured parameters must then be studied in a 1D framework. It is obvious that the cross variability estimation is not possible using such a data set and will than be neglected. The position of measurement stations for the 2 studied campaigns is the same. Nevertheless, the 10 minutes time interval sampling points show that the ship track is more variable in the western part of the estuary.

On the basis of the stations observations, the only spatial variability that can be quantified is the long-estuarine one. This can be established plotting the salinity or the temperature in function of a distance measured downstream toward the sea along the Scheldt from an origin point.

The position of the origin point has been arbitrarily defined the coordinates (323.800, -436.150), a few kilometers upstream the town of Antwerpen. This location determines the upstream border conditions of the model that put in relation distance and salinity or temperature and that are described here after.

The distance has been computed using a propagation algorithm. This algorithm works a in a raster mode accumulating the distance from a set of sources (origins) in function of the pixel resolution and taking into account the possible diagonal motion across these pixels and the weight of each pixel. In our case, the source grid consists in the vector to raster transformation of point cover containing the unique source or origin located in (323.800, -436.150). The allocation is then immediate and the distance is then unique. The pixels of the weight grid have the same value 1 in the water area. The propagation algorithm has been applied using a mask grid to allow the only motion along the Scheldt estuary. The working raster resolution is 0,125 km. The weight and the mask grids have been computed on the basis of the cover containing the coastlines in the Lorgna (Lambert Equal Area) cartographic reference system.

Nevertheless, some relatively great location incoherence’s have been stated comparing the 4 covers containing the sample points of the 2 campaigns (at the main stations and along the ship track with a 10 minutes sampling time interval). Some track points are in fact located in the continental area. We have then eroded the result of the vector to raster conversion of the polygons cover describing coastal lines. The erosion process has been performed using a spatial filtering method computing the focal maximum in a 3x3 convolution window. This process has produced a grid in which the water of the Scheldt has been connected with the water of the Scheldt-Rhine canal. The two areas have been disconnected adding a “dam” to the result of the filtering. Furthermore, using this propagation algorithm, the distance has been also computed upstream along the Scheldt, the canals flowing in the Zeebrugge and Oostende harbours, the Gent – Terneuzen canal and the Yser estuary. In this area, the resulting distance can’t be correlated to the salinity or the temperature. The mask has been modified to stop the distance propagation upstream along these canals and river.

Arcinfo processing to create the measurement points covers for the 9601 and 9626 campaigns

```
In the Arc module
generate c-9601 (stations cover)
generate c-9601tab (half an hour distance along track measurement points)
generate c-9629 (stations cover)
generate c-9629tab (half an hour distance along track measurement points)
build c-9601 points (creation of the point attribute table)
build c-9601tab points (creation of the point attribute table)
build c-9629 points (creation of the point attribute table)
build c-9629tab points (creation of the point attribute table)
addxy c-9601 point (creation of two coordinate control fields in the PAT)
addxy c-9629 point (creation of two coordinate control fields in the PAT)
addxy c-9629tab point (creation of two coordinate control fields in the PAT)
addxy c-9601tab point (creation of two coordinate control fields in the PAT)
```

Arcinfo processing to create the distance grid and its contouring

```
In the Arc module
dxfarc bcoast-lamb.dxf c-coast-lamb (importation of the coastal lines projected in the Lorgna
      reference system)
generate c-bordures (polygon created using the boundary coordinates of c-coast-lamb
ae (edition of c-bordures to merge c-bordures and c-coast-lamb)
clean c-coast-lamb (to create the polygon topology and build the polygon attribute table)
ae (edition of c-coast-lamb to modify the coastline position in function of the locations of
      the measurement points along the ship tracks of the 9601 and 9626 campaigns)
polygrid c-coast-lamb g-coast-lamb2 type
generate c-origine (creation of the cover containing the origin point located a few km
      upstream Antwerpen)
pointgrid c-origine g-origine # (rasterization)

In the Grid module
G-COAST-LAMB3 = focalmax ( g-coast-lamb2 )
```

```
G-MASK3 = select ( g-coast-lamb3 , 'value gt 0') (creation of a mask grid to not allow the propagation of distance from the origin in continental area)
Setmask (g-mask3) (remark : the setmask command can be used with the costdistance, costpath, idw, kriging, trend, linegrid, pointgrid and polygrid functions)
G-DISTANCE3 = costdistance ( g-origine , g-coast-lamb3 ) (distance grid computation)
```

In the Arc module

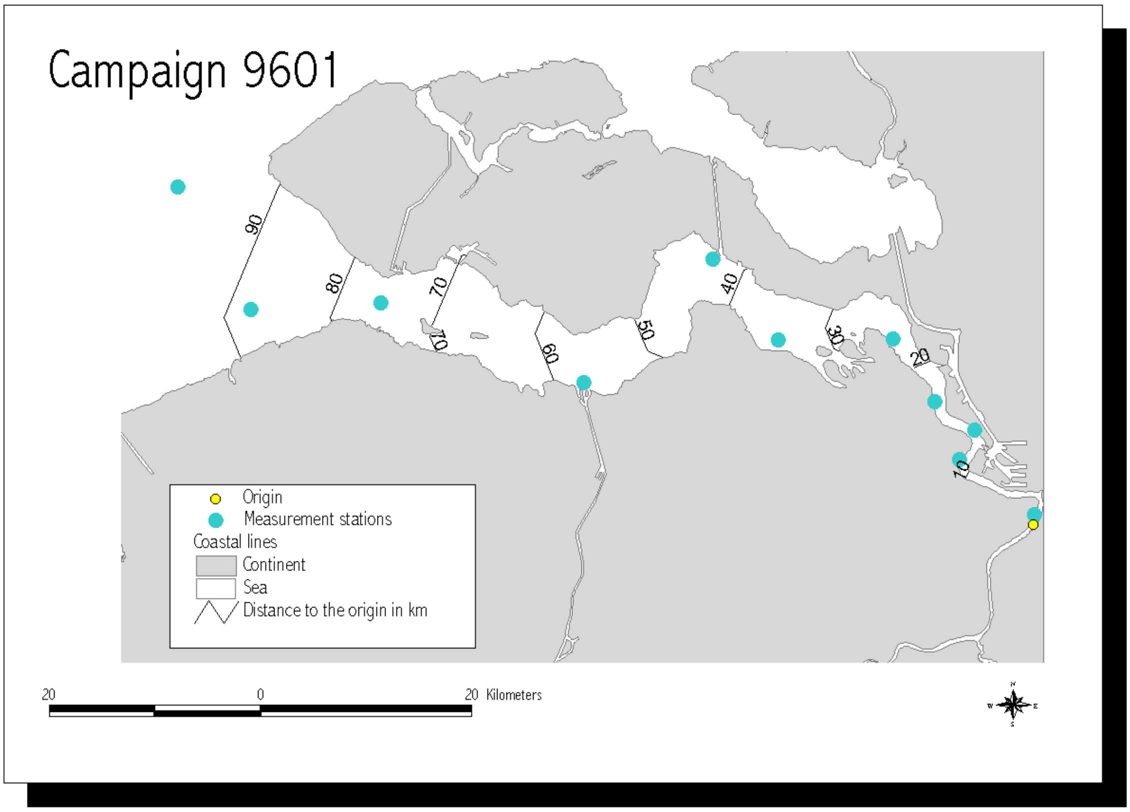
```
latticecontour g-distance3 c-distance3 10
clean c-distance3 # # # line
```

Arcinfo processing to compute the distance for each measurement point

In the Arc module

```
latticespot g-distance3 c-9601 distance3 (bilinear interpolation to extract the altitude at the measurement point in the distance grid)
latticespot g-distance3 c-9601tab distance3 (bilinear interpolation to extract the altitude at the measurement point in the distance grid)
latticespot g-distance3 c-9629tab distance3 (bilinear interpolation to extract the altitude at the measurement point in the distance grid)
latticespot g-distance3 c-9629 distance3 (bilinear interpolation to extract the altitude at the measurement point in the distance grid)
infodbase c-9601.pat 9601.dbf (exportation of the info file to the dbf > xls > sta format)
infodbase c-9601tab.pat 9601tab.dbf (exportation of the info file to the dbf > xls > sta format)
infodbase c-9629.pat 9629.dbf (exportation of the info file to the dbf > xls > sta format)
infodbase c-9629tab.pat 9629tab.dbf (exportation of the info file to the dbf > xls > sta format)
```

Map 4: Results of the contouring performed on the above computed distance grid using a mask grid



III. The Scheldt estuary

A. Water salinity in the Scheldt estuary

1. Campaign 9601

If we consider the first studied physical parameter of the water, the salinity, we can observe on figure 3 that it is strongly correlated with the distance from the origin. This statement is confirmed for the second studied campaign. We have then chosen to adjust a statistical model relating these variables.

a) Model definition and adjustment

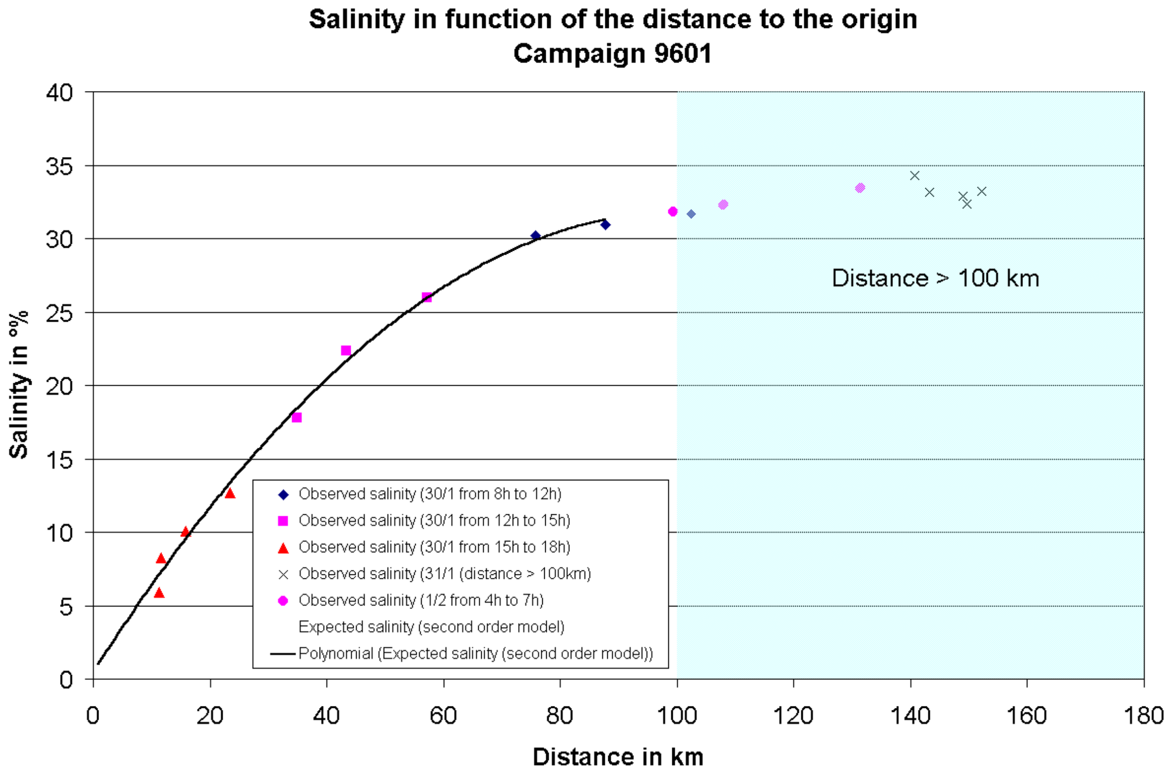
The salinity increases with the distance. From the origin toward the sea, the rate of increase goes down until a sill that is reached after 100 km (figure 3). There after, the rate of salinity increase is very low. We have then chosen to fit a second order function on the basis of the salinity measured at the Scheldt stations that are distant from the origin less than 100 km.

An iterative non linear method (quasi-Newtonian) has been used to adjust the coefficient of this function method. A convergence threshold for each coefficient has been fixed to stop the iterative process before a maximum number of iterations. The cost function to minimize is the sum of the square residuals. The results are given by the table I.

TABLE I

N = 10	Distance^2	Distance	Constant
Salinity (°%)	-0,0030895	0,6225063	0,451270
R	0,998	Explained variance	99,648 %

Figure 3: Salinity in function of the distance (campaign 9601)



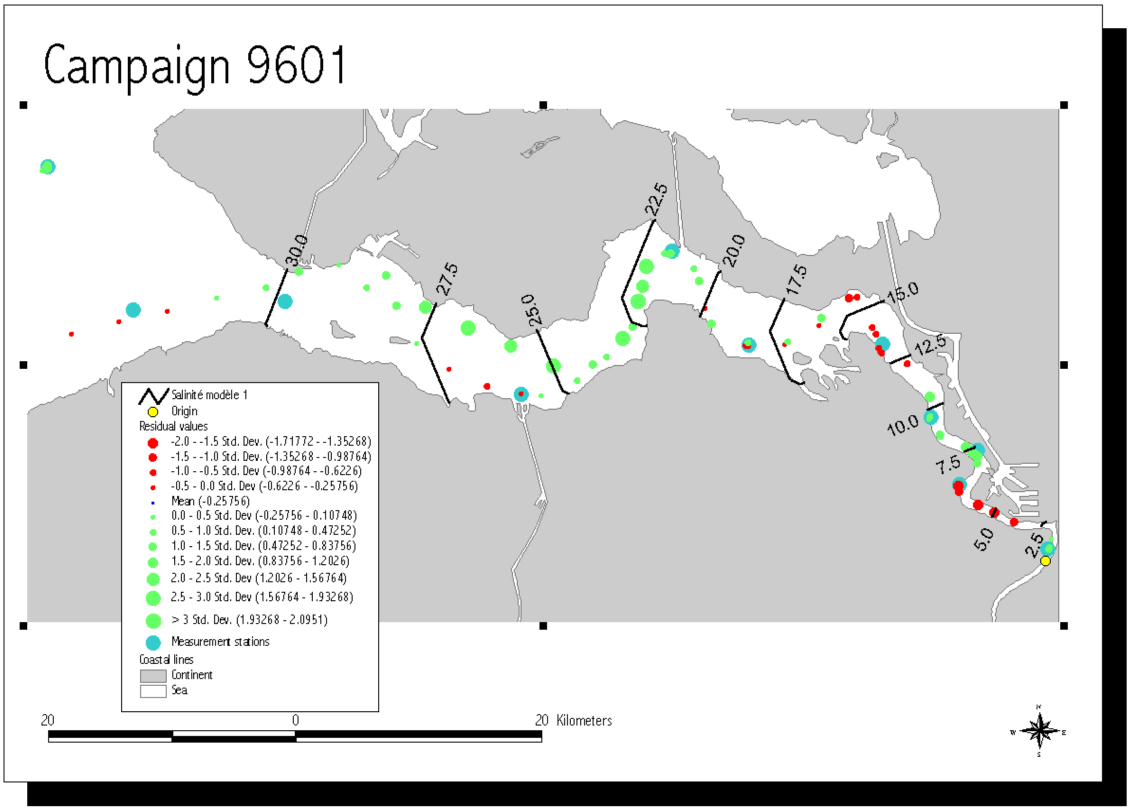
The model has then been used to estimate in salinity in function of the distance for each point (pixel) of the distance grid in the area of the Scheldt. The map 5 shows the result after the application of a contouring process with a 2.5 °% salinity interval.

This 1D process is similar to the 2D trend surface adjustment, but the second dimension has been neglected because of the lack of data across the Scheldt channel.

Arcinfo procedure to estimate the salinity values in function of the distance model using a grid surface modeling method and its contouring

```
In the Grid module
G-SAL-MOD1 = 0.451270 - 0.0030895 * sqr ( g-distance3 ) + 0.6225063 * ( g-distance3 )

In the Arc module
latticecontour g-sal-mod2 c-sal-mod2 2.5
clean c-sal-mod2 # # # line
```

Map 5: Salinity in the Scheldt estuary (campaign 9601) in function of the distance along the estuary and the residual values computed for the intermediate acquisition points at a 10 minutes sampling time interval along the ship track


Campaign 9601
30.0
27.5
25.0
22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
Salinité modèle 1
Origin
Residual values
-2.0 - -1.5 Std. Dev. (-1.71772 - -1.35268)
-1.5 - -1.0 Std. Dev. (-1.35268 - -0.98764)
-1.0 - -0.5 Std. Dev (-0.98764 - -0.6226)
-0.5 - 0.0 Std. Dev (-0.6226 - -0.25756)
Mean (-0.25756)
0.0 - 0.5 Std. Dev (-0.25756 - 0.10748)
0.5 - 1.0 Std. Dev. (0.10748 - 0.47252)
1.0 - 1.5 Std. Dev. (0.47252 - 0.83756)
1.5 - 2.0 Std. Dev. (0.83756 - 1.2026)
2.0 - 2.5 Std. Dev (1.2026 - 1.56764)
2.5 - 3.0 Std. Dev (1.56764 - 1.93268)
> 3 Std. Dev. (1.93268 - 2.0951)
Measurement stations
Coastal lines
Continent
Sea
20
0
20 Kilometers
N
W
E
S

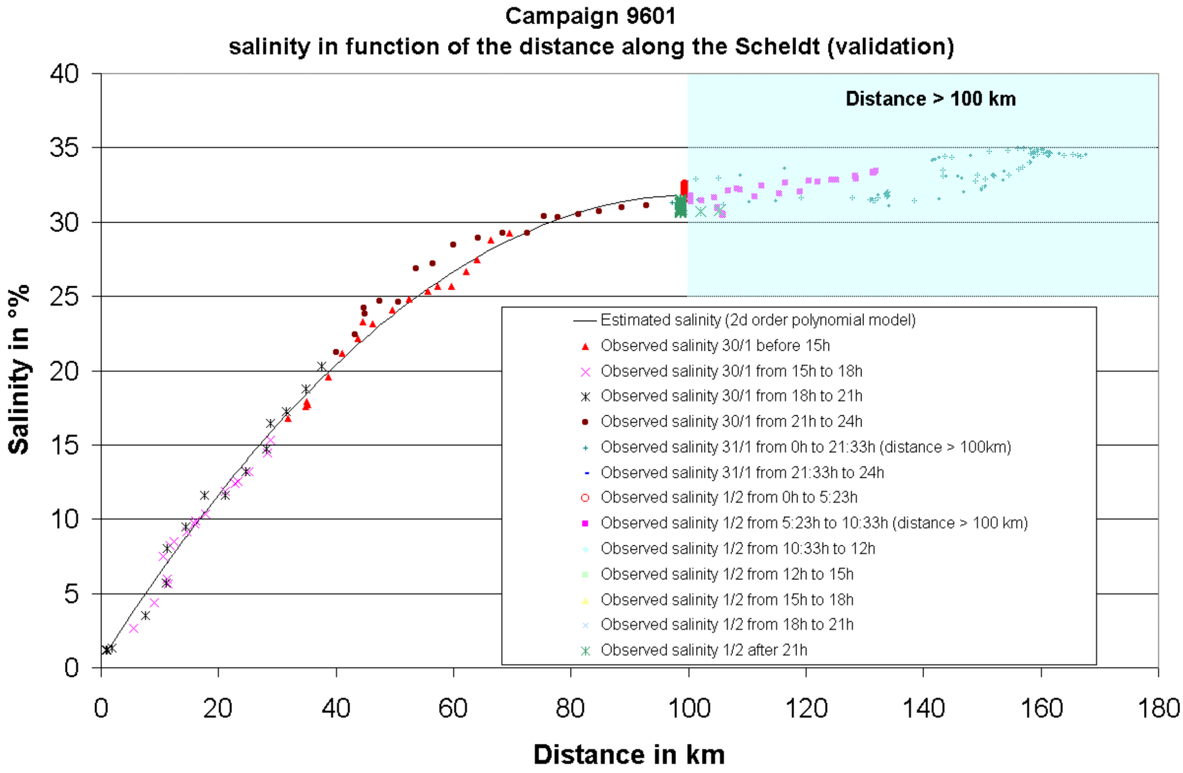
b) Model "validation"

The intermediate points data set with a sampling time interval of 10 minutes along the ship track have been used to validate the model. Nevertheless, on the scatter plot the salinity – distance (figure 4), if we take in consideration the acquisition dates and times for all these points, we state that all the measurements are not comparable on the temporal point of view. This observation is more deeply discussed for the campaign 9629. The magnitude of the residual values can then not be considered as a true validation process because it is affected by a local-temporal effect caused by the following phenomena:

- tide coefficient, meteorological conditions and, with a possible negligible effect in comparison with the tidal influences, hydrological conditions of the continental Scheldt;
- acquisition time and localization in relation with the tidal front propagation along the estuary.

These phenomena acting at two different spatio-temporal scales should be taken in consideration to correct the salinity values before adjusting the model. Nevertheless, we can observe on the figure 4 that the distribution of the residuals values is quite homogeneous. Furthermore, there is a strong autocorrelation effect between the successive measurements.

Figure 4: Scatter plot of the salinity measurements for the 10 minutes along track acquisitions in function of the distance and the second order polynomial function adjusted using the Scheldt stations measurements (campaign 9601)



2. *Campaign 9629*

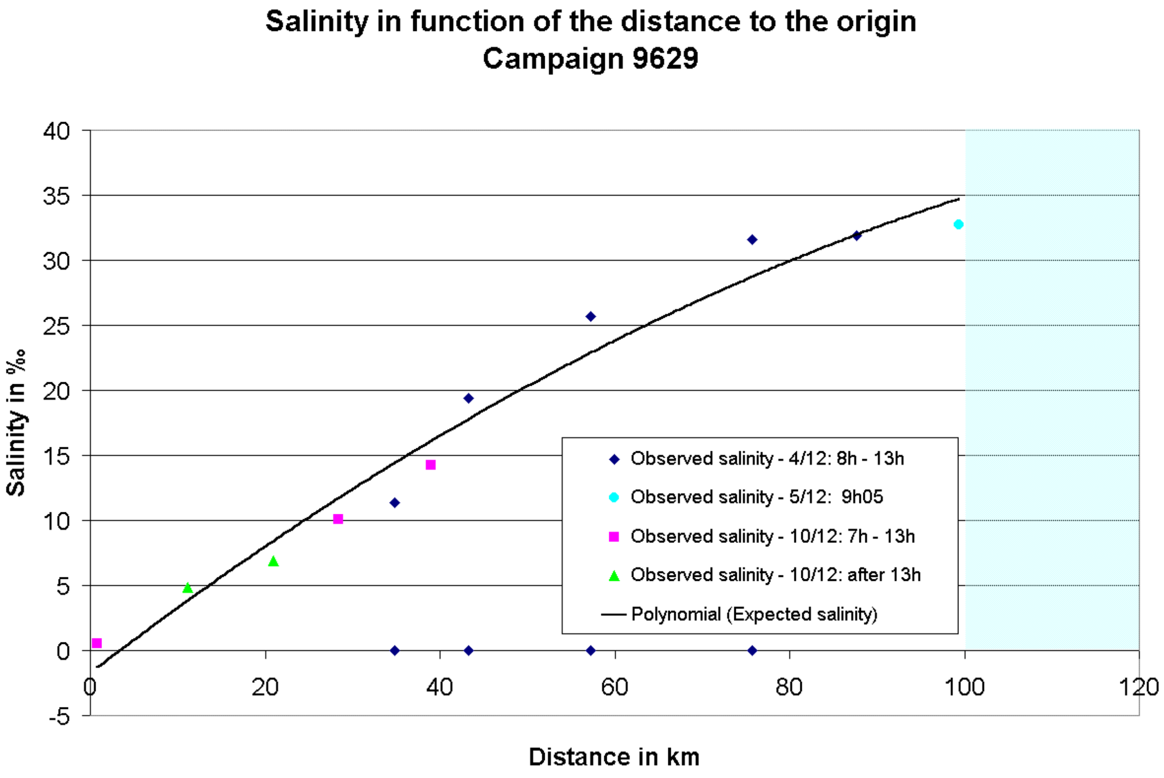
a) Model adjustment

The same kind of model has been adjusted using the Scheldt stations measurements for the campaign 9629. The results are shown on the table II and the figure 5.

TABLE II

N = 11	Distance^2	Distance	Constant
Salinity (°%)	-0,00144965	0,5156169	-1,742669
R	0,984	Explained variance	96,755 %

Figure 5: Salinity in function of the distance (campaign 9629)



The negative value of the intercept has no physical signification. It is obvious that this constant depends on the distribution of the data along the distance axe and of the salinity gradient along this axe that is influenced by the factors mentioned here above. Furthermore, the coefficient of the second order term is influenced by the density of the data points along the Scheldt estuary and on eastern and western sides of this estuary. A deeper analysis of the salinity variation on both sides of the graph should allow the definition of the border conditions to choice another formulation of the relationship between salinity and distance. In fact, it is well known that the seawater influence nearly reaches Gent. This means that the salinity can remain small but positive until this town during high tide period. For this campaign, this induces an upward concave aspect of the salinity trend towards the origin.

b) Model "validation"

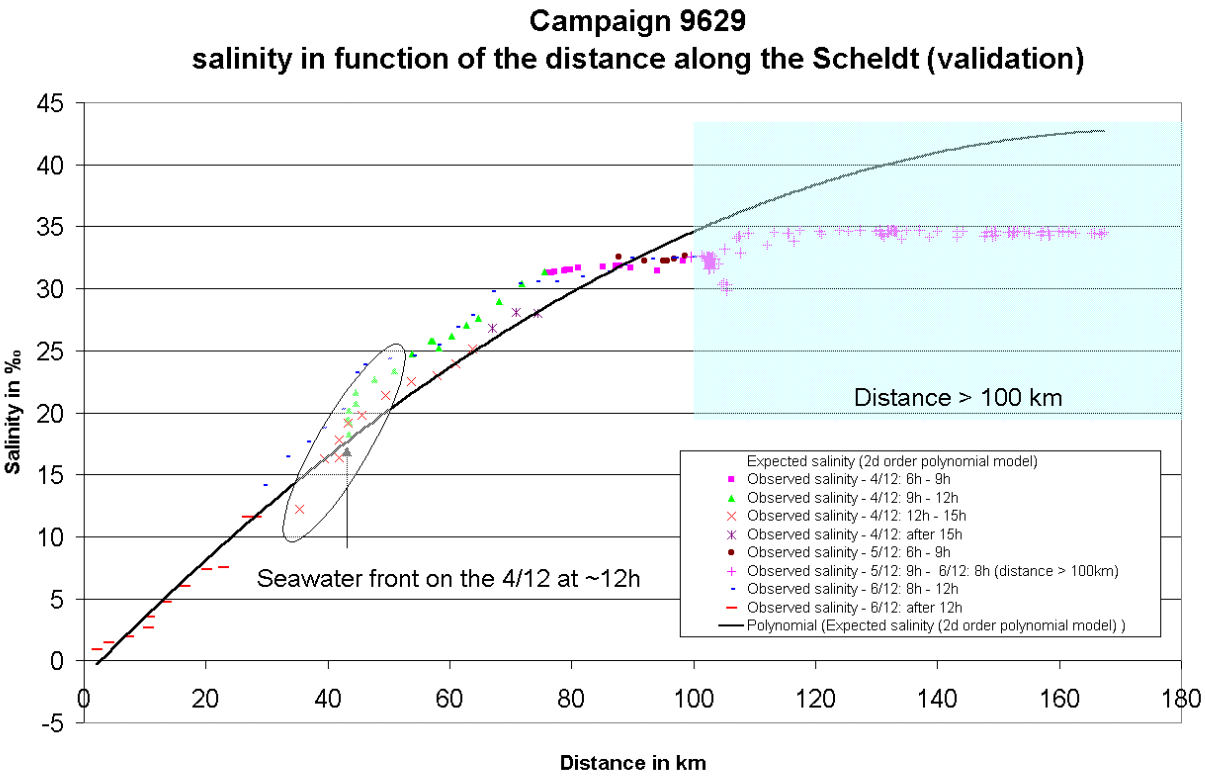
If we analyse the residual values in function of the distance (figure 6), we notice that they show a cyclic function with 0 values after ~15km, ~40 km and ~90km and the maximum gradient of these values is observed at ~40 km. The former statements suggest the apparent necessity to use a higher order polynomial function. This should justify the choice of third order polynomial model but the very small number of points to adjust the function doesn't allow its use.

Moreover, the acquisition dates and time of the data points are not comparable for all the points. For instance, if we consider the measurement from 8h till 13h of the 4/12, we can notice that the maximum rate of salinity variation in function of an increasing distance is localized between ~35km and ~45km. This is also visible on the figure 5. This rate is not so high for the first studied campaign. Nevertheless, on figure 6, it is obvious that the second order polynomial trending function fits well on the "full" data set. It is then not recommended to use any exact 1D or 2D interpolation method (IDW, linear or quintic TIN or kriging) to map the salinity “surface” in the Scheldt Estuary.

Between 40 and 60 km, the measured salinity after 12h on the 4/12 is lower than the one measured before. This could be partially explained by the effect of the different track position across the Scheldt but it can also be explained by the time difference between the acquisition moment and the upstream propagation of the tide along the Scheldt at the acquisition place. So the computation of the residual values doesn’t give true information about the precision of the adjusted spatial model.

The magnitude and trend of the salinity when the distance decreases, is influenced by the same factors then those ones above discussed. The magnitude is probably correlated with the importance of the discharge of propagation of marine water. The discharge of marine water depends on the tide coefficient and of the moment of salinity measurement with regard to the tidal state at this. As discussed here after, the position of the front of seawater fluctuates in a cyclic way along the distance axe. Neglecting the diffusion process that are probably impossible to determine using only the measurements performed at the stations, the salinity gradient through this front and the wavelength of the position fluctuation must be compared for the whole campaigns data set to explain the spatial distribution of the residual values along this axe.

Figure 6: Scatter plot of the salinity measurements for the 10 minutes along track acquisitions in function of the distance and the second order polynomial function adjusted using the Scheldt stations measurements (campaign 9629)



B. Water temperature in the Scheldt estuary

1. Campaigns 9601 and 9629

We have also briefly analysed the temperature variation along the Scheldt estuary to assess the possibility to apply the same kind of modelling for this parameter than the one exploited for the salinity. Unfortunately, the temperature - distance scatter plot is not so clearly structured as the salinity one. Moreover, it is probably more variable from one campaign to the other one, if we compare the plots of the 9601 and 9629 campaigns (figures 7 and 8).

The relationship between temperature and distance is quite obvious for the campaign 9601. This campaign has been realized in typical winter conditions. Relatively low water temperatures in the continental water area and higher ones in open sea characterize these conditions. So the effect of the urban and industrial area of Antwerpen is more visible until 20 km. From ~20 km until ~45 km, there is a decrease of the temperatures. After this distance until ~100 km, the temperature is less variable in space. After ~100km, the temperature of the water increases with the distance but the correlation is not significant because of the important dispersion. We are coming in the continental shelf shallow water area. This dispersion suggests the necessity to study the 2D variability. The mean temperature of this area is higher than the temperature of the estuary.

In the case of the campaign 9629, the variability of the temperature in our one-dimensional space is very small. There is no trend on the temperature along the Scheldt. There is only one station with a probable abnormally high value at the station S18.

If we compare the distribution of the 10 minutes time interval temperature samplings in function of the distance (figure 9), we can say for the case of the campaign 9601 and for a distance less than 100 km that the acquisition moment and track position across the Scheldt estuary has no effect on the cartographic process at the scale of the estuary length during the whole campaign. The temperature measured at the stations is sufficient to estimate intermediate values with a good precision regarding the temperature variability in the Scheldt estuary. A 1D interpolation process could then be applied on the basis of a temperature - distance array.

If we consider the same set of intermediate points along the ship track of the campaign 9629 (figure 10), it is possible to confirm the general trend stated for the stations measurements. But it is also obvious that the moment of the measure with regard to the tidal situation at the measurement date and time are more perceptible. This is particularly perceptible for the temperature measured on the 4/12 between 9 and 12h, on the 4/12 between 12 and 15h and on the 6/12 after 12h. In this case, if an exact interpolation process is performed on the basis of the stations temperature measurements in this area (between 25 and 70 km), then the residual values would be quite high in comparison with the temperature variability along the estuary. A shift along the *x*-axe of the different series of measurements is well visible. Nevertheless, the general shape of the temperature – distance distribution in the Scheldt estuary is very similar for the 2 campaigns.

Figure 7: Temperature - distance scatter plot along the Scheldt estuary for the campaign 9601

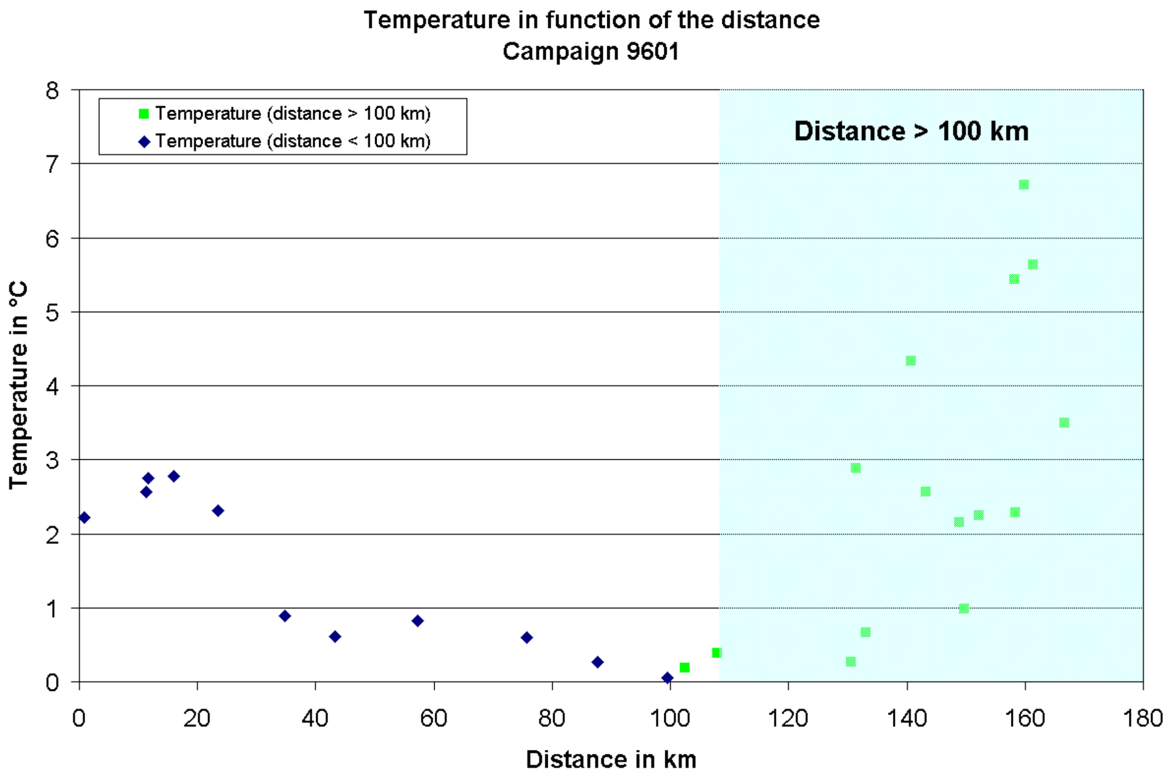


Figure 8: Temperature - distance scatter plot along the Scheldt estuary for the campaign 9629

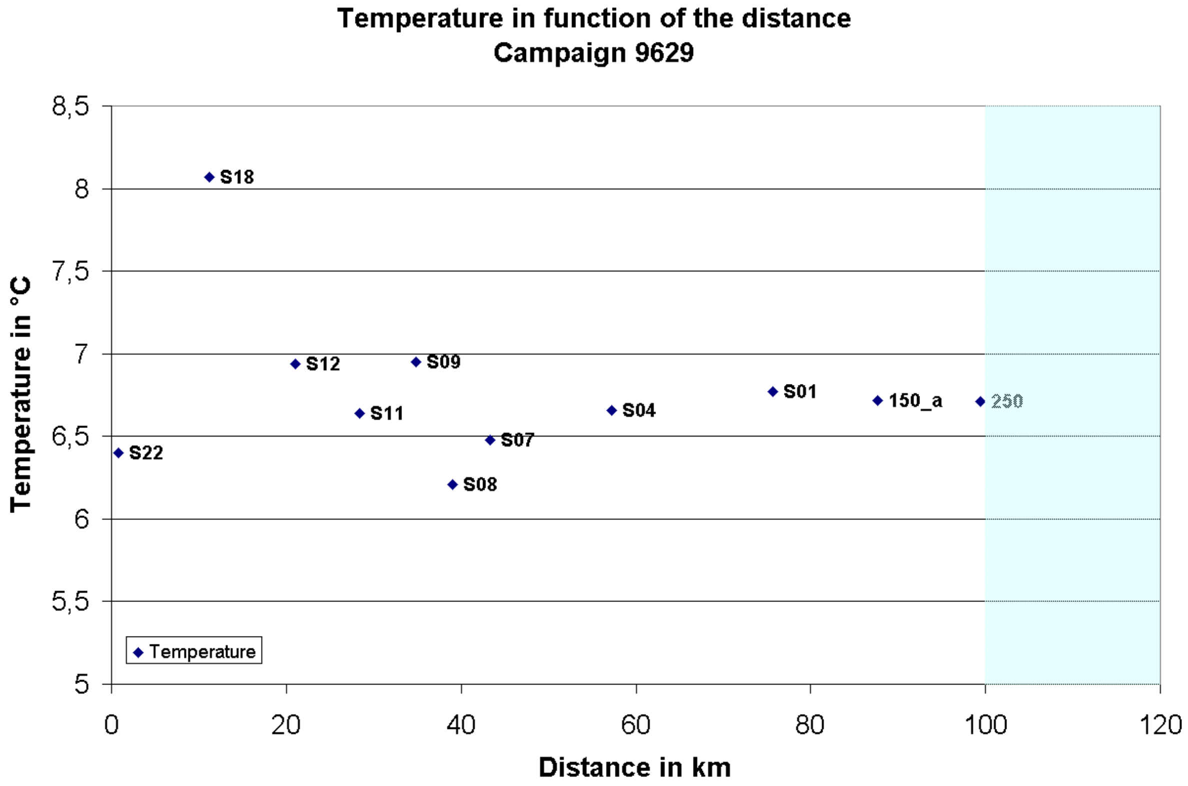


Figure 9 Scatter plot temperature in function of the distance for the 10 minutes time interval samplings along the ship track (campaign 9601)

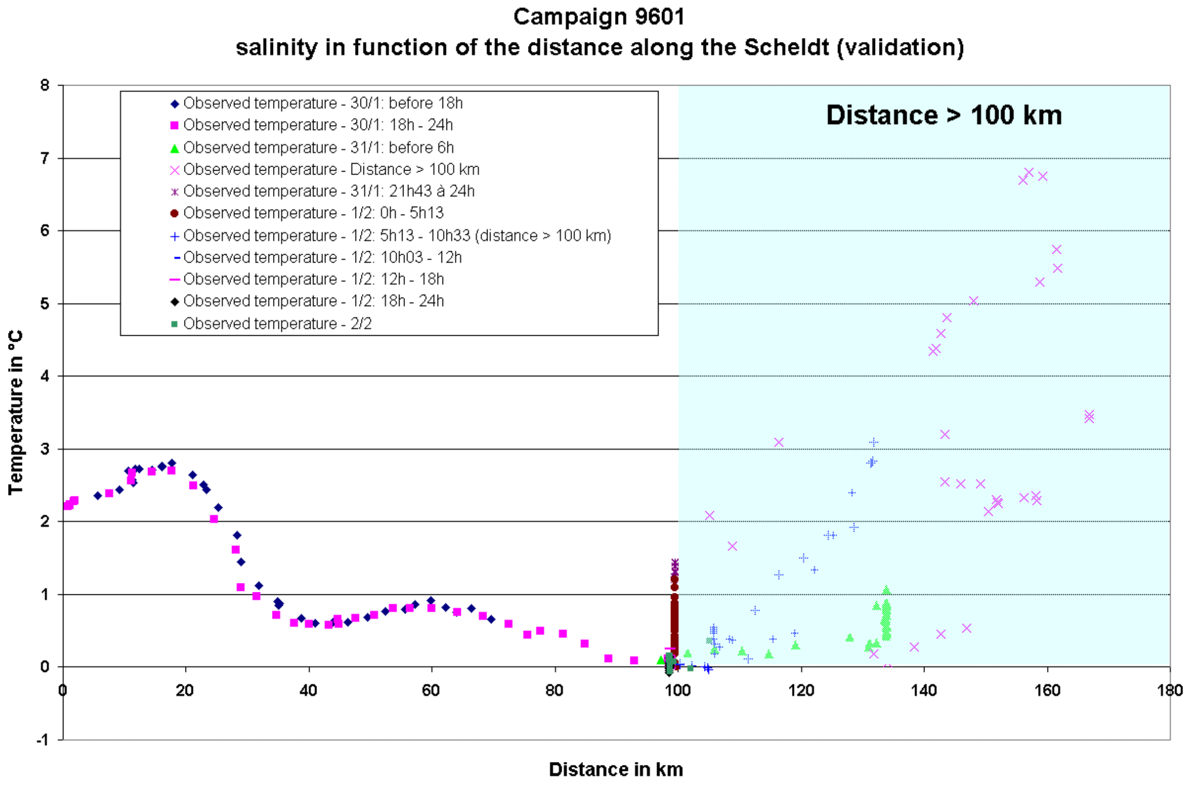
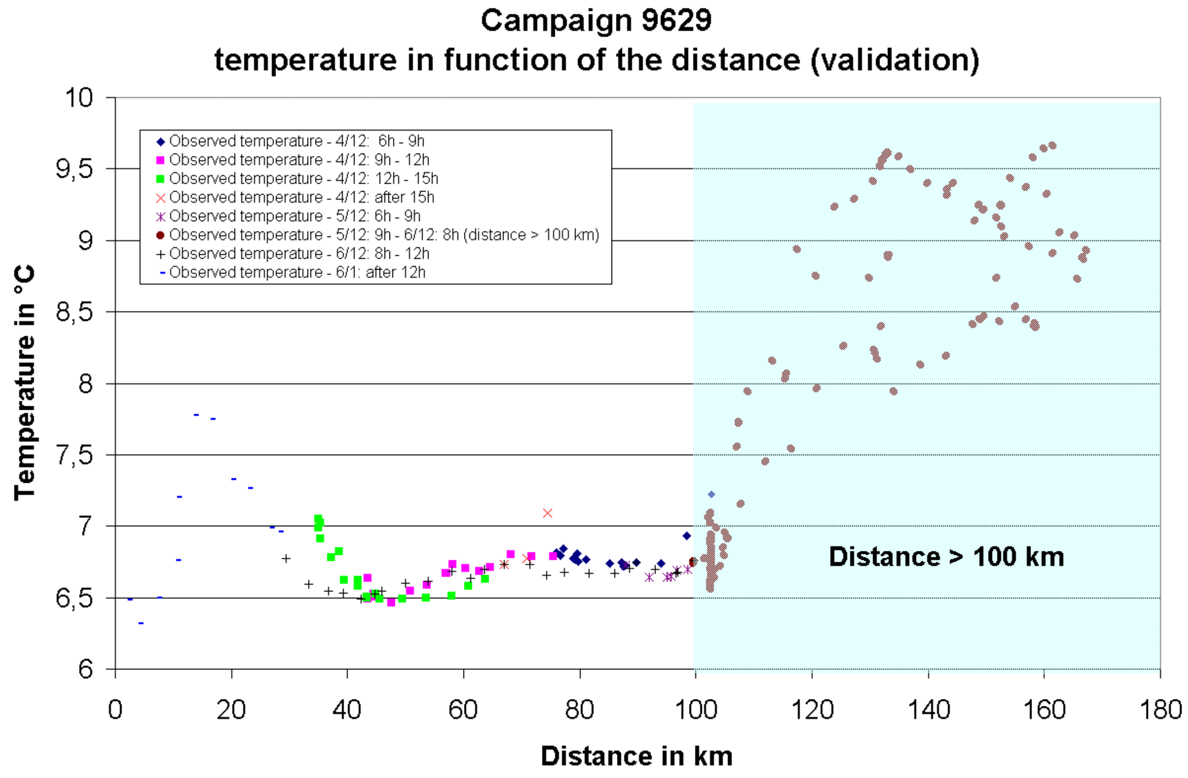


Figure 10: Scatter plot temperature in function of the distance for the 10 minutes time interval samplings along the ship track (campaign 9629)



IV. The continental shelf shallow water area

In this area, two kinds of surface mapping, exact interpolation, using kriging methods for instance, and trend surface, can be applied. Nevertheless, as demonstrated here above, exact interpolation process must be performed carefully because of the very high dispersion of the data points.

Before applying the kriging method, the spatial variability must be studied in order to choose correct parameter values in the kriging process. The construction of omnidirectional semivariograms has been described in former reports and the first statement concern the very little size of the data sets which doesn’t give well structured plots. In spite of this, we have developed a processing to compute directional semivariograms. The method is described by Isaaks & Srivastava (1989) and the processing has been performed using the next Statistica Basic programs. Nevertheless, Surfer 7 gives the dimensions and the orientation of the two axes of the ellipse fitted on the sill level visible on our computed directional semivariograms.

Statistica programs developed to compute the directional semivariograms on small data sets

```
RandomAccess;
Dim Donnees (100000,4);
Dim d(1);
Dim n(1);
Dim Aniso_n (10,10);
Dim Aniso_ns (10,10);
Dim Aniso (10,10);
Dim Aniso_deltax(10,10);
Dim Aniso_deltay(10,10);

DisplayNumericInputBox ('Distance de Recherche', 'D', d);
DisplayNumericInputBox ('Nombre de colonnes et de lignes', 'N paire et <= 10', n);
resolution:=(2*d(1))/n(1);

For i:=1 to n(1) do
      begin
      For j:=1 to n(1) do
        begin
        Aniso_deltax(i,j):=resolution*(i-0.5)-d(1);
        Aniso_deltay(i,j):=-resolution*(j-0.5)+d(1);
        {WriteLn (i,j,Aniso_deltax(i,j),Aniso_deltay(i,j))}
        end;
      end;

For i:=1 to NCases do
      begin
      donnees(i,1):=data(i,3);
      donnees(i,2):=data(i,4);
      donnees(i,3):=data(i,15);
      donnees(i,4):=data(i,18);
      {WriteLn(i,donnees(i,1),donnees(i,2),donnees(i,3),donnees(i,4))}
      {1 = x , 2 = y, 3 = Salinité; 4 = Température}
      end;

For i:=1 to NCases do
      begin
      for j:=1 to NCases do
        begin
        deltax:=(donnees(i,1)-donnees(j,1));
        deltay:=(donnees(i,2)-donnees(j,2));
        Distance:=sqrt(deltax**2+deltay**2);
        If (abs(deltax) < d(1)) and (abs(deltay) < d(1)) and (Distance<>0)  then
               begin
               semivar:=(donnees(i,4)-donnees(j,4))**2;
               colonne_id:=trunc((deltax+d(1))/resolution)+1;
               ligne_id:=trunc((-deltay+d(1))/resolution)+1;
               Aniso_n(colonne_id,ligne_id):=Aniso_n(colonne_id,ligne_id)+1;
               Aniso_ns(colonne_id,ligne_id):=Aniso_ns(colonne_id,ligne_id)+semivar;

        {WriteLn(deltax,deltay,semivar,colonne_id,ligne_id,Aniso_n(colonne_id,ligne_id),Aniso_n
      s(colonne_id,ligne_id))}
               end;
        end;
      end;

For i:=1 to n(1) do
      begin
      for j:=1 to n(1) do
        begin
        if Aniso_n(i,j)>0 then Aniso(i,j):=Aniso_ns(i,j)/(2*Aniso_n(i,j));
        end;
      end;
MatrixDisplay (Aniso, 'Semivariance');
MatrixDisplay (Aniso_n, 'Nombre de cas');
MatrixDisplay (Aniso_deltax, 'X_ Centre');
MatrixDisplay (Aniso_deltay, 'Y_Centre');

For i:=1 to n(1) do
      begin
      For j:=1 to n(1) do
        begin
        if Aniso_n(i,j) > 0 then
                begin
                WriteLn(Aniso_deltax(i,j),Aniso_deltay(i,j),Aniso_ns(i,j),
      Aniso_n(i,j),Aniso(i,j));
                end;
        end;
      end;
```

Statistica programs developed to compute the directional semivariograms on big data sets

```
RandomAccess;

Dim d(1);
Dim n(1);
Dim ncas(1);
Dim aniso_n(10,10);
Dim aniso_nsh(10,10);
Dim aniso_nsd(10,10);
Dim aniso_deltax(10,10);
Dim aniso_deltay(10,10);
Dim aniso_h (10,10);
Dim aniso_d (10,10);

ValMax (v1, 1, NCases, azimutmax );
ValMax (v2, 1, NCases, rangemax );
ValMin (v1, 1, NCases, azimutmin );
ValMin (v2, 1, NCases, rangemin );
largeur:=rangemax -rangemin;
hauteur:=azimutmax -azimutmin;
densite:=NCases/(largeur*hauteur);

WriteLn('Coordonnée maximum en azimutal :', azimutmax);
WriteLn('Coordonnée maximum en range : ', rangemax);
WriteLn('Coordonnée minimum en azimut :', azimutmin);
WriteLn('Coordonnée minimum en range :', rangemin );
WriteLn('Extension en azimut : ', hauteur);
WriteLn('Extension en range : ' , largeur);
WriteLn('Densité moyenne estimée : ', densite);
WriteLn('Si la hauteur est supérieure à la largeur,');
WriteLn('            le fichier de données doit être trié par Y croissant');
WriteLn('Si la largeur est supérieure à la hauteur,');
WriteLn('            le fichier de données doit être trié par X croissant');
WriteLn('Si ce n"est pas le cas, arrêter le programme,');
WriteLn('            trier les données et relancer le programme');

DisplayNumericInputBox ('Distance de recherche', 'd', d(1));
DisplayNumericInputBox ('Nombre de cellules', 'n', n(1));
DisplayNumericInputBox ('Nombre de cas à traiter pour le test ou nombre total de cas dans le
      fichier', 'Nombre de cas', ncas(1));

if  (hauteur  >  largeur)  then  nparmaille:=trunc((NCases*2*d(1)*largeur/(largeur*hauteur))+2)
      else nparmaille:=trunc((NCases*2*d(1)*hauteur/(largeur*hauteur))+2);

resolution:=2*d(1)/n(1);

WriteLn('Distance de recherche : ', d(1));
WriteLN('Nombre de points par cellule : ', nparmaille);
```

```
For i:=1 to n(1) do
      begin
      For j:=1 to n(1) do
        begin
        aniso_deltax(i,j):=resolution*(i-0.5)-d(1);
        aniso_deltay(i,j):=-resolution*(j-0.5)+d(1);
        end;
      end;

fin:=ncas(1)-trunc(nparmaille/2);

For i:=1 to fin do
      begin
      debut:=i;
      For j:=(i+1) to (i+trunc(nparmaille/2)) do
        begin
        deltarange:=data(i,2)-data(j,2);
        deltaazimut:=data(i,1)-data(j,1);
        distance:=(sqrt(deltarange**2+deltaazimut**2));
        if ((distance<d(1)) and (distance>0)) then
                begin
                semivar_h:=(data(i,3)-data(j,3))**2;
                semivar_d:=(data(i,4)-data(j,4))**2;
                colonne_id:=trunc((deltarange+d(1))/resolution)+1;
                ligne_id:=trunc((-deltaazimut+d(1))/resolution)+1;
                aniso_n(colonne_id,ligne_id):=aniso_n(colonne_id,ligne_id)+1;
                aniso_nsh(colonne_id,ligne_id):=aniso_nsh(colonne_id,ligne_id)+semivar_h;
                aniso_nsd(colonne_id,ligne_id):=aniso_nsd(colonne_id,ligne_id)+semivar_d;
                {if colonne_id>5 then WriteLn(i,j,colonne_id,ligne_id,deltarange,deltaazimut);}
                end;
        end;
      {WriteLn(i,j);}
      end;

For i:=1 to n(1) do
      begin
      for j:=1 to n(1) do
        begin
        if aniso_n(i,j)>0 then aniso_h(i,j):=aniso_nsh(i,j)/(2*aniso_n(i,j));
        if aniso_n(i,j)>0 then aniso_d(i,j):=aniso_nsd(i,j)/(2*aniso_n(i,j));
        end;
      end;
{MatrixDisplay (aniso, 'Semivariance');}
{MatrixDisplay (aniso_n, 'Nombre de cas');}
{MatrixDisplay (aniso_deltax, 'X_ Centre');}
{MatrixDisplay (aniso_deltay, 'Y_Centre');}

For i:=1 to n(1) do
      begin
      For j:=1 to n(1) do
        begin
        if aniso_n(i,j) > 0 then
                begin

        WriteLn(aniso_deltax(i,j),aniso_deltay(i,j),aniso_nsh(i,j),aniso_nsd(i,j),aniso_n(i,j),
      aniso_h(i,j),aniso_d(i,j));
                end;
        end;
      end;
```

Here are the results of this processing applied on the data set of the 9601 campaign: salinity and temperature measurements in the shelf area.

A. Campaign 9601

1. Semivariograms

Figure 11: Directional salinity semivariogram (continental shelf area stations – campaign 9601). Legend: red numbers: semivariance computed in each cell (15 km x 15 km); blue numbers: number of points in each cell; isolines: iso-semivariance curves (equidistance 0.5 ppm).

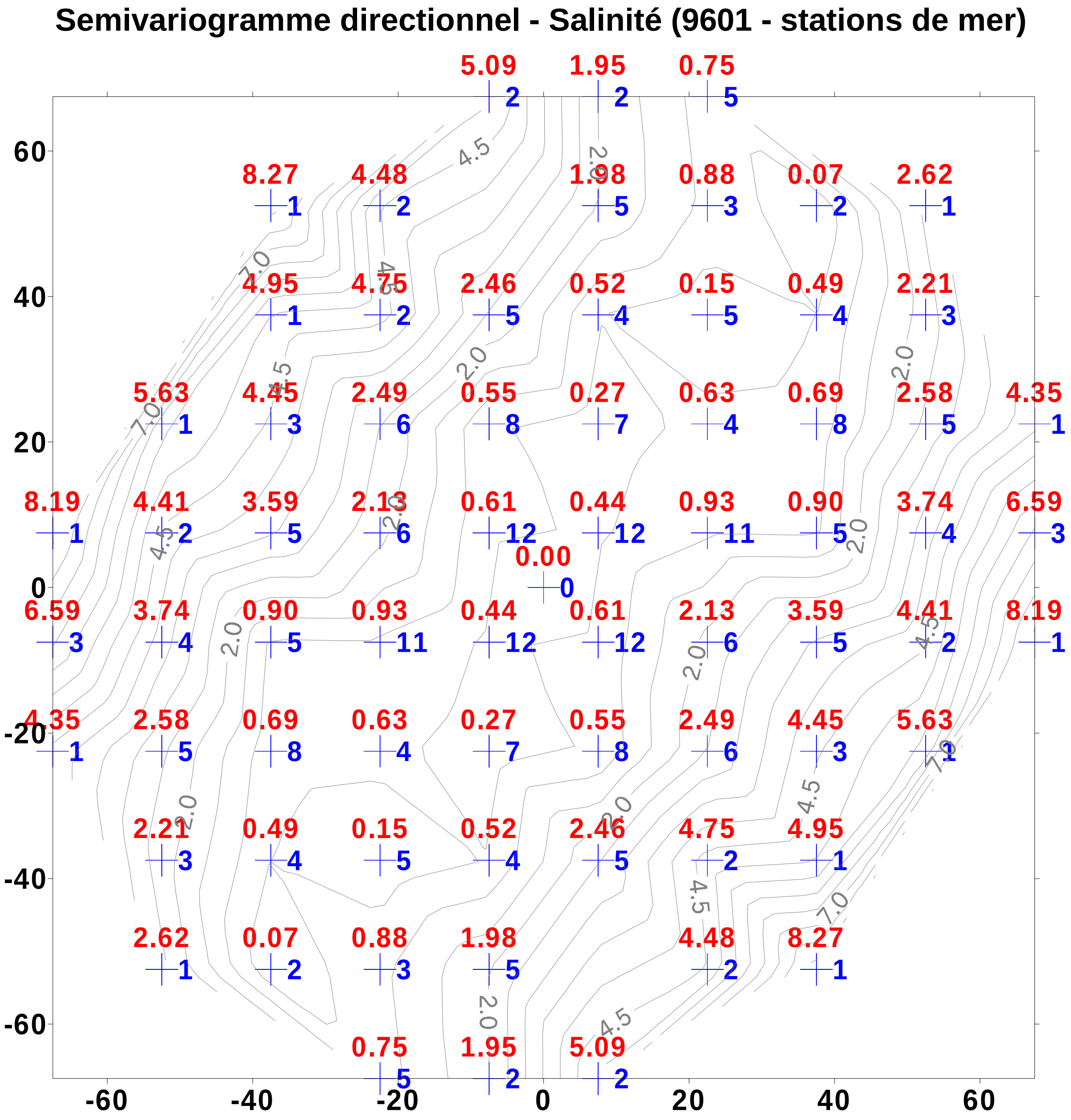
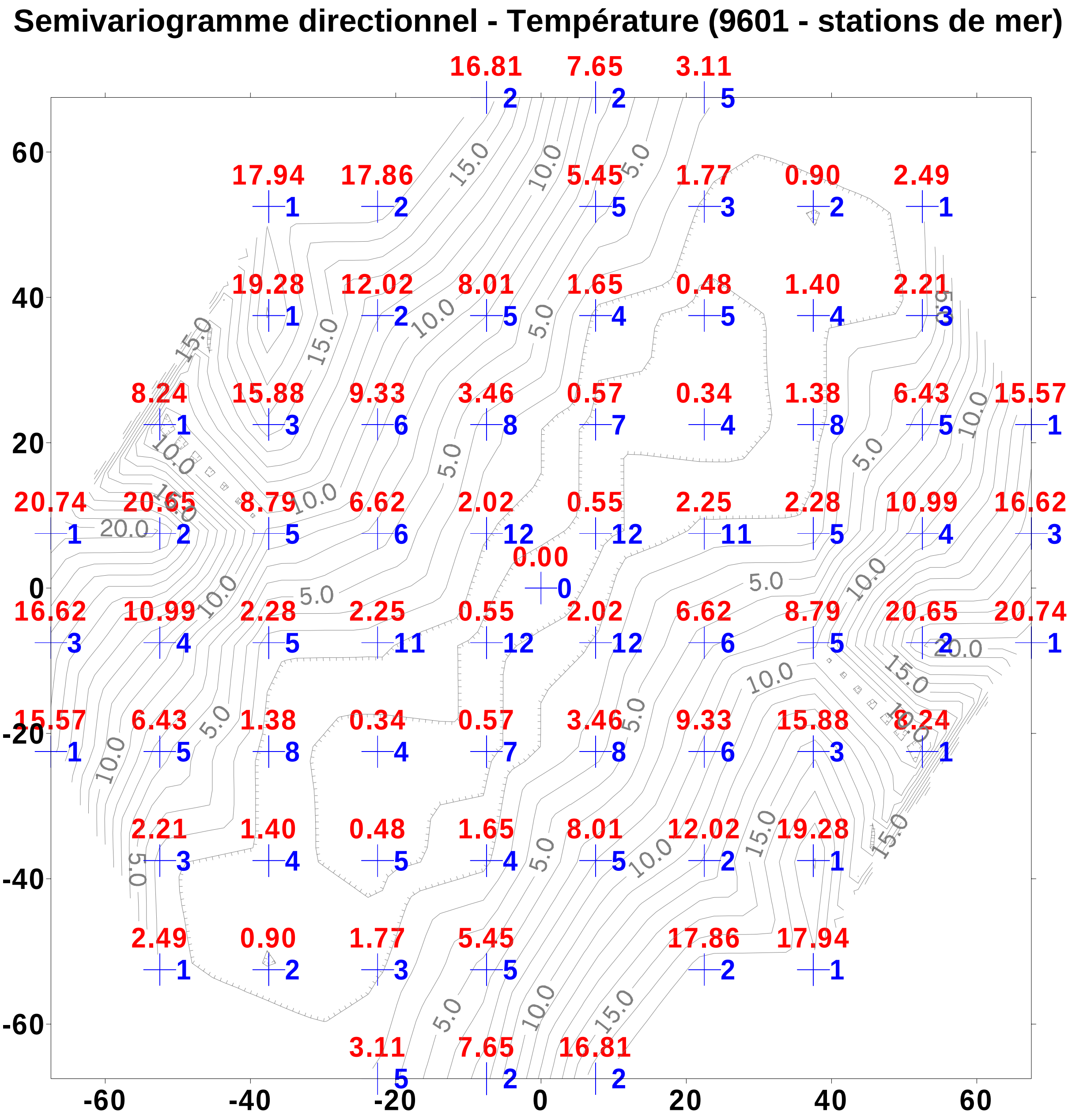


Figure 12: Directional salinity semivariogram (continental shelf area stations – campaign 9601). Legend: red numbers: semivariance computed in each cell (15 km x 15 km); blue numbers: number of points in each cell; isolines: iso-semivariance curves (equidistance 1 °C^2).



2. *Discussion*

The first statement concerns the very little number of point's couples in each cell. Even if Isaaks & Scrivastava (1989) tell us that the construction of direction semivariogram is not recommended when the omnidirectional semivariogram is not well structured and when the number of data is very little, we can observe on both figures that the semivariance increase is very low in the SW – NE direction. In this direction, the autocorrelation is well conserved for a relatively great distance for the two considered physical parameters. This is probably due to the possible maximum gradients of salinity and temperature in the perpendicular direction from the Scheldt estuary (and the coast) toward NW.

Moreover, it is possible to see on both figures that the semivariance increases nearly linearly with the distance in all the direction. The rate of this linear increase for the temperature is very high in the NW – SE direction and small in the perpendicular direction. A little decrease in the slope is observed when the distance becomes smaller but it is not well defined enough to be modelled.

The orientation and dimensions of the anisotropy ellipse given by Surfer 7 is reported here after and can be compared with the figures 11 and 12. The directional semivariograms computed for the salinity measured at the 18 sea stations using the Surfer 7 software give a value of 2 for the ratio between the major and minor axes of the ellipse. The orientation of the major axe of this ellipse is given by an anti-clockwise angle of 61° measured from the W-E direction. For the temperature, the ratio and angle are respectively 2 and 54°.

B. The precision of the interpolation process

The kriging method has been applied to interpolate water salinity and temperature in the continental shelf area for the campaign 9601 using the measurements coming from the stations acquisitions (18 stations). The Arcinfo 8.0.2 and Surfer 7 softwares have been used to perform this interpolation. The method has been described in former report but the aspect of the precision has not been discussed previously. We will focus the following discussion on this aspect.

Ordinary kriging method gives the possibility to compute the variance for each node of the computed grid. This variance is a good estimator of the precision. Both softwares allow the computation of this grid. Nevertheless, every software has its own characteristics that are summarized in the table III.

TABLE III

ARCINFO 8.0.1	**SURFER 7**
POSSIBILITY TO USE A MASK GRID	IMPOSSIBLE TO USE A MASK GRID
VERY POOR TOOLS TO ANALYZE OMNIDIRECTIONAL SEMIVARIOGRAMS	VERY HIGH PERFORMANCE TOOLS FOR THE ANALYSIS OF OMNIDIRECTIONAL AND DIRECTIONAL SEMIVARIOGRAMS
IMPOSSIBILITY TO TAKE THE ANISOTROPY IN CONSIDERATION	POSSIBILITY TO FIX THE ANISOTROPY PARAMETERS

The variance computing method is described by Issaks & Srivastava (1989). It will be summarized and illustrated in the next report. This is the first method exploited to determine the precision of the interpolation process in function of the geographic position. Nevertheless, this kind of validation is not independent of the interpolation process.

Another validation method of the interpolation process can also be used, if we can consider that all the measurements coming from the stations and from the 10 minutes time interval samplings along the ship track are comparable. In fact, this hypothesis is not verified in the Scheldt estuary is no temporal correction is applied. Considering the results of one time/advection correction computed simulating the net Lagrange drift during the whole period of the campaign 17-18 April 2000 for the stations measurements (Scory, results presented on the 3/7/2000), it is probably acceptable in the continental shelf area. This method is independent of interpolation process. Indeed, the first data set used to weight the station measurements during the kriging. It is then possible to use the second set of data to validate the interpolated values. We have then compared the results of this interpolation with the 10 minutes along track acquisitions to validate the kriging process.

The variances extracted from the variance grid for all these points using the Arcinfo LATTICESPOT function already described have been chosen to assess the difference between the interpolated values extracted from the interpolation resulting grid using the same function and the measured values.

The results are shown and discussed here after.

The ordinary kriging interpolation process of the Arcinfo 8.0.2 software has been applied to the salinity data set of the 18 stations localized in the continental shelf area. The procedure is given here under.

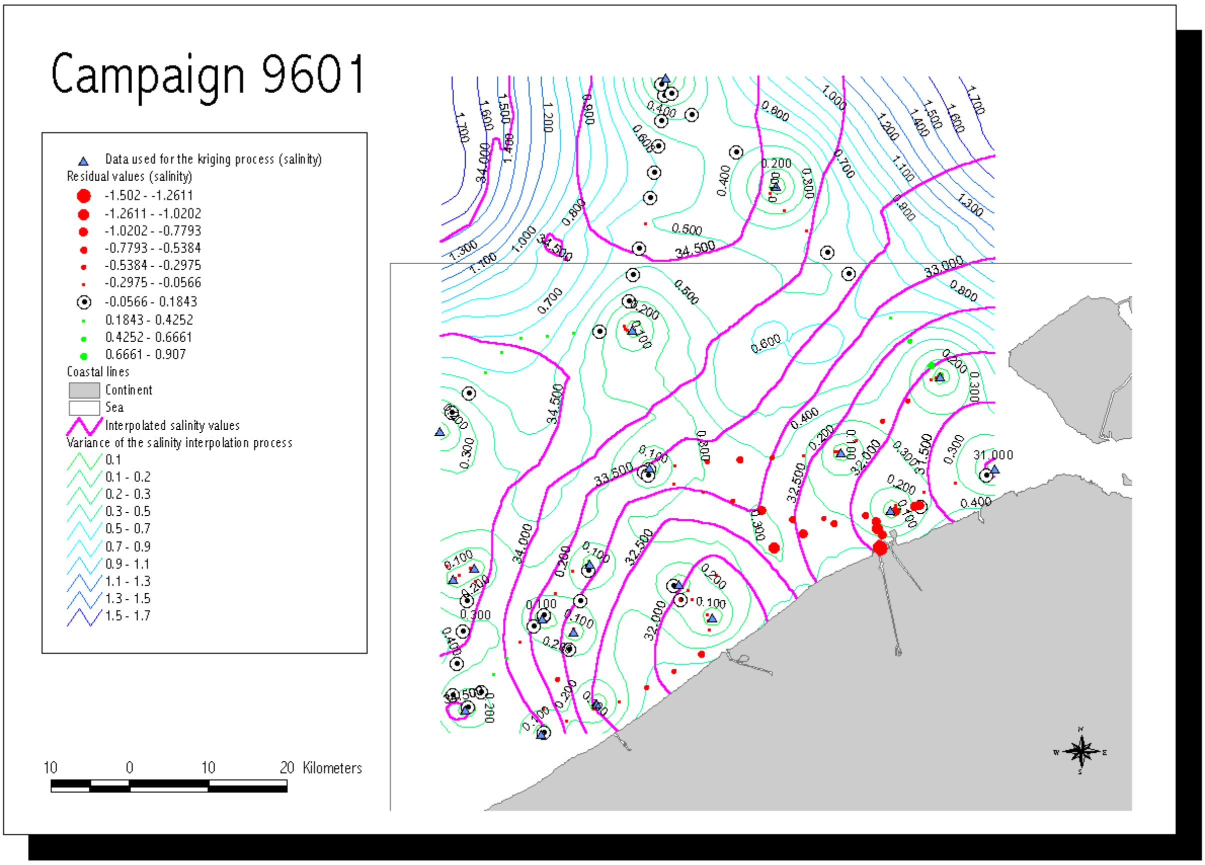
Arcinfo kriging interpolation and contouring processing

```
G-KR9601-SAL = kriging ( c-krmer9601 , psalat31_v , # , GRID , g-9601var-sal , LINEAR , sample
      , 10 , 200 , 0.125 ) (the first parameter of the function, c-krmer9601, is the name of
      the data set point cover, the second, psalat31_v, is the name of the item containing the
      data for interpolation, # means the possibility to use a barrier cover or file, the GRID
      options means that no semivariogram file is stored but only a grid file containing the
      results of the interpolation is produced, g-9601var-sal is the name of the predicted
      semivariance grid file, LINEAR means that a linear model is fitted on the computed
      semivariances, sample means that the number data used for the interpolation is fixed by
      a sample - the alternative is a search radius from the node, 10 is the minimum number of
      points used in the interpolation process, 200 is the maximum radius in map units (km)
      that determine the area around the node in which the 10 points must be found, 0.125 is
      the resolution of the output grids)
G-KR9601-SAL = kriging ( c-krmer9601 , psalat31_v , # , GRID , g-9601var-sal , LINEAR , sample
      , 10 , 200 , 0.125 )
latticecontour g-kr9601-sal c-kr9601-sal 0.5 32 (contouring)
latticecontour g-9601var-sal c-9601var-sal 0.002 0.568 (contouring)

Linear semivariance model (this model is a function of the computed grid resolution, the
      distance, h, is computed in pixels units):
      C0 = 0.514
      C = 0.010
      a = 53.250
      Sill = 0.524

G-KR9601-S2 = kriging ( c-krmer9601 , psalat31_v , # , GRID , g-9601var-s2 , LINEAR , sample ,
      18 , 300 , 1 )
G-KR9601-S2 = kriging ( c-krmer9601 , psalat31_v , # , GRID , g-9601var-s2 , LINEAR , sample ,
      18 , 300 , 1 )

Linear semivariance model (this model is a function of the computed grid resolution, the
      distance, h, is computed in pixels units):
      C0 = 0.000
      C = 1.537
      a = 31.000
      Sill = 1.537

latticecontour g-9601var-s2 c-9601var-s2 0.1 0.1
latticecontour g-kr9601-s2 c-kr9601-s2 0.5 32
```

Arcinfo processing applied to extract the interpolated salinity and variance for the 10 minutes along track data set

```
latticespot g-kr9601-s2 c-9601tab kr_s2
latticespot g-9601var-s2 c-9601tab var_s2
```

Map 6: Results of the interpolation process using the ordinary kriging method of the Arcinfo 8.0.2 software describe here above and classed posting of the signed difference between the observed salinity (10 minutes interval data set) and the interpolated salinity using the method described above



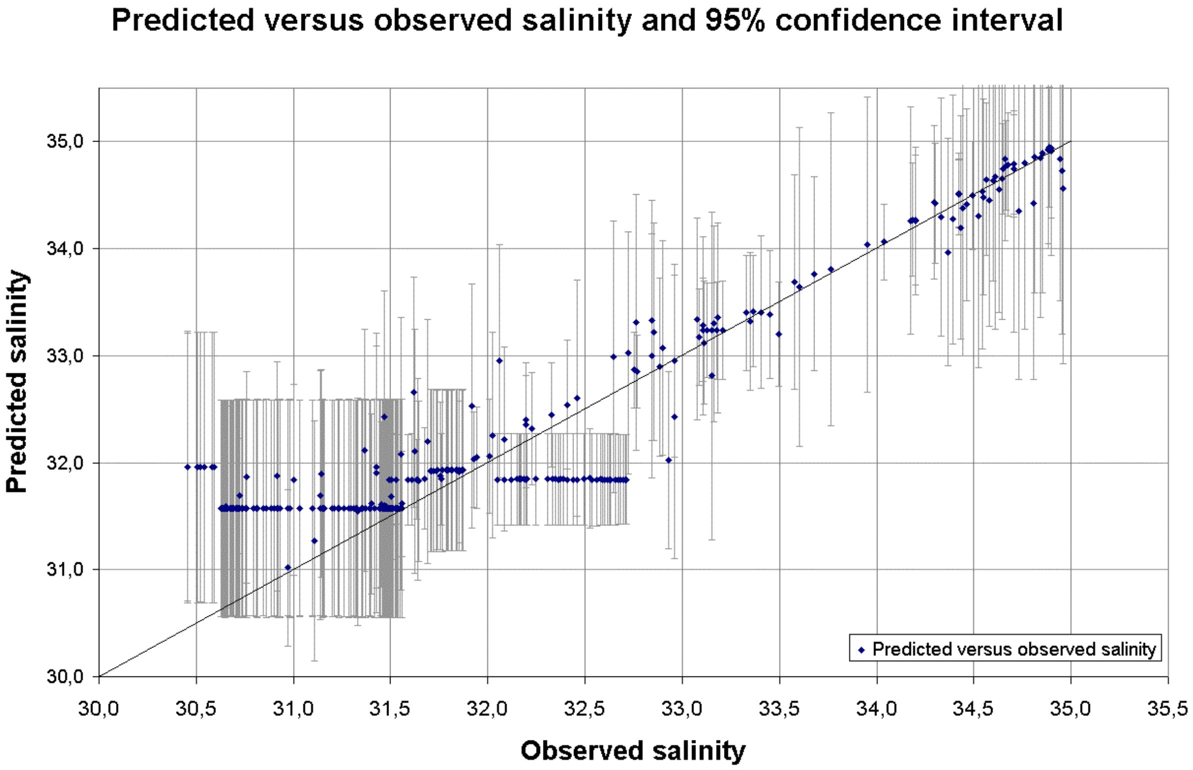
The map 6 shows the spatial distribution of the interpolation error. For the studied case, the gradients in the near coast area of Zeebrugge are too variable with respect to the spatial sampling density to allow an accurate interpolation process. In the other zones, the computed residual values are very low (from –0,5 – overestimation - to 0,4 ‰ - underestimation) and the magnitude order of the square root of the kriging process variance is the same than the absolute value of the residual values.

The scatter plot of the figure 10 illustrates the relation between the predicted and the observed values. It gives quantitative information about the global imprecision of the interpolation process independently of the localization of the validation points.

Nevertheless, we must remind that all the former interpolation and validation analysis are based on the hypothesis that the temporal variability of the salinity can be neglected regarding the spatial variability. That means that all the measurements are comparable.

Figure 13: Scatter plot of predicted values versus observed values and confidence interval at a 5% level around the interpolated value computed using the predicted semivariance for all the validation points. The confidence interval is given by the next formula:

$\hat{s} \pm 2 * \sqrt{\hat{\gamma}}$



UCS

Universitair Centrum voor Statistiek

Integrated and Dynamical Oceanographic Data Management - IDOD

Scientific Report
January 2001

Contribution of KUL-UCS

University Center of Statistics
Katholieke Universiteit Leuven
de Croylaan 52b
B-3000 Leuven
Belgium

KATHOLIEKE UNIVERSITEIT LEUVEN

1. Introduction

During the fourth year of the IDOD project the work of UCS has concentrated solely on the further development, integration, testing and documentation of the two following software components:

1. SAT: the statistical analysis tool that allows to perform statistical analyses on data retrieved from the database;

2. SQC: a stand-alone program that allows to screen data in the database and perform a statistical quality control on those data.

The concepts of both program have been described in earlier scientific reports: for SAT, we refer the reader to the scientific report 2000, for SQC the reader is referred to the scientific report 1999.

In appendix to this scientific report a detailled description is given of the various software packets:

1. Appendix A contains an installation manual for the IDOD/UCS software;

2. Appendix B: describes the statistical analysis tool and its functionalities;

3. Appendix C: contains an updated user manual for the statistical quality control program SQC);

4. Appendix D: details the structure of an Access database that is used by SQC to store the definition of datasets, quality control schemes, etcetera;

5. Appendix E: details the structure of an Access database that is used to interchange statistical results obtained in SAT to the SQC program.

A case study application of both the SAT program and the SQC program to nutrient data stored in the IDOD database is under development. However this report is still under production and will be released during the first semester of the upcoming year.

2. Summary

During this year, all effort has been devoted to the testing, integration and documentation of the software for the statistical analysis tool (SAT) and the statistical quality control (SQC). This has resulted in an operational software package that is relatively stable and well documented. Documentation of the programs is appended to this report.

In the upcoming year, the various programs will be further applied to real data and the results of these applications will be documented to demonstrate the use and functionality of the software.

UCS
Universitair Centrum voor Statistiek

Integrated and Dynamical Oceanographic Data Management - IDOD

Scientific Report: Appendices

January 2000

Contribution of KUL-UCS

University Center of Statistics
Katholieke Universiteit Leuven
de Croylaan 52b
B-3000 Leuven
Belgium

KATHOLIEKE UNIVERSITEIT LEUVEN

- Appendix A: Installation Manual
- Appendix B: Statistical Analysis Tool: Manual
- Appendix C: Statistical Quality Control Program: Manual
- Appendix D: Structure of QCData.mdb
- Appendix E: Structure of Statresults.mdb

Appendix A:
IDOD Software

Installation Manual

UCS

Universitair Centrum voor Statistiek

IDOD Software

Installation Manual

Version 1: Draft

01/11/2000

Table of Contents

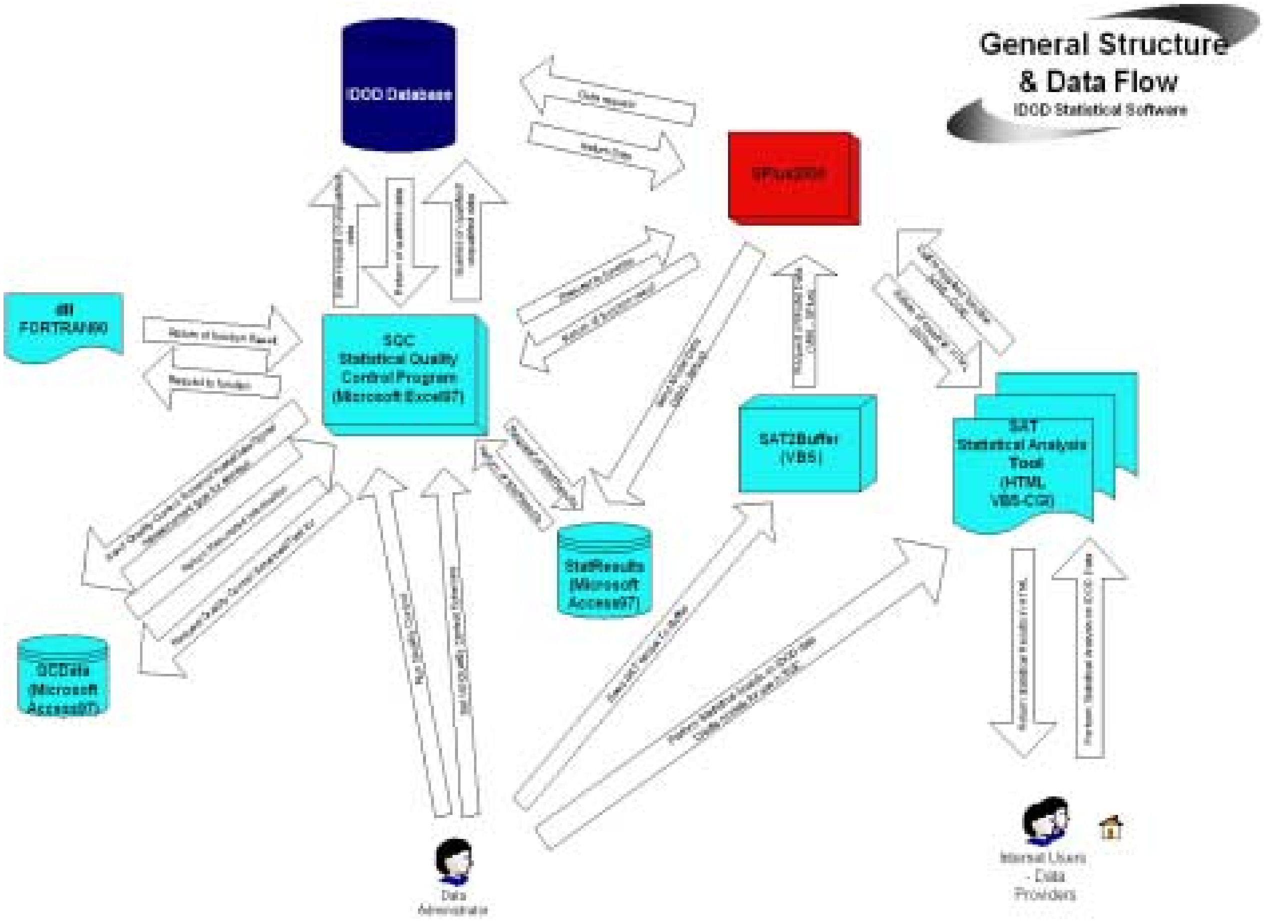
Introduction

The IDOD software is a group of software components developed to interact with the IDOD database to make statistical analyses and to perform statistical quality control procedures on the data.

The installation procedures of the different programs and the necessary updates of the different configuration files, depending on the computer running the software, are described in this manual.

IDOD Software structure

The IDOD Software structure is given in the picture below.



The following color codes are used:

UCS Developed software	light blue
BMM Devolped software	dark blue
Commercial Software	red

Necessary Software

The IDOD software is developed in different programming languages and uses different commercial software as backbone.

The following programs third-party software component need to be installed on the server machine to run the IDOD software.

- Office97 Professional Edition (Microsoft Company)
- SPlus2000 (MathSoft), a statistical software program. More information on this software can be found on the website of the MathSoft Company (http://www.mathsoft.com)
- Apache Web Server (version 1.3.12, Apache Organisation). More information on this software can be found on website of the Apache Organisation (http://www.apache.org)

To develop and/or maintain the IDOD software following development software platforms should be available

- Office97 Professional Edition (Microsoft Company)
- Visual Basic 5 (Microsoft Excel, Microsoft Access)
- Visual Fortran90 (Microsoft Company)
- SPlus2000 (MathSoft)
- HTML Editor (HotMetalPro6, SoftQuad company, http://www.HotMetalPro.com)

The UCS developed software consists of following groups of software

- SAT: software to perform statistical analysis on IDOD data and develop statistical models for use in SQC
- Sat2Buffer: software to store developed statistical software in supporting database
- SQC: software to perform statistical quality control
- Supporting databases, database to store the different models

In the remainder of this document the configuration and the setup of the different programs and files for the IDOD software is explained.

The setup is split up in 3 parts. The setup of the third-party software, the configuration of the setup files depending on the users computer and the setup of the UCS developed software components.

Setup of third-party software

Setup of Microsoft Office97 Professional Edition

The IDOD software is partly written in Microsoft Excel97 en uses Microsoft Access97 databases to store the different models in.

It is recommended that a full installation of the Microsoft Office97 Professional software is done to make sure all the components are available

Setup of SPlus2000 Professional

SPlus2000 is a statistical software package used as backbone in the SAT software.
For the installation of this software component, the custom setup procedure should be chosen.

Following options should be installed for the SAT software package

- Program Files
- User Interface and preference files
- Microsoft ODBC Files
- Windows Files for Program
- Development System Files
- Excel Add-in files

The other components can be installed but are not necessary. The option "Setup Disk Images" is not useful. This includes the bitmap files used in the setup procedure.

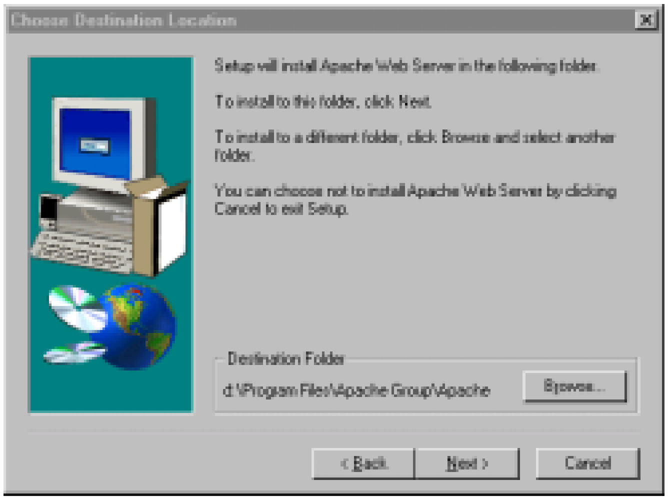
Setup of Apache Web Server (version 1.3.12)

The Apache Web server is a common used http server. This, or another webserver, is necessary to make the CGI functions of the SAT software possible.

The setup program can be followed. Afterwards, the configuration file should be manually adapted.

Setup Apache Webserver

Simply running this setup program and accepting the licence information gives the choice of the installation path of the software (here d:\Program Files\Apache Group\Apache)



Choose Destination Location

Setup will install Apache Web Server in the following folder.

To install to this folder, click Next.

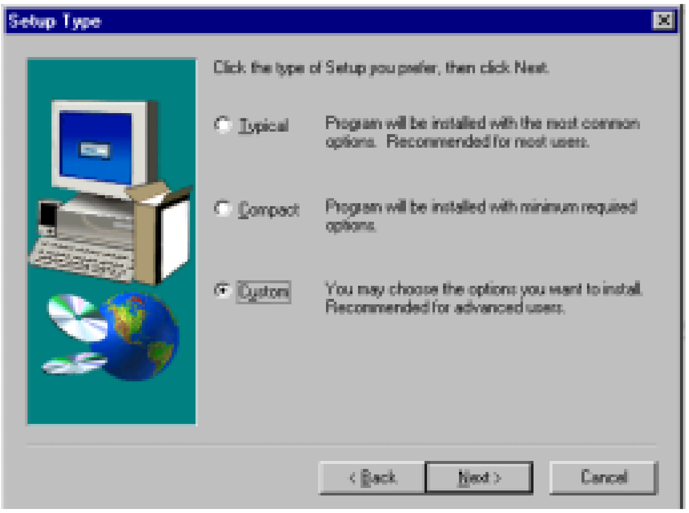
To install to a different folder, click Browse and select another folder.

You can choose not to install Apache Web Server by clicking Cancel to exit Setup.

Destination Folder

d:\Program Files\Apache Group\Apache

Next the setup type is chosen. Here it is recommended to choose the Custom setup Type.



Setup Type

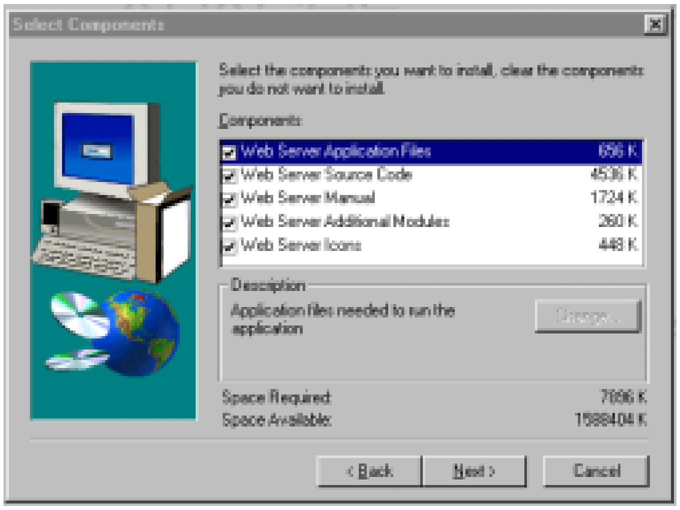
Click the type of Setup you prefer, then click Next.

○ Typical — Program will be installed with the most common options. Recommended for most users.

○ Compact — Program will be installed with minimum required options.

● Custom — You may choose the options you want to install. Recommended for advanced users.

It is recommended to install all components of the Apache Software.



Select Components

Select the components you want to install, clear the components you do not want to install.

Components

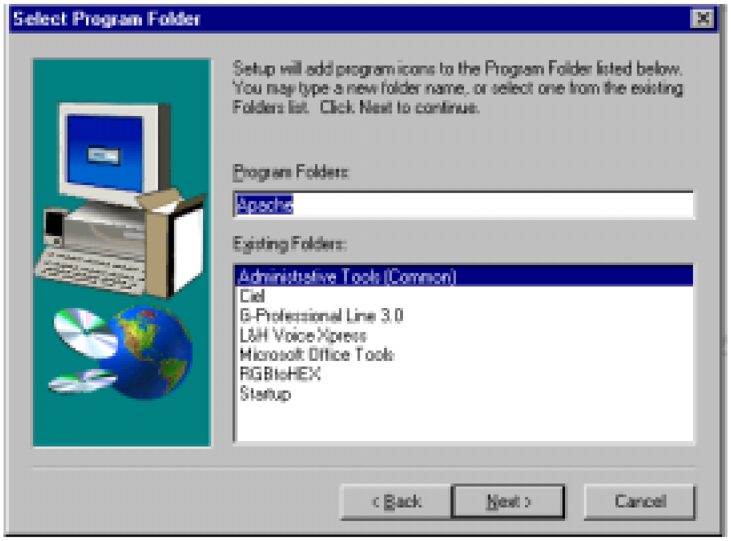
☑ Web Server Application Files	696 K
☑ Web Server Source Code	4536 K
☑ Web Server Manual	1724 K
☑ Web Server Additional Modules	260 K
☑ Web Server Icons	448 K

Description

Application files needed to run the application

Space Required: 7696 K
Space Available: 1588404 K

Shortcuts will be made in the program menu.



Select Program Folder

Setup will add program icons to the Program Folder listed below. You may type a new folder name, or select one from the existing Folders list. Click Next to continue.

Program Folders:

Apache

Existing Folders:

Administrative Tools (Common)
Ciel
G-Professional Line 3.0
L&H Voice Xpress
Microsoft Office Tools
RGBioHEX
Startup

Setup of configuration files

In the configuration file of the Apache Webserver, some computer specific settings should be changed. This is the file httpd.config, located in the ..\httpd\conf subdirectory of the installation path.

Note:
The settings for the configuration file depend on the computer configuration. This section only indicates the changes necessary. One should ask the system administrator, what the settings should be

- First of all the server name (or IP address) should be indicated. These are the following lines (the bold sections should be adapted)

```
#
# ServerName allows you to set a host name which is sent back to clients for
# your server if it's different than the one the program would get (i.e., use
# "www" instead of the host's real name).
#
# Note: You cannot just invent host names and hope they work. The name you
# define here must be a valid DNS name for your host. If you don't understand
# this, ask your network administrator.
# If your host doesn't have a registered DNS name, enter its IP address here.
# You will have to access it by its address (e.g., http://123.45.67.89/)
# anyway, and this will make redirections work in a sensible way.
#
#ServerName new.host.name
#Aanpassing
 ServerName 212.123.11.3
```

- Next the root directory of the webpages should be indicated. This is the directory where the SAT html files reside. This is changed in the following section of the http.config file:

```
#
# DocumentRoot: The directory out of which you will serve your
# documents. By default, all requests are taken from this directory, but
# symbolic links and aliases may be used to point to other locations.
#
#Aanpassing
#DocumentRoot "C:/Program Files/Apache/httpd/htdocs"
DocumentRoot "E:/Noordzeeproject/IDOD Software/Software/SATWWW"

#
# Each directory to which Apache has access, can be configured with respect
# to which services and features are allowed and/or disabled in that
# directory (and its subdirectories).
#
# First, we configure the "default" to be a very restrictive set of
# permissions.
#
<Directory />
    Options FollowSymLinks
    AllowOverride None
</Directory>

#
# Note that from this point forward you must specifically allow
# particular features to be enabled - so if something's not working as
# you might expect, make sure that you have specifically enabled it
# below.
#

#
# This should be changed to whatever you set DocumentRoot to.
#
#Aanpassing
#<Directory "C:/Program Files/Apache/httpd/htdocs">
<Directory " E:/Noordzeeproject/IDOD Software/Software/SATWWW ">
```

- Finally the cgi redirecting directory should be indicated. This is the directory where the CGI executables are available.

```
#
# ScriptAlias: This controls which directories contain server scripts.
# ScriptAliases are essentially the same as Aliases, except that
# documents in the realname directory are treated as applications and
# run by the server when requested rather than as documents sent to the client.
# The same rules about trailing "/" apply to ScriptAlias directives as to
# Alias.
#
#Aanpassing
#ScriptAlias /cgi-bin/ "C:/Program Files/Apache/httpd/cgi-bin/"
ScriptAlias /cgi-bin/ "E:/Noordzeeproject/IDOD Software/Software/SATWWW/cgi-bin"
#
# "C:/Program Files/Apache/httpd/cgi-bin" should be changed to whatever your ScriptAliased
# CGI directory exists, if you have that configured.
```

```
#
#Aanpassing
#<Directory "C:/Program Files/Apache/httpd/cgi-bin">
#    AllowOverride None
#    Options None
#</Directory>

<Directory "E:/Noordzeeproject/IDOD Software/Software/SATWWW/cgi-bin">
    AllowOverride None
    Options None
</Directory>
```

Setup of IDOD software

Setup of the configuration files of the SAT software

The SAT software needs 2 configuration files which should be located in the root of the C-Drive.

These are the following files:

- IDODConfigSPLUS.txt
  This file consist of necessary file paths for the SPLUS Program
- IDODCONFIG.txt
  This file consist of necessary file paths for the CGI exe's

The content of these files is computer/system dependent. An example is given below.

- IDODConfigSPLUS.txt

```
WEBDIRSPLUS, E:\\Noordzeeproject\\IDOD Software\\Software\\SATWWW\\
SPLUSDIR,D:\\Program Files\\sp2000
IDODINIT, E:\\Noordzeeproject\\IDOD Software\\Software\\SAT\\General functions Splus\\Function SPLUSInit\\SplusInit.ssc
DATABASE, E:\\Noordzeeproject\\IDOD Software\\Software\\IDOD Database\\Seawater.mdb
DATABASEDIR, E:\\Noordzeeproject\\IDOD Software\\Software\\IDOD Database
```

- IDODCONFIG.txt

```
DATABASE,E:\Noordzeeproject\IDOD Software\Software\IDOD Database
WEBDIR,E:\Noordzeeproject\IDOD Software\Software\SATWWW\
APACHEDIR,D:\Apache\Apache.exe
SPLUSDIR,D:\Program Files\sp2000\cmd\SPLUS.exe
```

Setup of libraries for SPLUS

The SAT software needs some extra libraries, next to the standard ones.
This are the following directories:

- Mass
- S2HTML

To install these libraries, copy them to the following path:

Application path SPLUS\library\   e.g. d:\Program Files\SP2000\Library

Finally in the SPLUS home directory a Directory IDOD should be created.

Setup of SAT software

The SAT software consists of a folder with html files and a folder with CGI-exe files. The location of these files is the same as indicated in the Apache Webserver Configuration files (see chapter 4.3.2).

The SAT software has 1 configuration file. This file is located under the following directory:
..\IDOD Software\Software\SAT\General functions Splus\Function SPLUSInit

In this .ssc (splus script) file the location of the different scripts on the computer are indicated.

This file looks like this:

```
#########################################################
## function to initialize all the SPLUS sesssions     ###
##This function should only be run at an installation ###
#########################################################

SPlusInit<-function(IDODdb="d:\\Program Files\\SP2000\\IDOD\\_Data"){

.getPath<-function(name){
    temp<-
    scan("c:\\IDODConfigSPLUS.txt",what=list(character(),character()),sep=",",mu
    lti.line=T)
   rownr<-pmatch(name, as.matrix(temp[[1]],nrow=1))
   return(as.matrix(temp[[2]])[rownr])
  }

  IDODdb<-paste(.getPath("SPLUSDIR"),"\\Idod\\_data",sep="")
  print (IDODdb)
  attach(IDODdb, name = "IDOD")
  print("DB attached")


######
## Location of the different SPlus scripts used in the SAT program.
## These functions must be run once to make the objects available for the VB-
    Cgi scripting
######

source("E:\\Noordzeeproject\\IDOD Software\\Software\\SAT\\Function
    StepWise\\SPlus\\combinations.ssc")
source("E:\\Noordzeeproject\\IDOD Software\\Software\\SAT\\General functions
    Splus\\Function SPLUSInit\\DataRetrieveDataInit.ssc")
source("E:\\Noordzeeproject\\IDOD Software\\Software\\SAT\\General functions
    Splus\\Function SPLUSInit\\DataRetrieveDataVariableInit.ssc")
source("E:\\Noordzeeproject\\IDOD Software\\Software\\SAT\\General functions
    Splus\\Function Variable_check\\exists.ssc")
source("E:\\Noordzeeproject\\IDOD Software\\Software\\SAT\\General functions
    Splus\\Function Variable_check\\existsvariable.ssc")
source("E:\\Noordzeeproject\\IDOD Software\\Software\\SAT\\Function Normality
    Check\\SPlus\\Gaussfit.ssc")
source("E:\\Noordzeeproject\\IDOD Software\\Software\\SAT\\General functions
    Splus\\Function SPLUSInit\\GetPath.SSC")
source("E:\\Noordzeeproject\\IDOD Software\\Software\\SAT\\Function
    Correlation\\CorrelationPlot\\Splus\\IDODCorrPlotWWW.ssc")
source("E:\\Noordzeeproject\\IDOD Software\\Software\\SAT\\Function
    Correlation\\CorrelationMatrix\\SPlus\\IDODcorWWW.ssc")
source("E:\\Noordzeeproject\\IDOD Software\\Software\\SAT\\Function SQC
    Distribution Test\\Splus\\IDODRegression.ssc")
source("E:\\Noordzeeproject\\IDOD Software\\Software\\SAT\\Function
   Regression\\Splus\\IDODRegression.ssc")
source("E:\\Noordzeeproject\\IDOD Software\\Software\\SAT\\Function
   SendData\\Splus\\IDODSendData.ssc")
source("E:\\Noordzeeproject\\IDOD Software\\Software\\SAT\\Function Summary
   Statistics\\Summary Plots\\SPlus\\IDODSumPlotWWW.ssc")
source("E:\\Noordzeeproject\\IDOD Software\\Software\\SAT\\Function Summary
   Statistics\\Summary Statistics\\SPlus\\IDODSumStatWWW.ssc")
source("E:\\Noordzeeproject\\IDOD Software\\Software\\SAT\\Function
   Trend\\Splus\\IDODtrendWWW.ssc")
source("E:\\Noordzeeproject\\IDOD Software\\Software\\SAT\\Function
   Variogram\\SPlus\\IDODVariogram.ssc")
source("E:\\Noordzeeproject\\IDOD Software\\Software\\SAT\\Function SQC
   Spatial\\Splus\\IDODvarioWWW.ssc")
source("E:\\Noordzeeproject\\IDOD Software\\Software\\SAT\\Function dataSPLUS
   retrieval\\RetrieveDataSets\\SPlus\\RetrieveDataSets.ssc")
source("E:\\Noordzeeproject\\IDOD Software\\Software\\SAT\\Function dataSPLUS
   retrieval\\RetrieveVariablesInDataset\\Splus\\RetrieveVariables.ssc")
source("E:\\Noordzeeproject\\IDOD Software\\Software\\SAT\\General functions
   Splus\\Function SPLUSInit\\SPlusInit.ssc")
source("E:\\Noordzeeproject\\IDOD Software\\Software\\SAT\\Function SQC
   Distribution Test\\Splus\\SQCDistribution.ssc")
source("E:\\Noordzeeproject\\IDOD Software\\Software\\SAT\\Function SQC
   Distribution Test\\Splus\\StepWiseInteraction.ssc")
source("E:\\Noordzeeproject\\IDOD Software\\Software\\SAT\\Function
   StepWise\\SPlus\\StepWiseRegression.ssc")
source("E:\\Noordzeeproject\\IDOD Software\\Software\\SAT\\Function SQC
   Regression\\Splus\\SubSetRegression.ssc")
source("E:\\Noordzeeproject\\IDOD Software\\Software\\SAT\\Function
   Transformation\\SPlus\\transfo.ssc")
source("E:\\Noordzeeproject\\IDOD Software\\Software\\SAT\\Function View
   Dataset\\Splus\\view.ssc")

print("IDOD is ready")
}

SPlusInit()
```

The different source lines can be made by means of the Excel program makePathNameForSplus.xls in the systemtools directory. In this program the startdirectory to search for the .ssc files should be indicated. The list will be made in the excel sheet. This can be copied in the existing SPLUSInit.ssc file.

IMPORTANT:
The line with source (…/SPlusInit.ssc) should be deleted in the SPLUSInit.ssc file

Setup of SQC software

The SQC software consists of a folder different Excel files. No special setup is required

Setup of SPLUS2Buffer software

The SPlus2Buffer Software is a VB exe files and no special setup is required

Appendix B: Statistical Analysis Tool

Manual

UCS

Universitair Centrum voor Statistiek

Statistical Analysis Tool

Manual

Version 1: Draft

1/11/2000

Table of Contents

List of figures

Introduction

The statistical analysis tool program SAT is developed to perform statistical analyses on data in the IDOD database. The SAT program is also used to derive models for use in the statistical quality control program SQC.
In this manual a description is given of the functionalities of the program.

Note:
SAT is developed to run in a web browser such as Netscape or Explorer. Since features as javascript, cgi, CSS, etc. are used only versions 4 or higher of these browsers will be able to run the program.

MAIN MENU

The SAT program has several sections in the main menu (see figure 1). The user can choose the appropriate section by a single click.

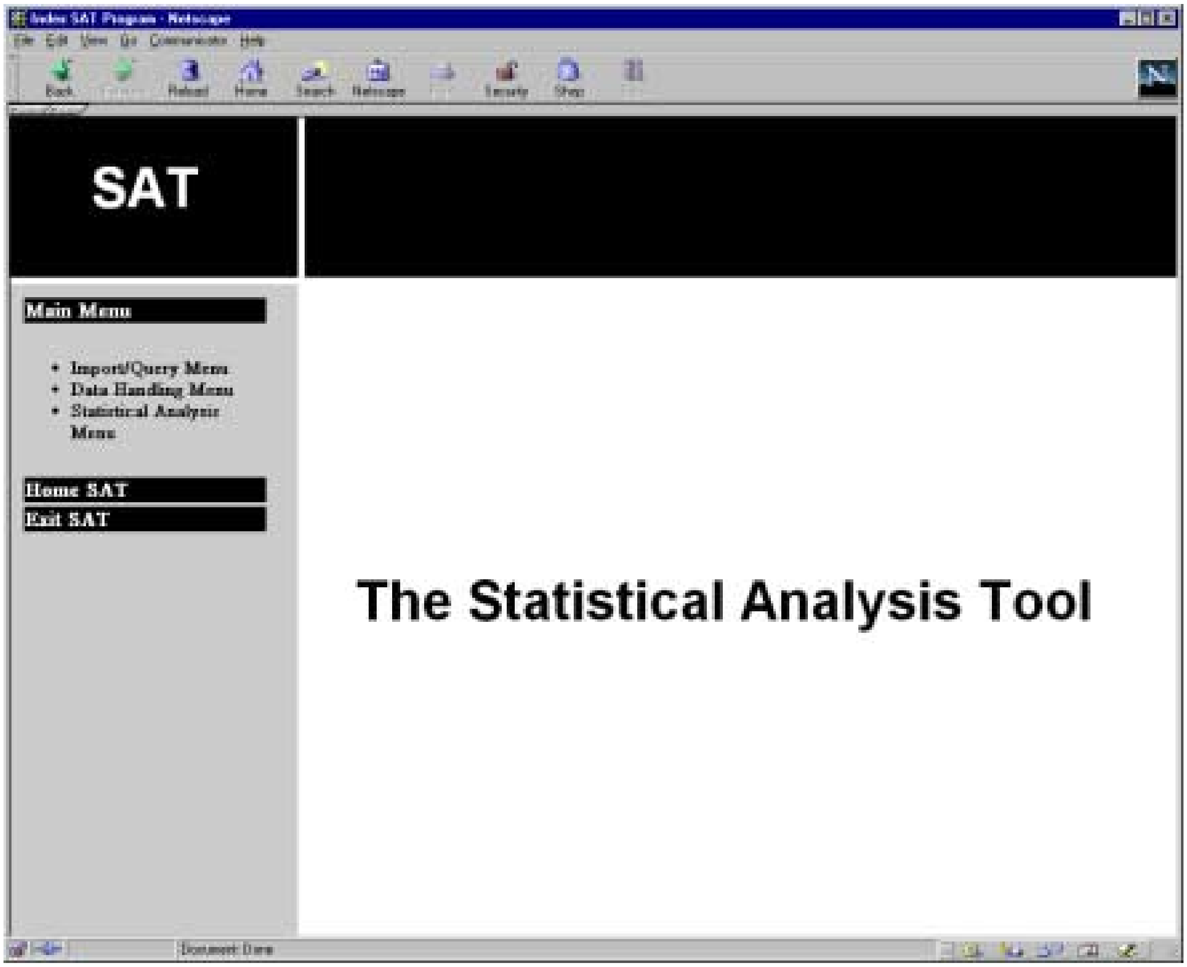


Figure 1: Main Menu

- **Section: Import/Query Menu**
  The “**Import/Query Menu**” section is used to retrieve data from the IDOD database and to send it to the SAT program.

- **Section: Data Handling Menu**
  The “**Data Handling Menu**” section is used to perform some basic data manipulation actions. The user is able to make transformations of the data, to change a numerical data type to a categorical variable and to view (or download) the selected data.

- **Section: Statistical Analysis Menu**
  In the “Statistical Analysis Menu” section, the user is able to perform statistical analyses on the selected data. Supported functionalities are Summary Statistics, Correlation Analyses, Trend Analysis, Regression Analysis, Spatial Analysis, PCA/Factor Analysis. This section is also used to derive models for the statistical quality control program.

Each of these sections is further detailed in this manual.

IMPORTANT NOTE:
In SAT all the commands and fill-in fields are CASE SENSITIVE. This is due to the underlying statistical program SPLUS2000. Please be careful when filling in data sets and variable names.

IMPORT/QUERY MENU

In the “**Import/Query Menu**” section (see figure 2), the user is able to import data from the IDOD database. After the query has been specified the data are retrieved from the IDOD database. The function “Send Data to SAT” makes those data available for further use in the SAT program.

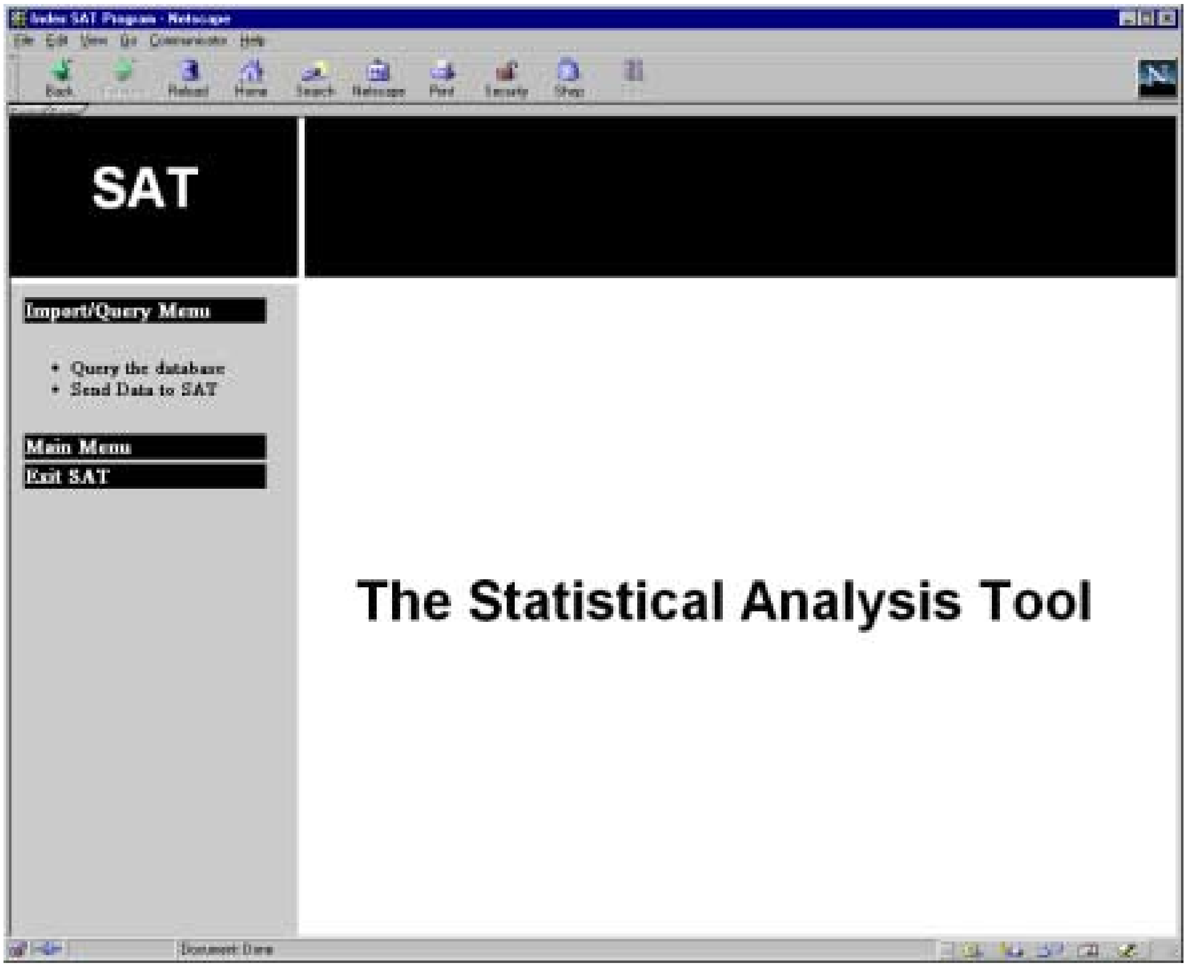


Figure 2: Import/Query Menu

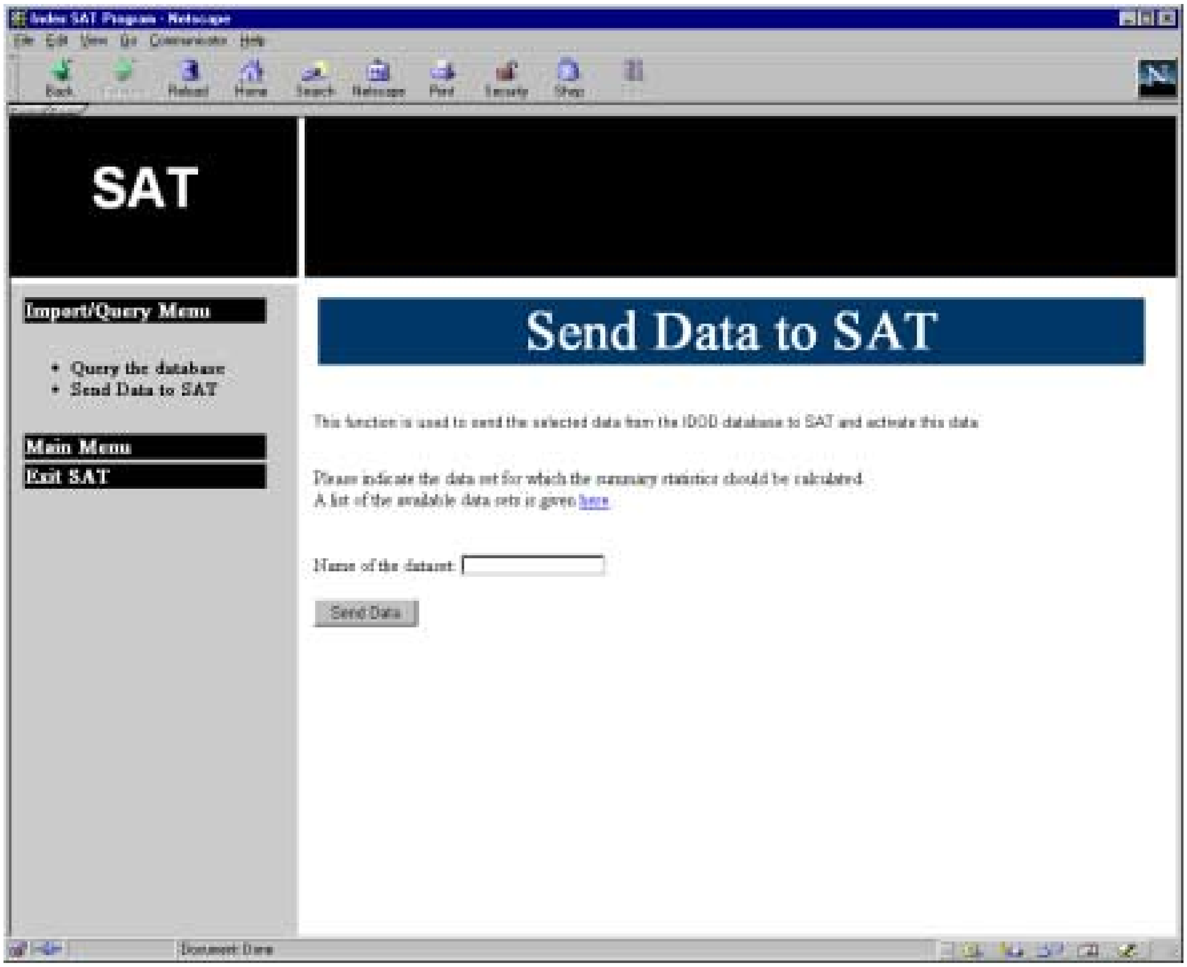
Step 1: Query the database

Note:
The choice to query the database is not operational at this time. The implementation will be completed by BMM/Ulg at a later stage.

At the present time, selecting data is possible off-line using the first prototype IDOD database, stored in Microsoft Access97.

Step 2: Send Data to SAT

The option "**Send Data to SAT**" in the menu activates the screen shown in figure 3. To activate the data for SAT, the user only has to provide a name for the data set and push the button "Send Data". A message is given to indicate that the data set is imported and activated in SAT.



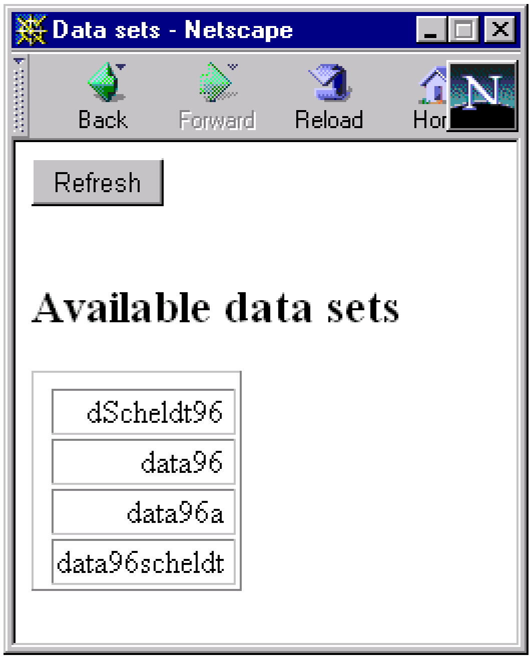
This function is used to send the selected data from the IDOD database to SAT and activate this data

Please indicate the data set for which the summary statistics should be calculated.
A list of the available data sets is given here

Name of the dataset:

Figure 3: Send Data to SAT function

The names of the data sets, already active in SAT, can be viewed by following the appropriate "here" link in this menu. This link opens a popup browser with the names of the active data sets. (see figure 4).



Available data sets

dScheldt96
data96
data96a
data96scheldt

Figure 4: Pop-up browser with available datasets

DATA HANDLING MENU

The "**Data Handling Menu**" section is used to perform basic operations on the variables in the active data sets.
After selection of the “**Data Handling Menu**” option in the “**Main Menu**” the user has the choice between 3 functions (see figure 5).

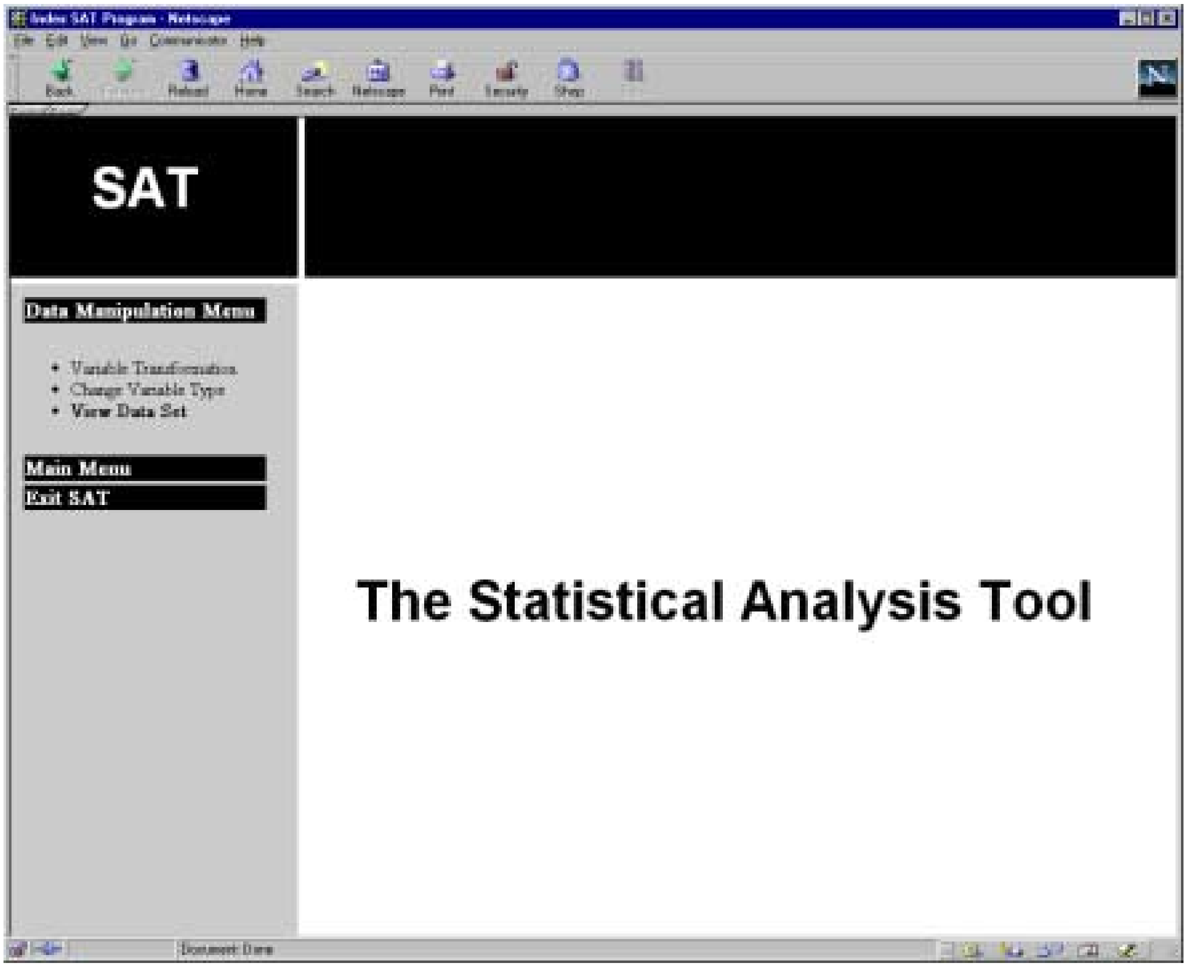
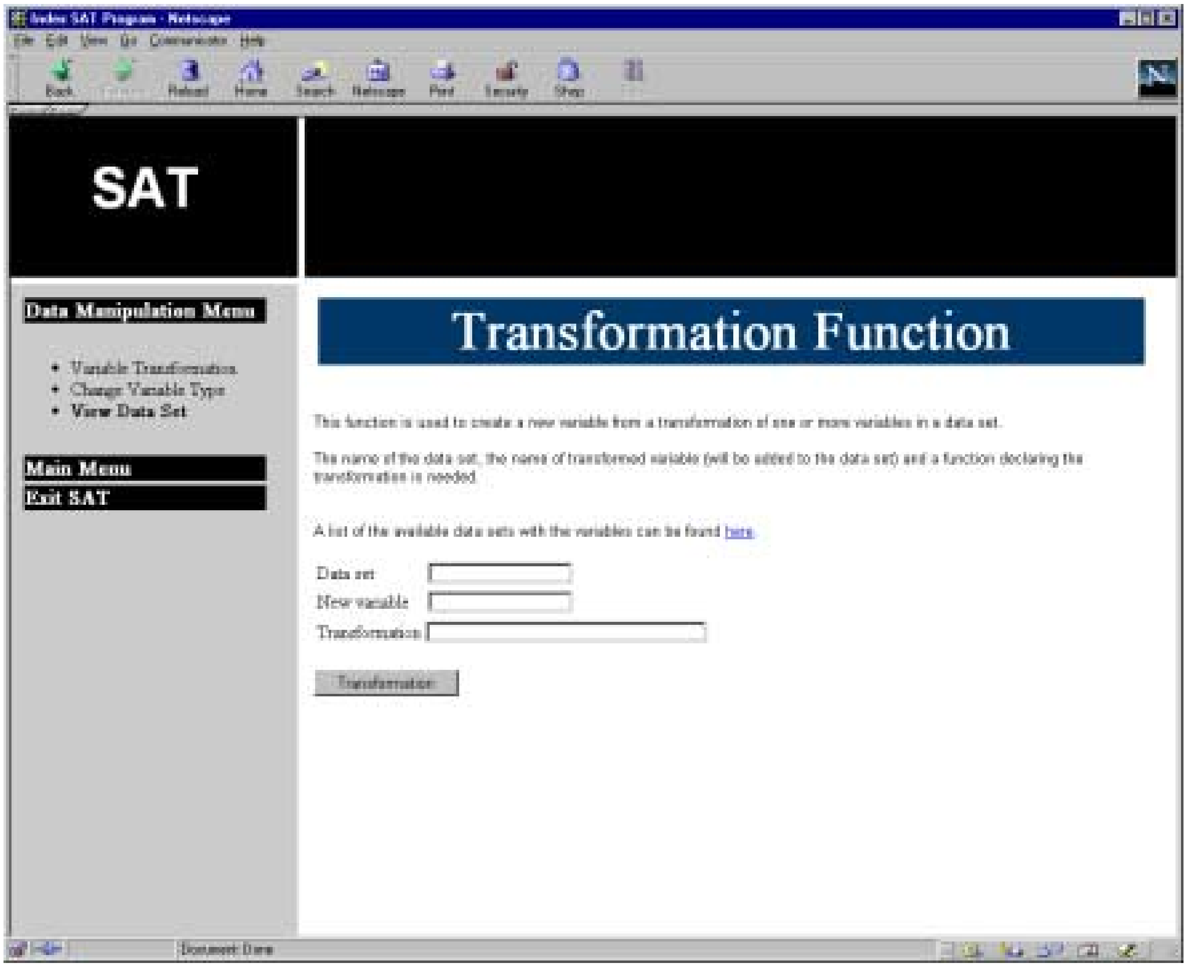

Index SAT Program - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop
SAT
Data Manipulation Menu
Variable Transformation
Change Variable Type
View Data Set
Main Menu
Exit SAT
The Statistical Analysis Tool
Document: Done

Figure 5: Data Handling Menu

Variable transformation

The function “**Variable Transformation**” is used to insert a transformed variable into the selected data set. (see figure 6)



Transformation Function

This function is used to create a new variable from a transformation of one or more variables in a data set.

The name of the data set, the name of transformed variable (will be added to the data set) and a function declaring the transformation is needed.

A list of the available data sets with the variables can be found here.

Data set
New variable
Transformation

Figure 6: Variable Transformation Function

The user has to fill in 3 fields to create a transformed variable:

- Data set: the name of the data set in which the original variable is located
- New variable: the name of the transformed variable which the users wants to use
- Transformation: the transformation function. In this function unary (on one variable), binary (on two variables) and combinations of these operators can be used. The following operators are allowed
  - Unary:

^ (power)	e.g. AMON^3
log (natural logarithm)	e.g. log(AMON)
log10 (logarithm base 10)	e.g. log10(AMON)
exp (exponential)	e.g. exp(AMON)
sqrt (square root)	e.g. sqrt(AMON)
/ (division)	e.g. AMON/2
* (multiplication)	e.g. AMON*2

  - Binary:

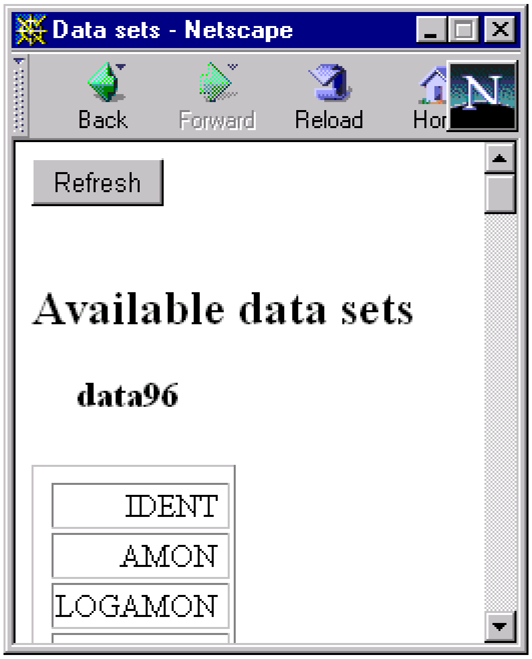
+ (sum)	e.g. AMON+NTRI
- (minus)	e.g. AMON-NTRI
/ (division)	e.g. AMON/NTRI
* (multiplication)	e.g. AMON*NTRA

- Combination of the unary and binary operators

After the button “**Transformation**” is pressed, the new variable is added as the last column in the data set.

Note:

> The names of the existing data sets and their variables can be viewed by following the “here” link, which shows the names of the data sets in a popup browser. An example of the form of this window is given in figure 7.



Available data sets

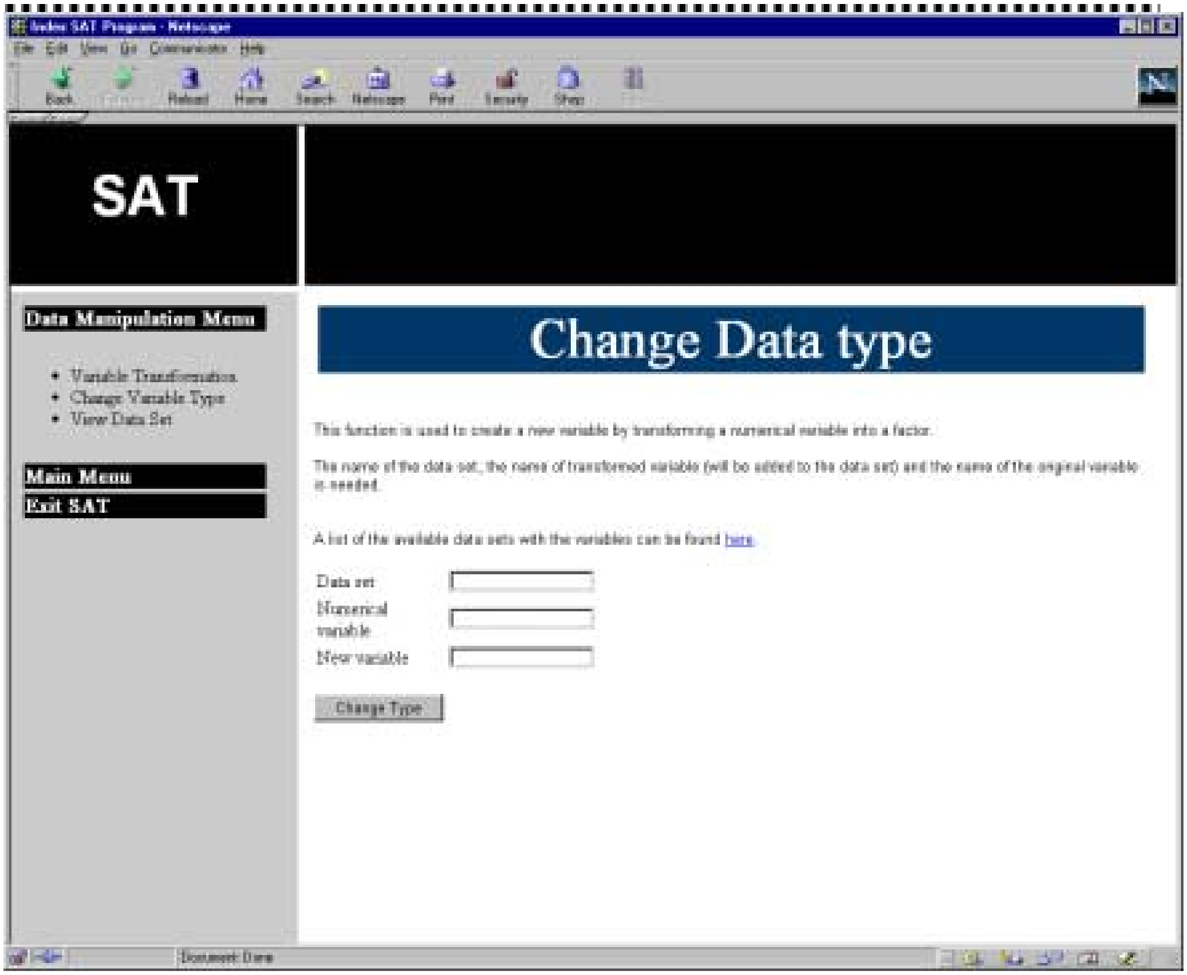
data96

IDENT
AMON
LOGAMON

Figure 7: Existing data sets and variables

Change Variable Type

The function “Change Variable Type” is used to transform a numerical variable in a factor
variable. (see figure 8)


Index SAT Program - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Stop
SAT
Data Manipulation Menu
Variable Transformation
Change Variable Type
View Data Set
Main Menu
Exit SAT
Change Type
Document: Done

Change Data type

This function is used to create a new variable by transforming a numerical variable into a factor.

The name of the data set, the name of transformed variable (will be added to the data set) and the name of the original variable is needed.

A list of the available data sets with the variables can be found here.

Data set

Numerical variable

New variable

Figure 8: Change Variable Type function

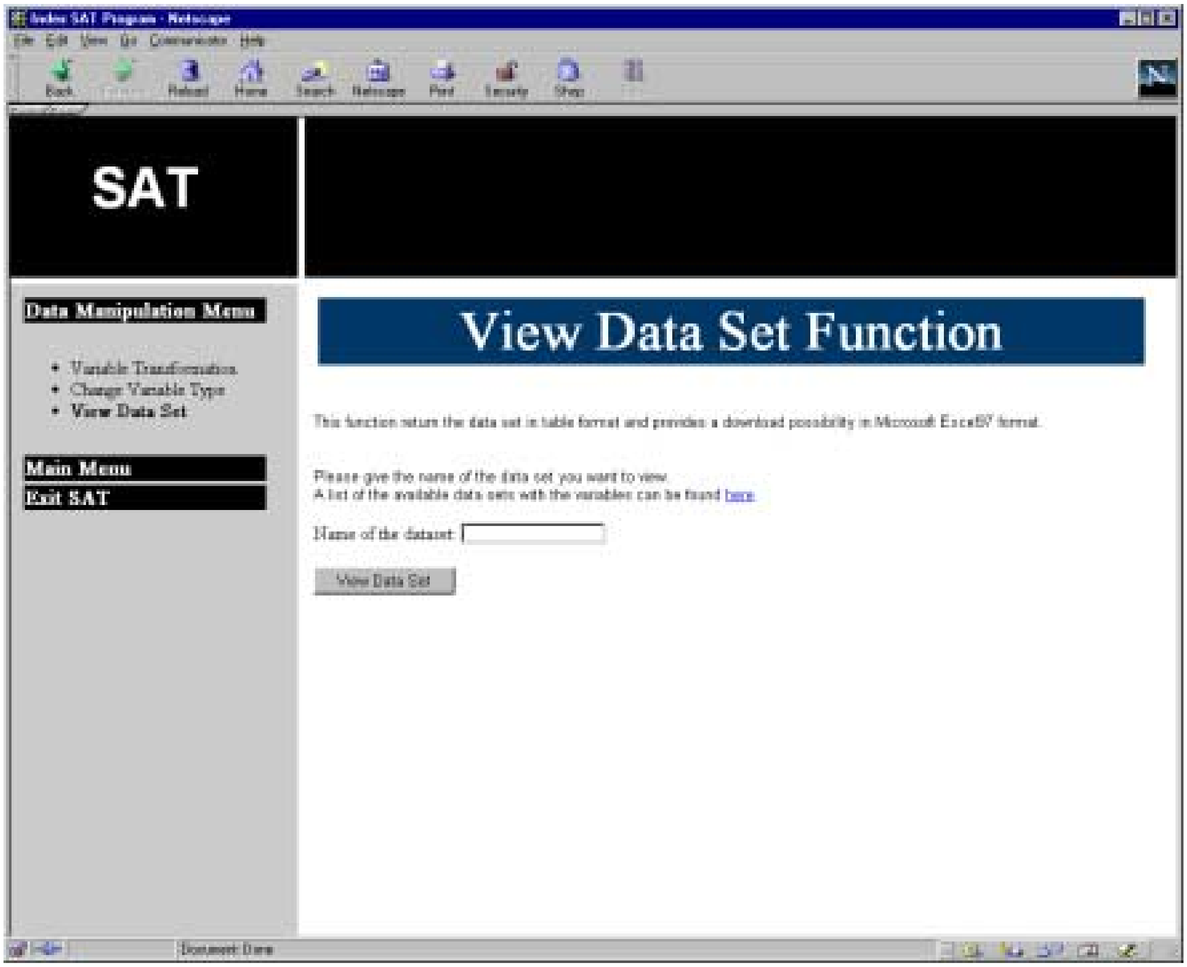
The user has to fill in following fields:

- Data set: the name of the data set in which the original variable is located
- Variable: the name of the numerical variable
- New variable: the name of the new factor variable, which is added to the data set.

The factor variables are possible necessary for regression models.

View Data Set

The “**View Data Set**” function is used to represent the data set in a tabular form. The user only has to indicate the name of the data set. A table with the data is given and a link to download the data set as a Microsoft Excel sheet is available.



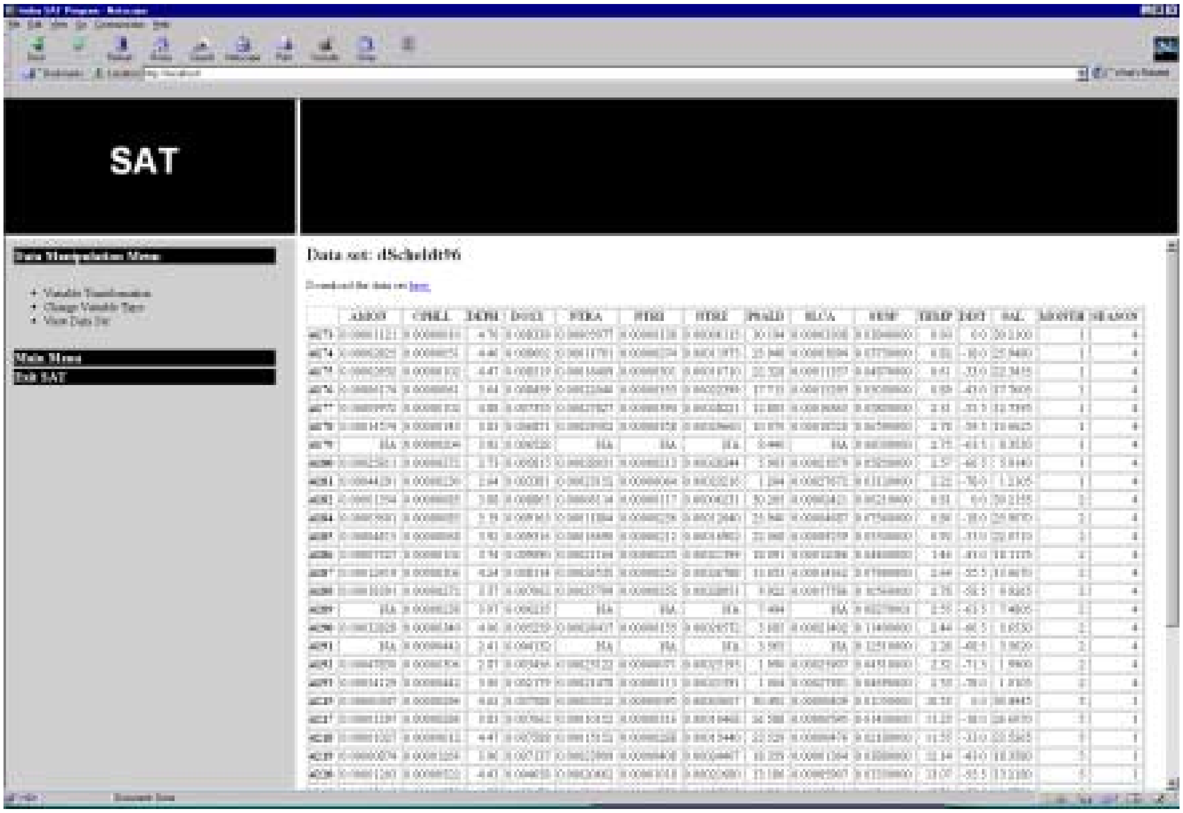
This function return the data set in table format and provides a download possibility in Microsoft Excel97 format.

Please give the name of the data set you want to view.
A list of the available data sets with the variables can be found here.

Name of the dataset:

Figure 9: View Data Set function

An example of the result of this function is shown in figure 10.



SAT

Data Manipulation Menu

- Variable Transformation
- Change Variable Type
- View Data Set

Main Menu

Exit SAT

Data set: dSchelde96

Download the data set here.

	AMON	CPHLL	DEPH	DOXY	NTRA	NTRI	NTRZ	PHALD	SLCA	TEMP	TRBP	DIST	SAL	MONTH	SEASON
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1	4
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1	4
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1	4
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1	4
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1	4
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1	4
[illegible]	NA	[illegible]	[illegible]	[illegible]	NA	NA	NA	[illegible]	NA	[illegible]	[illegible]	[illegible]	[illegible]	1	4
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1	4
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1	4
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2	4
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2	4
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2	4
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2	4
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2	4
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2	4
[illegible]	NA	[illegible]	[illegible]	[illegible]	NA	NA	NA	[illegible]	NA	[illegible]	[illegible]	[illegible]	[illegible]	2	4
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2	4
[illegible]	NA	[illegible]	[illegible]	[illegible]	NA	NA	NA	[illegible]	NA	[illegible]	[illegible]	[illegible]	[illegible]	2	4
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2	4
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2	4
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	3	1
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	3	1
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	3	1
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	3	1
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	3	1

Figure 10: Result of view data set function

STATISTICAL ANALYSIS MENU

The “**Statistical Analysis Menu**” option is used to perform basic statistical procedures and model building on the active data sets. In this section, it is also possible to derive models for use in the statistical quality control program (SQC).

Different types of statistical analysis are available (see figure 11). At this present moment not every function is implemented yet. As soon as data are available that allows the functions to be tested, implementation will be completed.

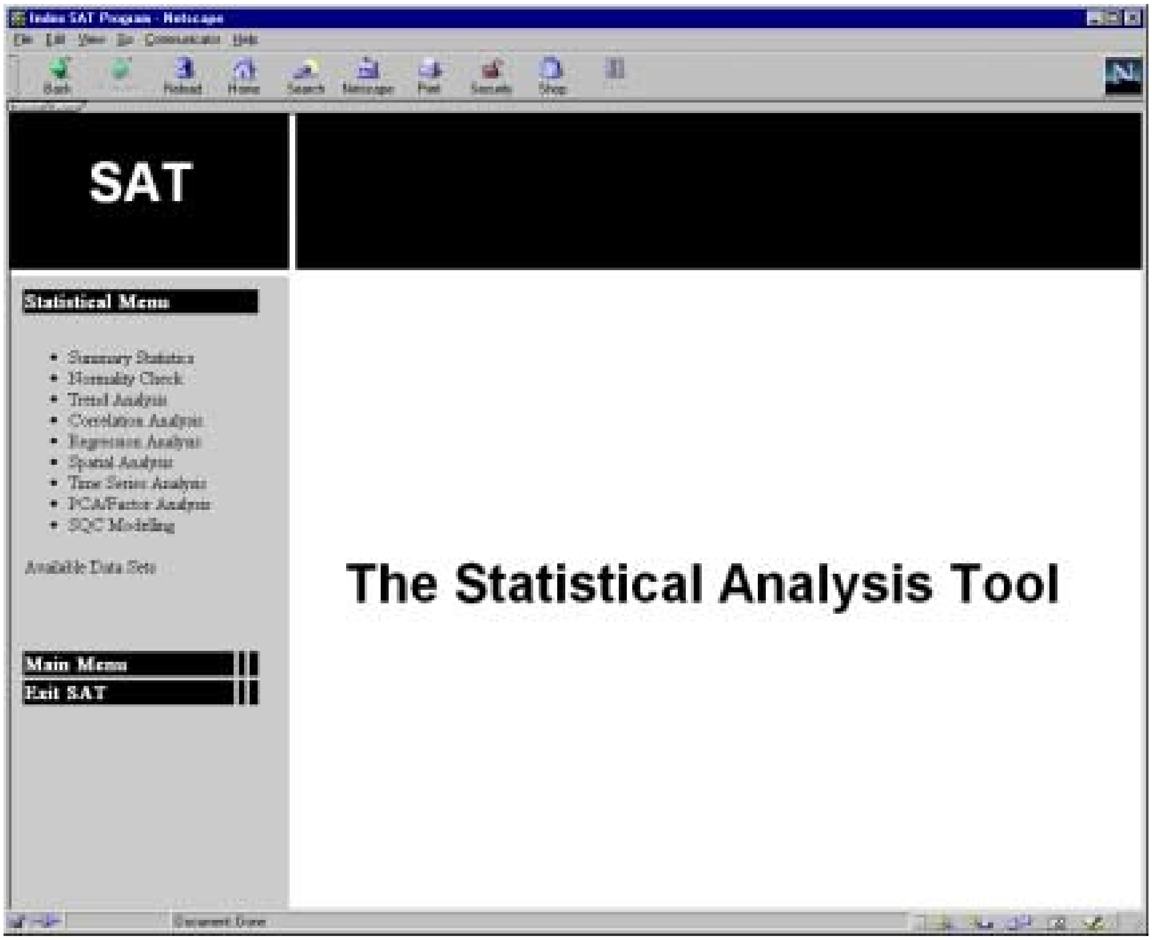


Figure 11: Statistical Analysis Menu

The following types of analysis are available.

- Summary Statistics
- Normality Check
- Trend Analysis:
- Correlation Analysis
- Regression Analysis
- Spatial Analysis
- Time Series Analysis
- PCA/Factor Analysis
- SQC Modeling

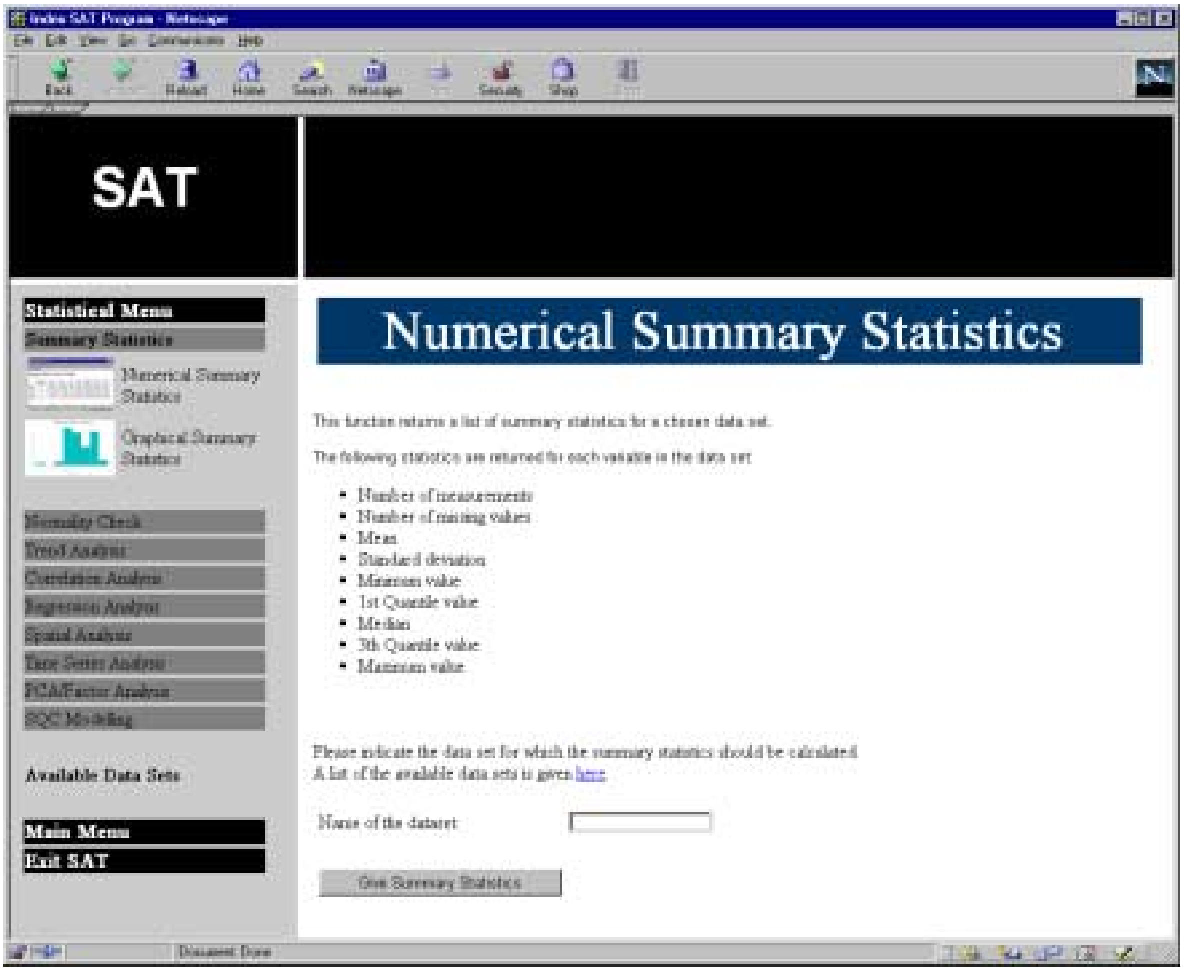
The functionality of these functions is explained in the following paragraphs.

Summary Statistics

The “Summary Statistics” section groups basic exploratory data analysis tools. These functions are used to provide the user some insight in the data he/she whishes to analyze. The section is split up in 2 parts that are given detailed next.

Numerical Summary Statistics

The “Numerical Summary Statistics” function is used to derive basic statistics for the variables in the data set.



Numerical Summary Statistics

This function returns a list of summary statistics for a chosen data set.

The following statistics are returned for each variable in the data set

- Number of measurements
- Number of missing values
- Mean
- Standard deviation
- Minimum value
- 1st Quantile value
- Median
- 3th Quantile value
- Maximum value

Please indicate the data set for which the summary statistics should be calculated.
A list of the available data sets is given here.

Name of the dataset

Give Summary Statistics

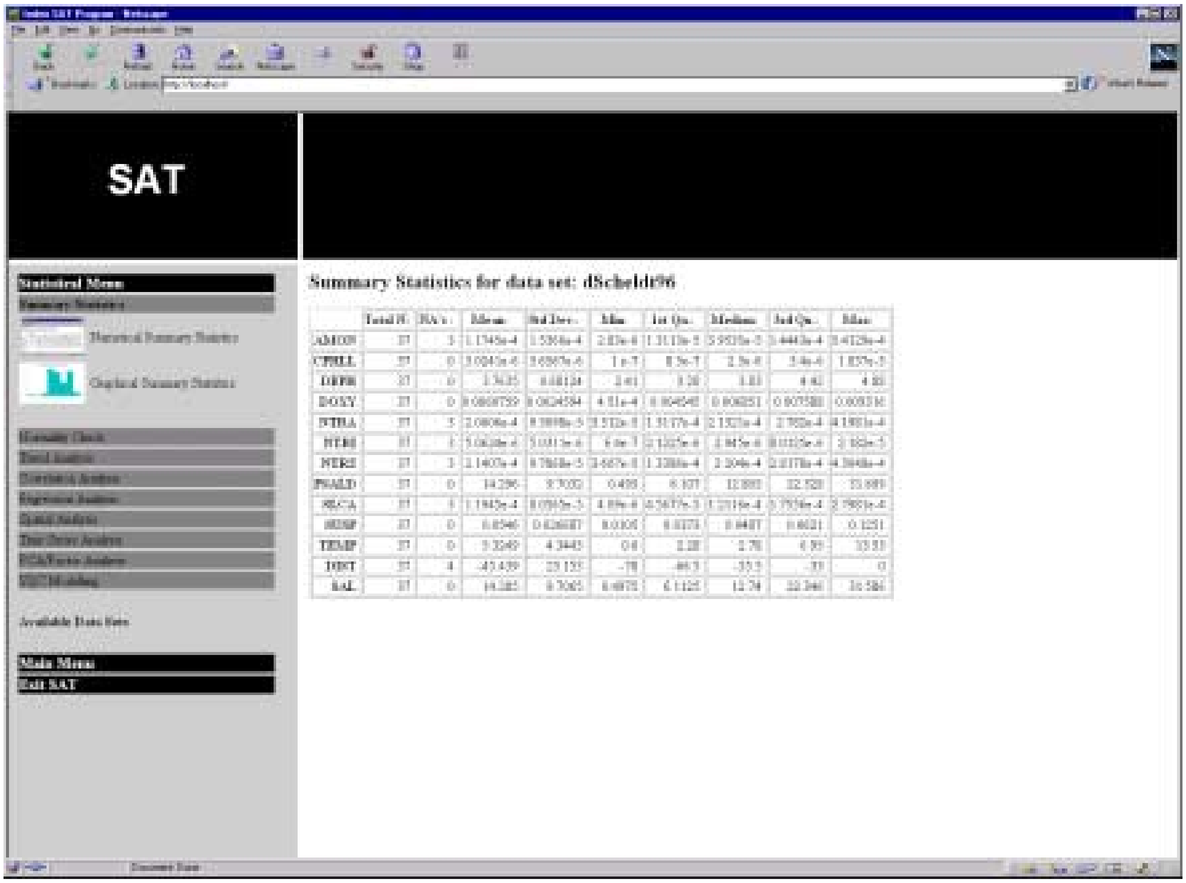
Figure 12: Numerical Summary Statistics function

The following statistics are given for a chosen data set:

- Number of measurements
- Number of missing values
- Mean
- Standard Deviation
- Minimum Value
- 1ste Quartile
- Median
- 3rd Quartile
- Maximum Value

The user should specify the name of the data set that should be analyzed. Pushing the button “**Give Summary Statistics**” results in a table with the statistics for each variable in the chosen data set.

An example of the output op this function is given in figure 13.



Summary Statistics for data set: dScheldt96

	Total N	NA's	Mean	Std Dev.	Min	1st Qu.	Median	3rd Qu.	Max
AMON	37	3	1.1745e-4	1.5364e-4	2.83e-6	1.3113e-5	5.9323e-5	1.4443e-4	5.4129e-4
CPHLL	37	0	3.0243e-6	3.6987e-6	1e-7	8.5e-7	2.5e-6	3.4e-6	1.837e-5
DEPTH	37	0	3.7635	0.68124	2.41	3.38	3.83	4.42	4.88
DOXY	37	0	0.0060759	0.0024594	4.51e-4	0.004540	0.006251	0.007580	0.009316
NTRA	37	3	2.0806e-4	8.3898e-5	3.512e-5	1.3177e-4	2.1321e-4	2.702e-4	4.1981e-4
NTRI	37	3	5.0636e-6	3.0313e-6	6.6e-7	2.1325e-6	3.945e-6	8.0325e-6	3.102e-5
NTRS	37	3	2.1407e-4	8.7868e-5	3.667e-5	1.3388e-4	2.204e-4	2.8178e-4	4.3848e-4
PHALD	37	0	14.296	9.7032	0.498	6.107	12.888	22.520	35.889
SLCA	37	3	1.1943e-4	8.0593e-5	4.89e-6	4.5677e-5	1.2316e-4	1.7556e-4	2.7981e-4
SUSP	37	0	0.0546	0.036687	0.0105	0.0273	0.0487	0.0621	0.1251
TEMP	37	0	5.3249	4.3443	0.6	2.28	2.78	6.39	15.53
DIST	37	4	-45.439	23.133	-78	-66.5	-33.5	-33	0
SAL	37	0	14.385	9.7065	0.4875	6.1125	12.74	22.346	31.586

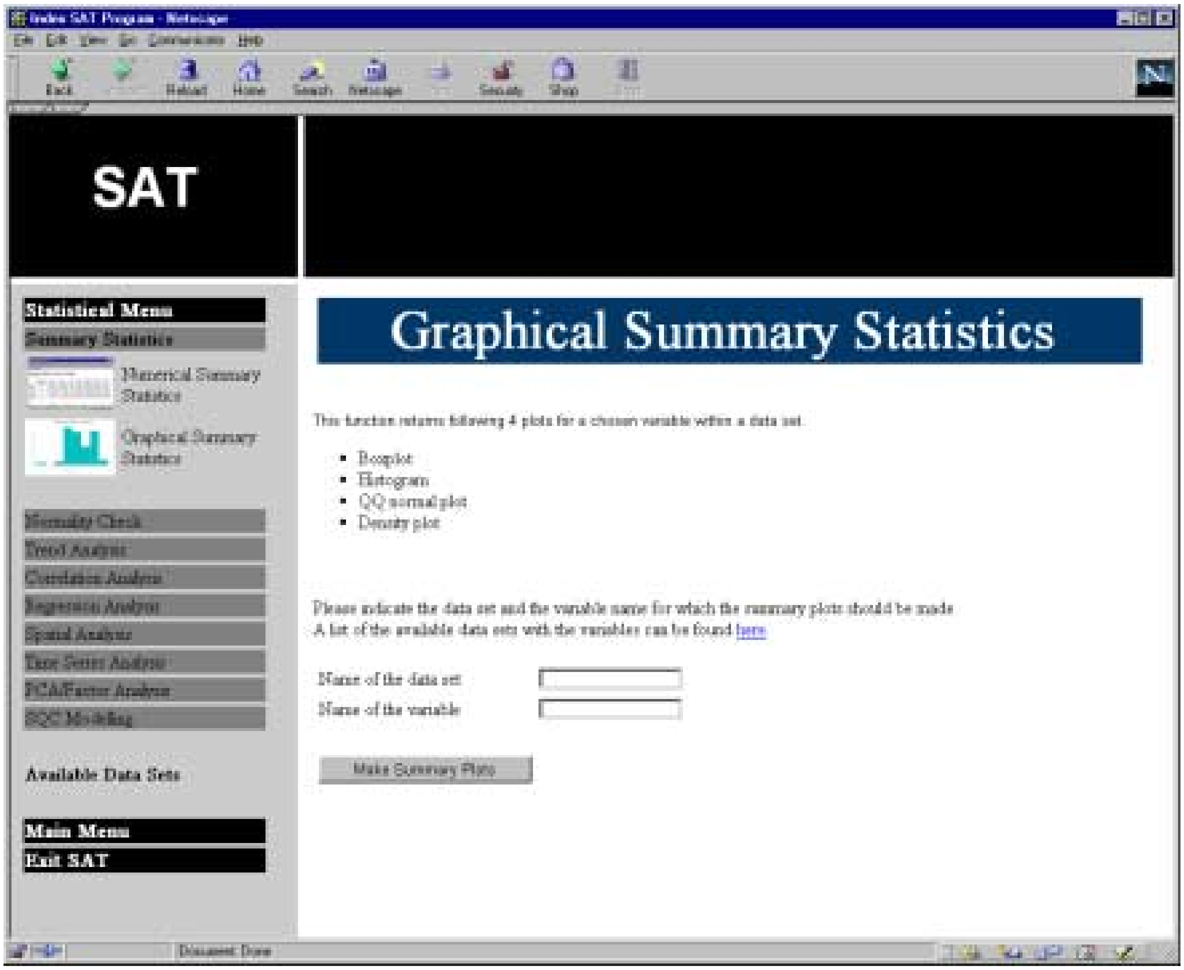
Figure 13: Example of the output of the Summary Statistics Function

Graphical Summary Statistics

The “**Graphical Summary Statistics**” function can be used to graphically summarize the data for a given variable.

Following graphs are created:

- Boxplot
- Histogram
- QQ Normal Plot
- Density Plot



Graphical Summary Statistics

This function returns following 4 plots for a chosen variable within a data set.

- Boxplot
- Histogram
- QQ normal plot
- Density plot

Please indicate the data set and the variable name for which the summary plots should be made
A list of the available data sets with the variables can be found here

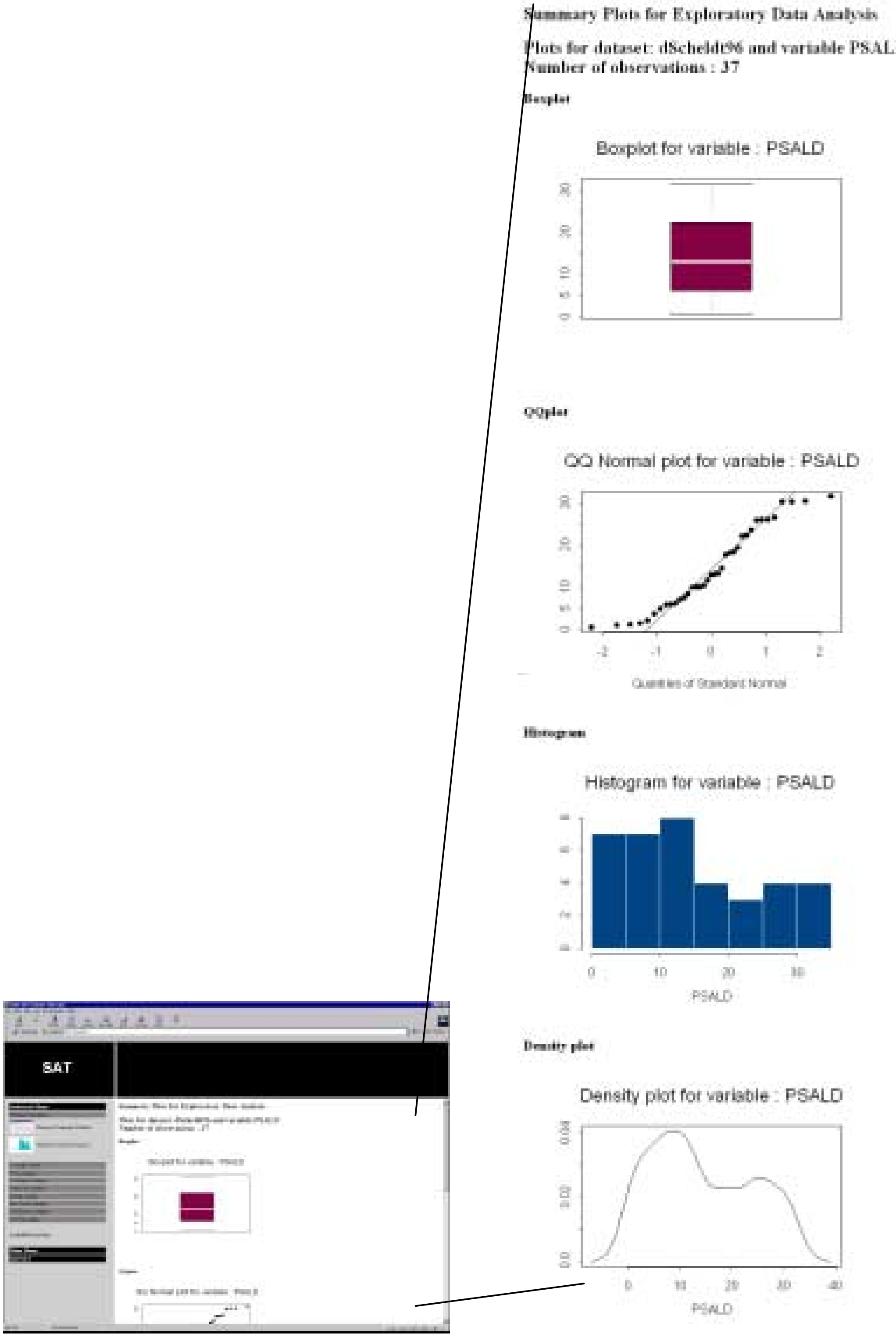
Name of the data set

Name of the variable

Figure 14: Graphical Summary Statistics function

The user should provide the name of the data set and the variable. Pushing the button “**Make Summary Plots**” results in a visual representation of the data.

An example of the result of this function is given in figure 15.



Summary Plots for Exploratory Data Analysis

Plots for dataset: dScheldt96 and variable PSALD
Number of observations : 37

Boxplot

Boxplot for variable : PSALD
0
5
10
20
30

QQplot

QQ Normal plot for variable : PSALD
0
5
10
20
30
-2
-1
0
1
2
Quantiles of Standard Normal

Histogram

Histogram for variable : PSALD
0
10
20
30
PSALD

Density plot

Density plot for variable : PSALD
0.0
0.02
0.04
0
10
20
30
40
PSALD
SAT

Figure 15: Example of the Graphical Summary Statistics function

Normality Check

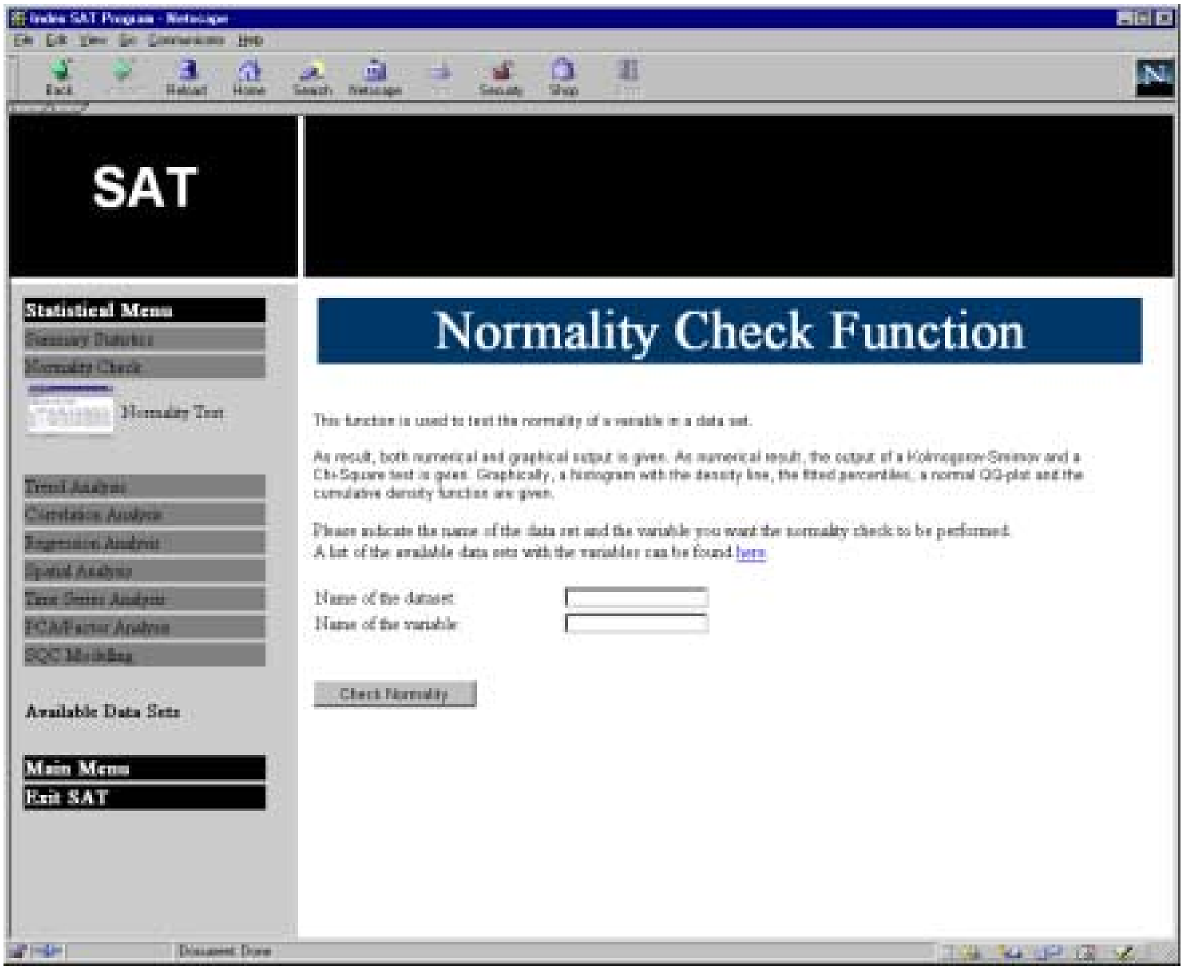
The “Normality Check” section allows to visually and numerically verify whether the data of a given variable can be assumed normally distributed.

Normality Test

The normality of a variable (Gaussian distribution) is be tested by means of the “**Normality Test**” function.

The user should indicate the name of the data set and the name of the variable of interest.

The Kolmogorov-Smirnov and the Chi-square test are performed.
In addition to these numerical results also graphical representations of the variable are derived. A histogram with density line, the fitted percentiles, a normal QQ-plot and the cumulative density plot are given.



Normality Check Function

This function is used to test the normality of a variable in a data set.

As result, both numerical and graphical output is given. As numerical result, the output of a Kolmogorov-Smirnov and a Chi-Square test is given. Graphically, a histogram with the density line, the fitted percentiles, a normal QQ-plot and the cumulative density function are given.

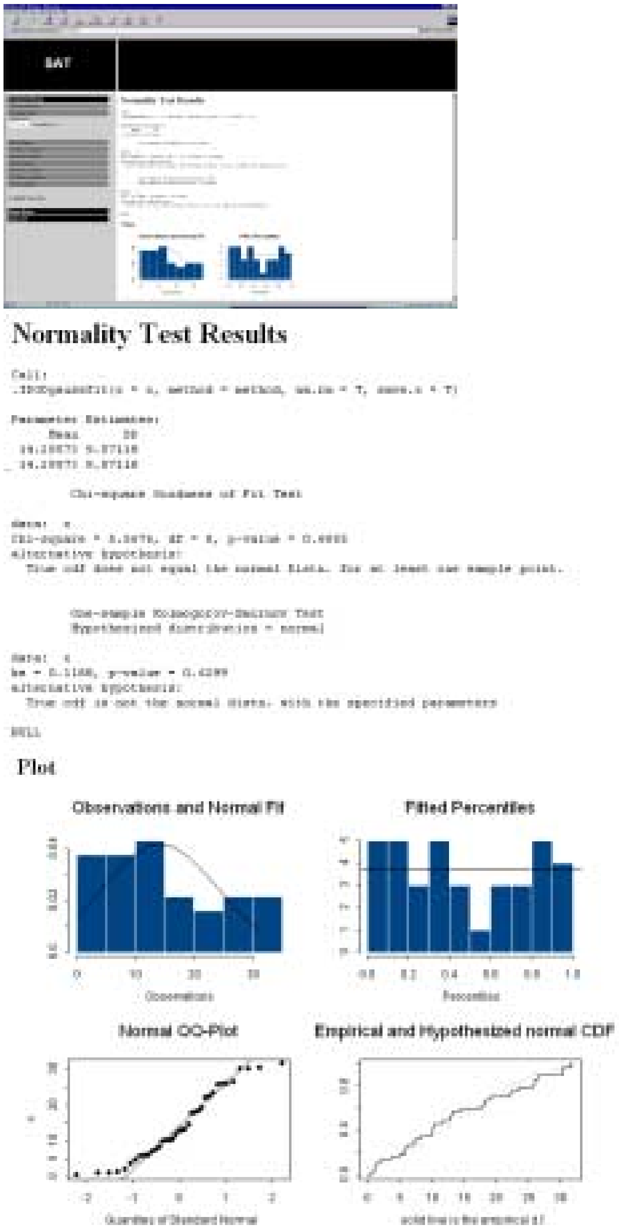
Please indicate the name of the data set and the variable you want the normality check to be performed.
A list of the available data sets with the variables can be found here

Name of the dataset
Name of the variable

Check Normality

Figure 16: Normality Check function

An example of the result of this function is given in following figure.



Normality Test Results

```
Call:
.IDOXgaussfit(x = x, method = method, se.fit = T, save.x = T)

Parameter Estimates:
     Mean      SD
 19.25571 9.87118
_ 19.25571 9.87118

        Chi-square Goodness of Fit Test

data:  x
Chi-square = 3.5679, df = 5, p-value = 0.6001
alternative hypothesis:
  True cdf does not equal the normal Distn. for at least one sample point.


        One-sample Kolmogorov-Smirnov Test
        Hypothesized distribution = normal

data:  x
ks = 0.1168, p-value = 0.4289
alternative hypothesis:
  True cdf is not the normal distn. with the specified parameters

NULL
```

Plot

Observations and Normal Fit
0
10
20
30
Observations
Fitted Percentiles
0.0
0.2
0.4
0.6
0.8
1.0
Percentiles
Normal QQ-Plot
-2
-1
0
1
2
Quantiles of Standard Normal
Empirical and Hypothesized normal CDF
0
5
10
15
20
25
30
solid line is the empirical d.f.

Figure 17: Example result of the normality check

Trend Analysis

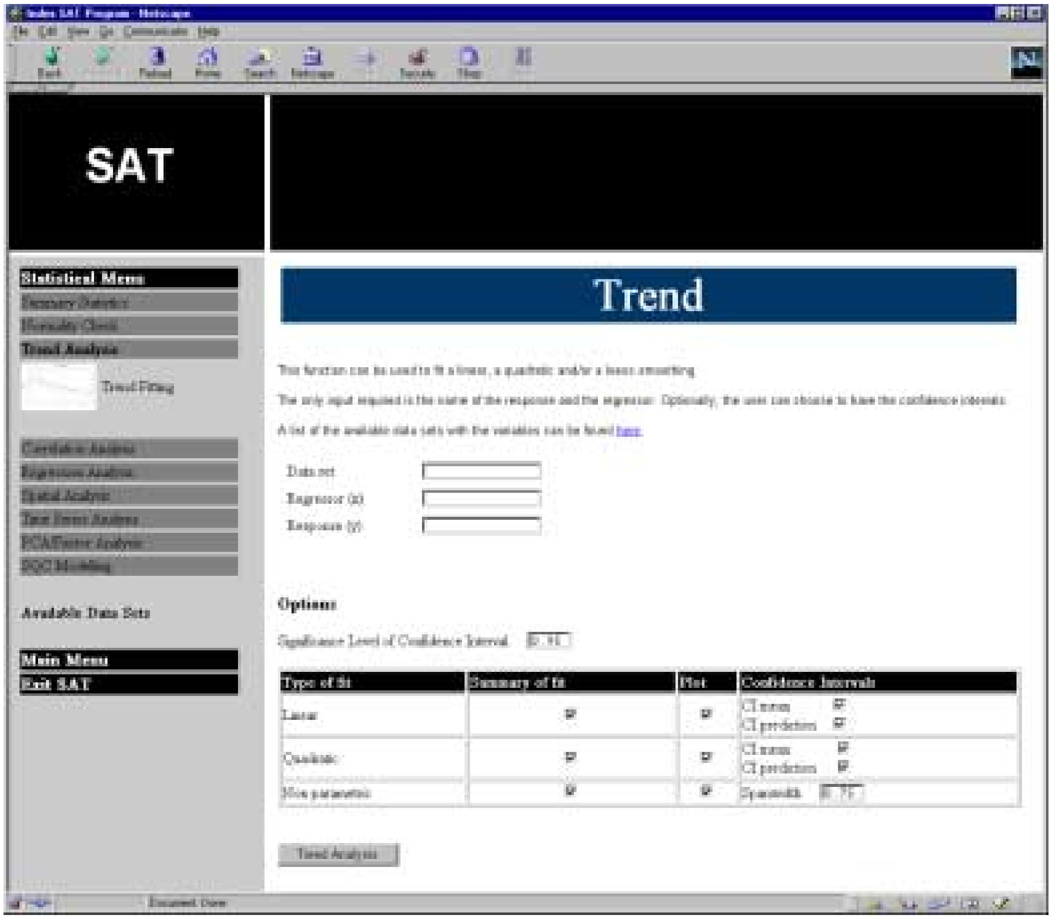
Trend Fitting

The trend fitting function is used to derive the relationship between a variable of interest (the response variable) and another variable such as distance (regressor variable).

The user should provide the name of the data set, the name of the regressor variable and the response variable.

The user can specify following options:

- Significance level of the confidence interval: this option indicates the probability level of the plotted confidence interval
- Linear: a linear regression model is used in this case, the user can indicate whether he desires a numerical summary, and/or a plot and he can indicate which type of confidence intervals he/she wishes to see on the plot.
- Quadratic: a quadratic regression model is used. The same options as in the linear case are available
- Non-parametric: a non-parametric loess smoothing is performed. The span width of the model is to be chosen by the user.



Trend

This function can be used to fit a linear, a quadratic and/or a loess smoothing

The only input required is the name of the response and the regressor. Optionally, the user can choose to have the confidence intervals

A list of the available data sets with the variables can be found here.

Data set
Regressor (x)
Response (y)

Options

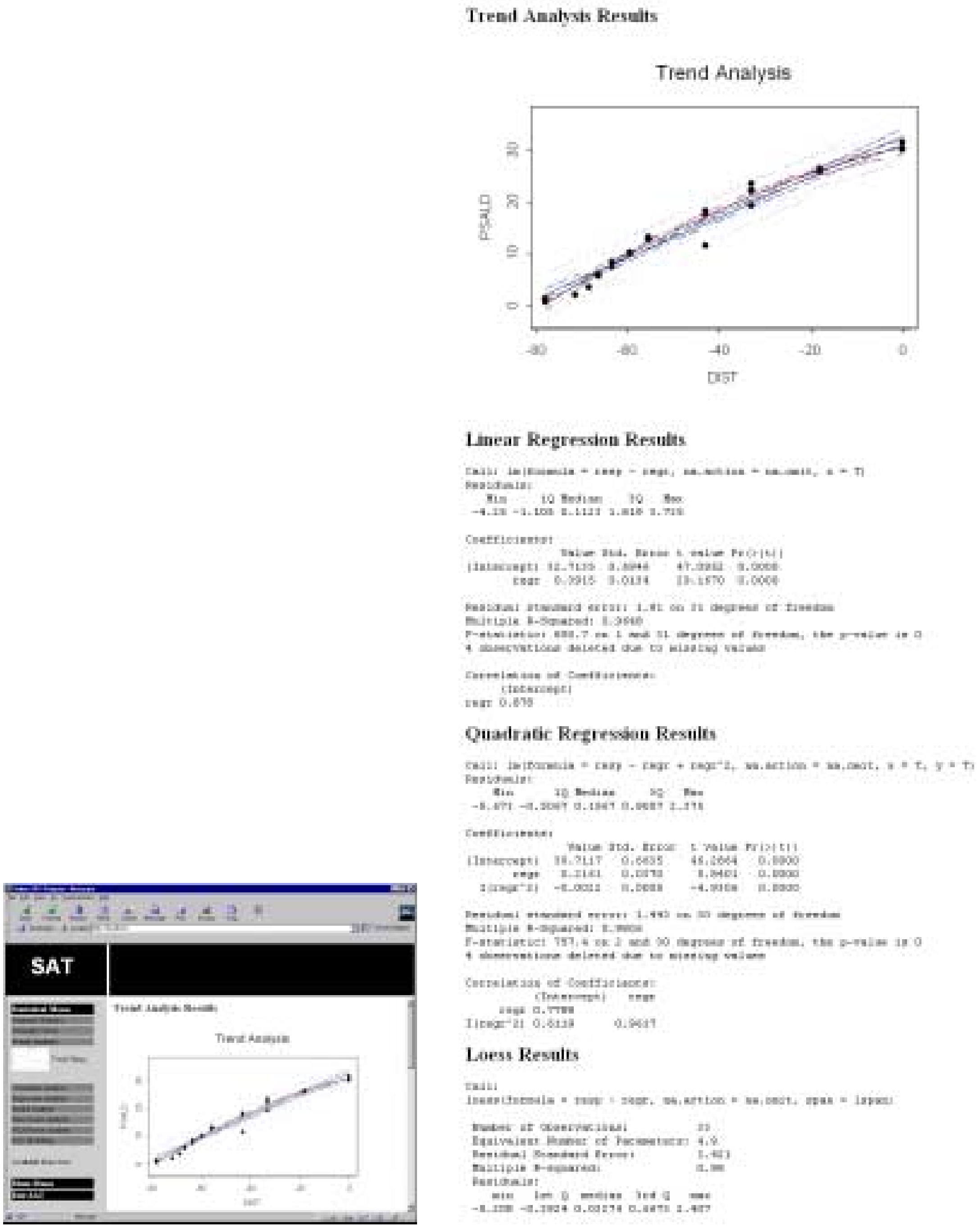
Significance Level of Confidence Interval 0.95

Type of fit	Summary of fit	Plot	Confidence Intervals
Linear	☑	☑	CI mean ☑ CI prediction ☑
Quadratic	☑	☑	CI mean ☑ CI prediction ☑
Non parametric	☑	☑	Spanwidth 0.75

Trend Analysis

Figure 18: The trend fitting function

An example of the result of this function is given in figure 19.



Trend Analysis Results

Linear Regression Results

```
Call: lm(formula = resp ~ regr, na.action = na.omit, x = T)
Residuals:
   Min     1Q Median    3Q   Max
 -4.35 -1.105 0.1123 1.018 3.715

Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept) 32.7135  0.6946    47.0952  0.0000
       regr  0.3915  0.0134    29.1870  0.0000

Residual standard error: 1.91 on 31 degrees of freedom
Multiple R-Squared: 0.9649
F-statistic: 851.7 on 1 and 31 degrees of freedom, the p-value is 0
4 observations deleted due to missing values

Correlation of Coefficients:
     (Intercept)
regr 0.878
```

Quadratic Regression Results

```
Call: lm(formula = resp ~ regr + regr^2, na.action = na.omit, x = T, y = T)
Residuals:
    Min      1Q Median     3Q   Max
 -5.671 -0.3087 0.1967 0.8687 2.376

Coefficients:
                Value Std. Error  t value Pr(>|t|)
(Intercept)  30.7117   0.6635    46.2864   0.0000
       regr   0.2161   0.0270     8.0401   0.0000
  I(regr^2)  -0.0012   0.0003    -4.9308   0.0000

Residual standard error: 1.493 on 30 degrees of freedom
Multiple R-Squared: 0.9806
F-statistic: 757.4 on 2 and 30 degrees of freedom, the p-value is 0
4 observations deleted due to missing values

Correlation of Coefficients:
          (Intercept)    regr
     regr 0.7789
I(regr^2) 0.6128      0.9627
```

Loess Results

```
Call:
loess(formula = resp ~ regr, na.action = na.omit, span = lspan)

 Number of Observations:          33
 Equivalent Number of Parameters: 4.9
 Residual Standard Error:         1.423
 Multiple R-squared:              0.98
 Residuals:
    min   1st Q   median  3rd Q   max
 -5.208 -0.3924 0.03274 0.4870 2.407
```

Figure 19: Example result of the trend fit function

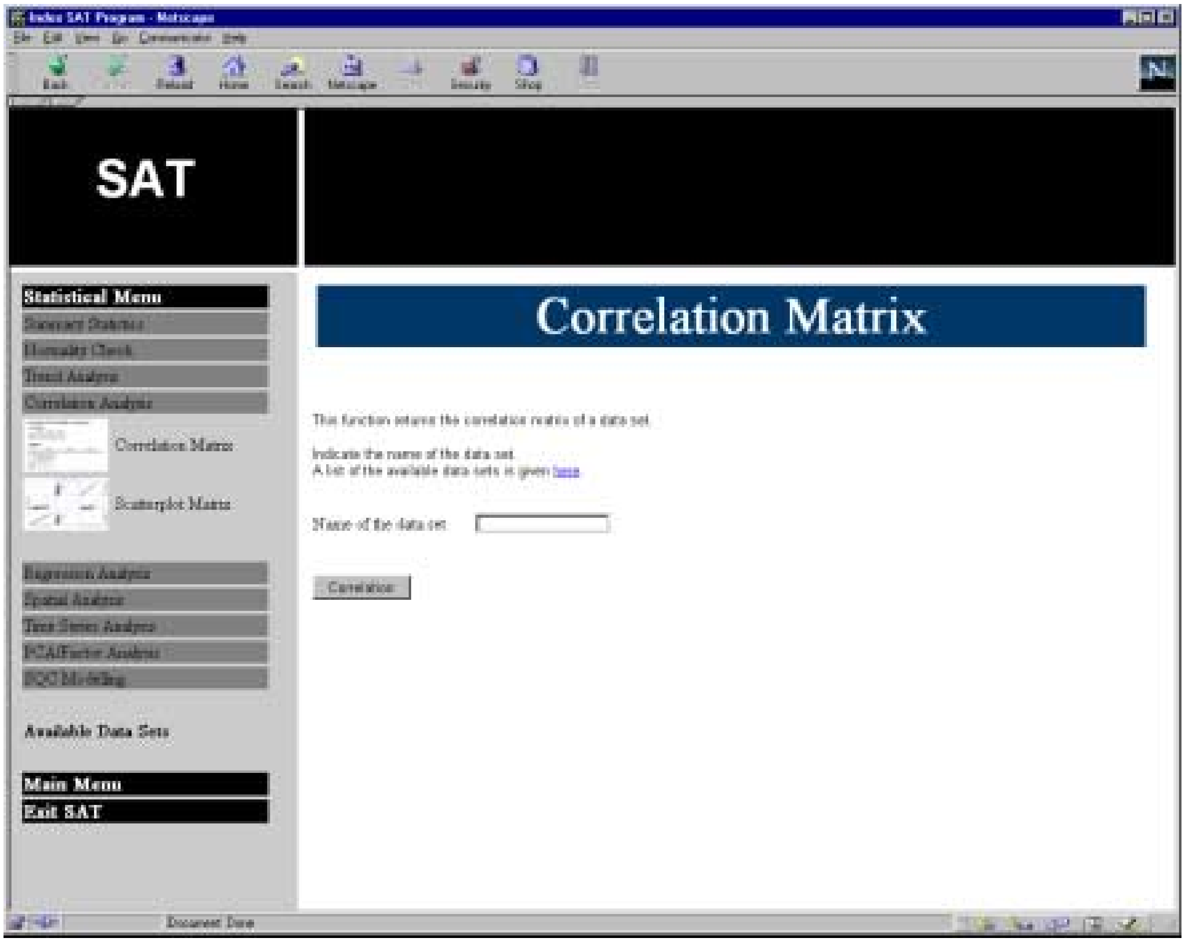
Correlation Analysis

The correlation matrix section of the statistical analysis functions consists of 2 functions. The Correlation Matrix function gives the numerical results of a correlation analysis, while the Scatter plot matrix function derives a visual representation that illustrates the correlation between the variables in a data set.

Correlation Matrix

The correlation matrix function derives the correlation between the variables in a data set. The correlation matrix is calculated as well as the variance of each variable and the number of observations for each pair of variables.

The user should provide the name of the data set to which this option should be applied.



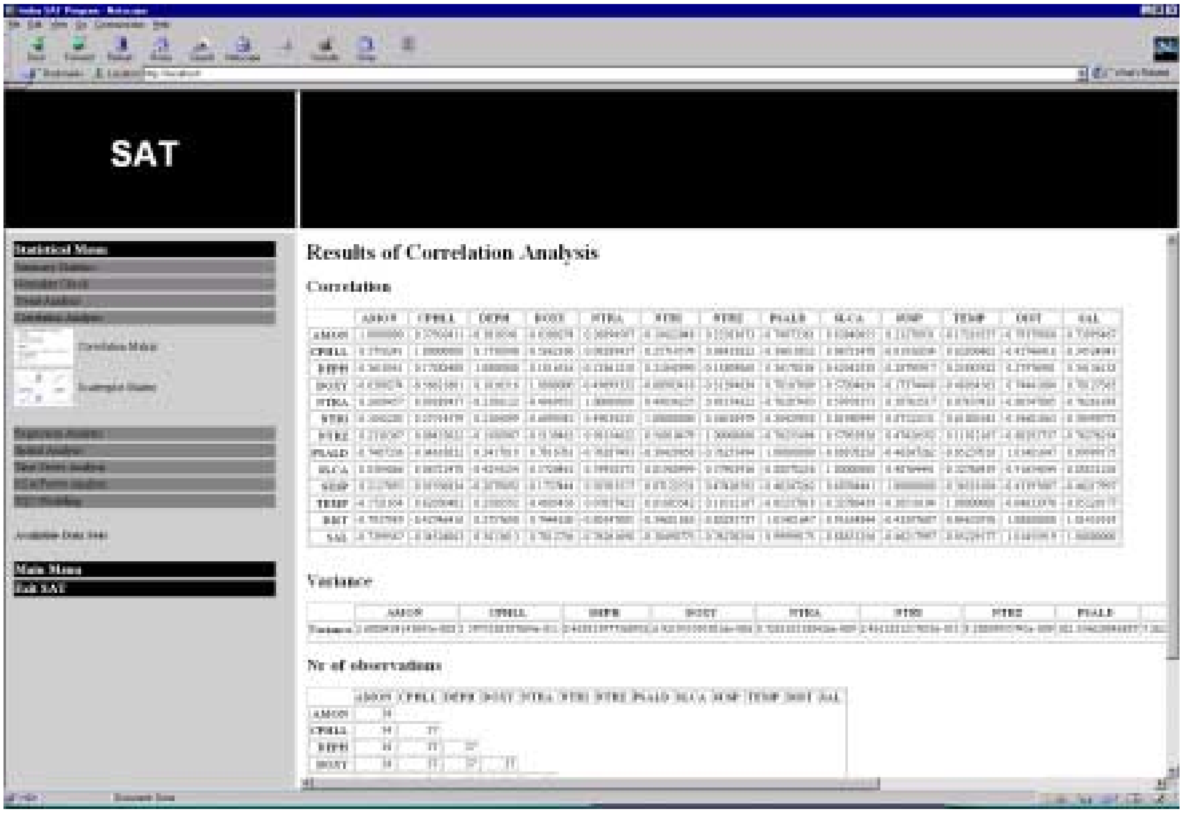
This function returns the correlation matrix of a data set.

Indicate the name of the data set.
A list of the available data sets is given here

Name of the data set

Figure 20: The correlation matrix function

An example of the result of this function is given in figure 21.



Results of Correlation Analysis

Correlation

	AMON	CPHLL	DEPH	DOXY	NTRA	NTRI	NTRZ	PSALD	SLCA	SOSP	TEMP	DOUT	SAL
AMON	1.0000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
CPHLL	[illegible]	1.0000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
DEPH	[illegible]	[illegible]	1.0000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
DOXY	[illegible]	[illegible]	[illegible]	1.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
NTRA	[illegible]	[illegible]	[illegible]	[illegible]	1.0000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
NTRI	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1.0000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
NTRZ	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1.0000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
PSALD	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1.0000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
SLCA	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1.0000000	[illegible]	[illegible]	[illegible]	[illegible]
SOSP	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1.0000000	[illegible]	[illegible]	[illegible]
TEMP	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1.0000000	[illegible]	[illegible]
DOUT	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1.0000000	[illegible]
SAL	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1.0000000

Variance

	AMON	CPHLL	DEPH	DOXY	NTRA	NTRI	NTRZ	PSALD
Variance	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Nr of observations

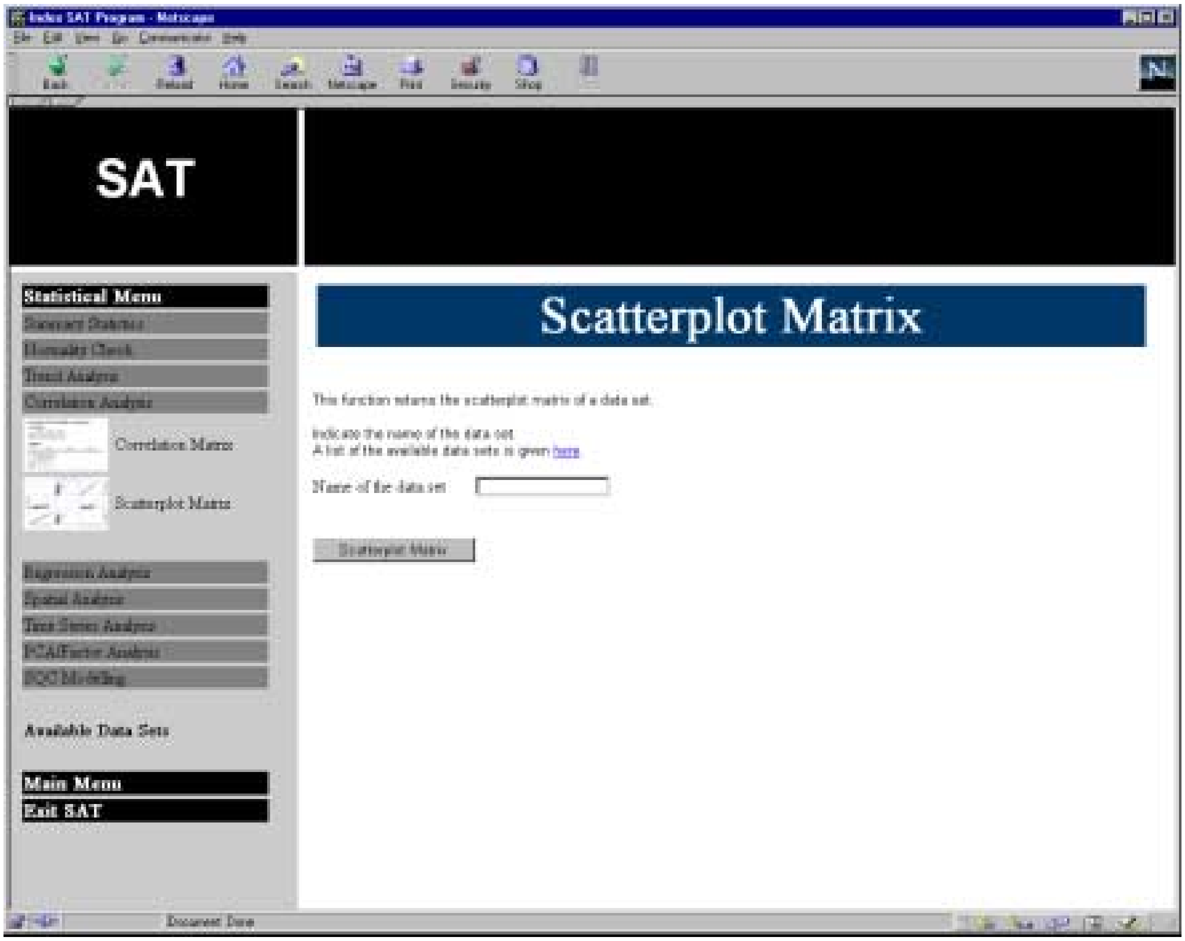
	AMON	CPHLL	DEPH	DOXY	NTRA	NTRI	NTRZ	PSALD	SLCA	SOSP	TEMP	DOUT	SAL
AMON	18												
CPHLL	18	77											
DEPH	18	77	97										
DOXY	18	77	77	77									

Figure 21: Example result of the correlation function

Scatter plot Matrix

The scatter plot matrix function is used to visually illustrate the correlation between the different variables in a data set.
For each pair of variables a scatter plot is made en shown in a matrix form.

As input the name of the data set should be given.



This function returns the scatterplot matrix of a data set.

Indicate the name of the data set.
A list of the available data sets is given here.

Name of the data set

Figure 22: Scatter plot matrix function

An example of the result of this function is given in figure 23.

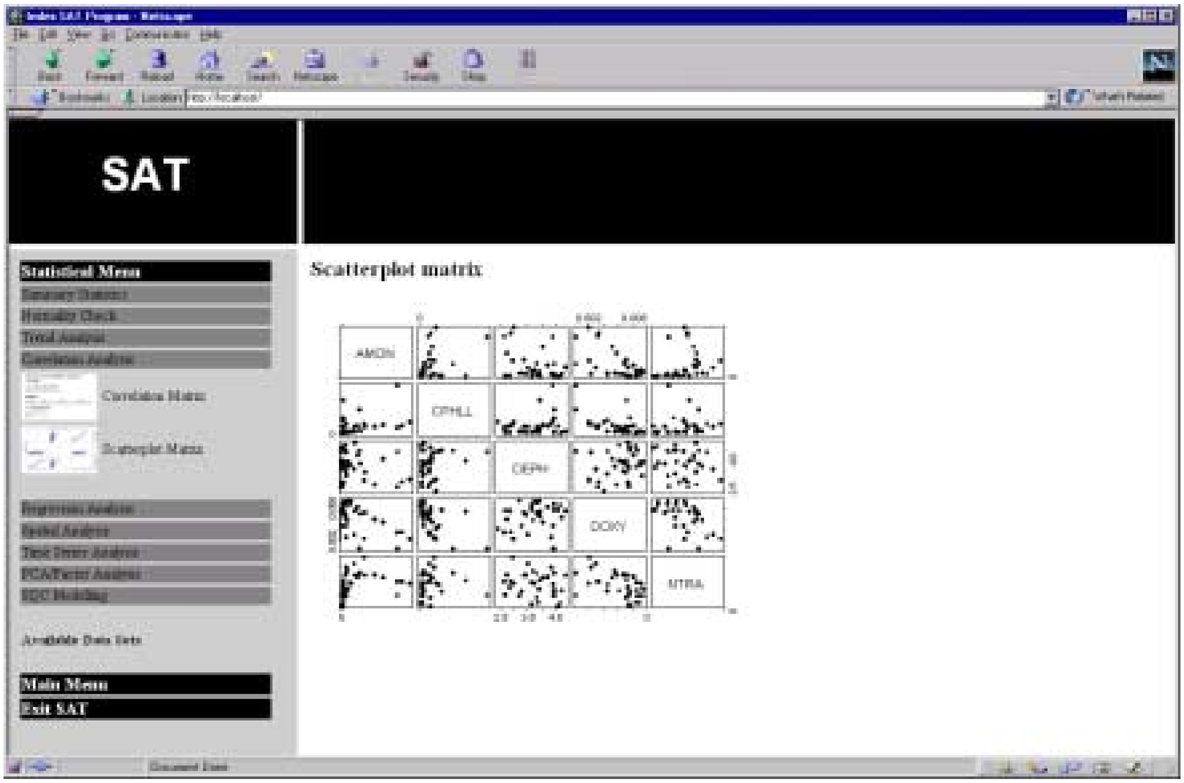


Figure 23: Example of the scatter plot matrix function

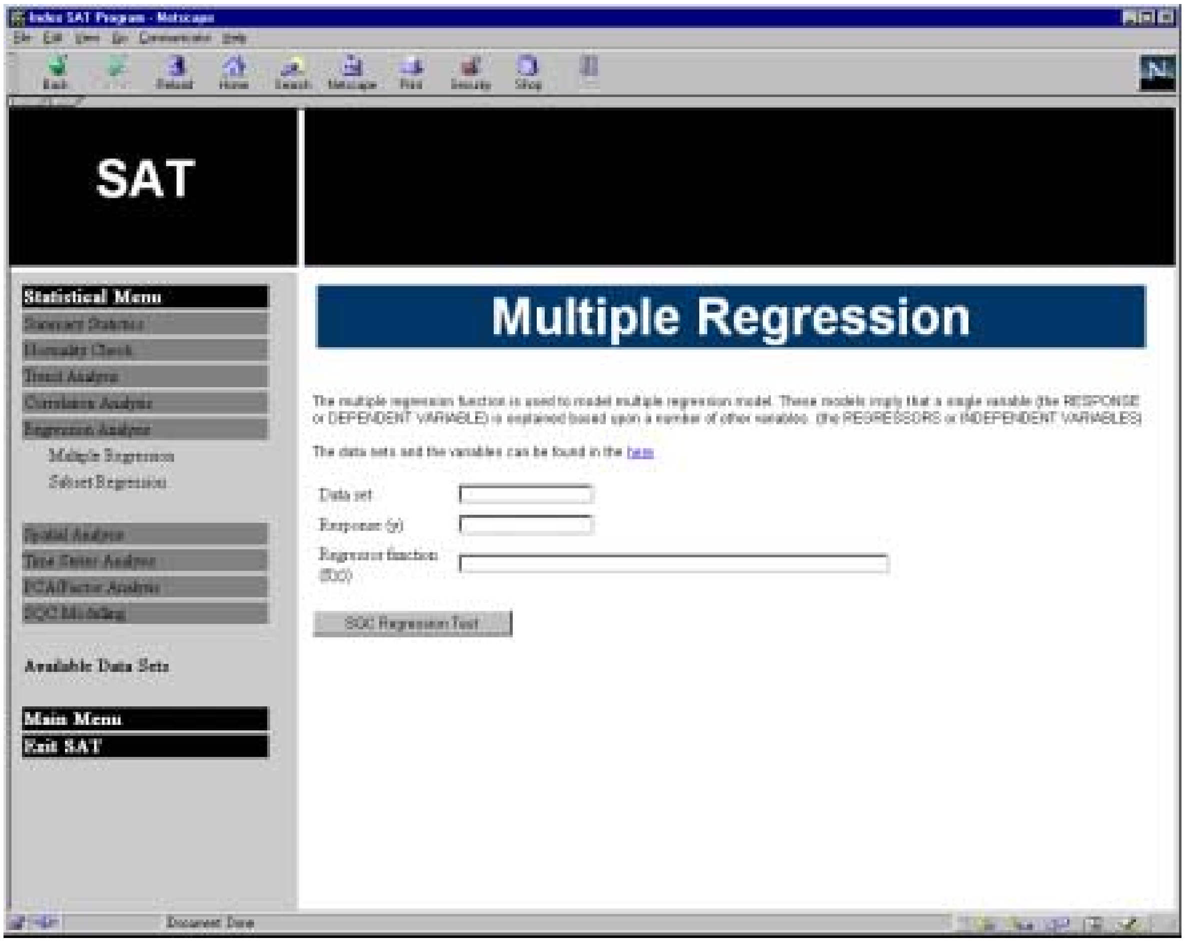
Regression Analysis

The regression analysis section comprises 2 different analyses: the first analysis, the multiple regression function, is used to derive the parameters of a user-defined regression model. The second analysis, the subset regression function, is used to find a subset of regressor variables out of a list of possible regressors such that all the regressor variables are significant at a given level.

Multiple Regression

The multiple regression function is used to perform regression with more then 1 regressor variable.

The data set should be given, the name of the response variable and the regression function.
The regression function specifies the regression model and the variables to be considered as regressors.
For instance:
NTRI+NTRA
indicates that the regression is performed using the regressors NTRI and NTRA
NTRI+NTRA+NTRI:NTRA
indicates that the regression is performed using the regressors NTRI and NTRA
and that also interaction between NTRI and NTRA is allowed



The multiple regression function is used to model multiple regression model. These models imply that a single variable (the RESPONSE or DEPENDENT VARIABLE) is explained based upon a number of other variables. (the REGRESSORS or INDEPENDENT VARIABLES)

The data sets and the variables can be found in the here

Data set
Response (y)
Regressor function (f(x))

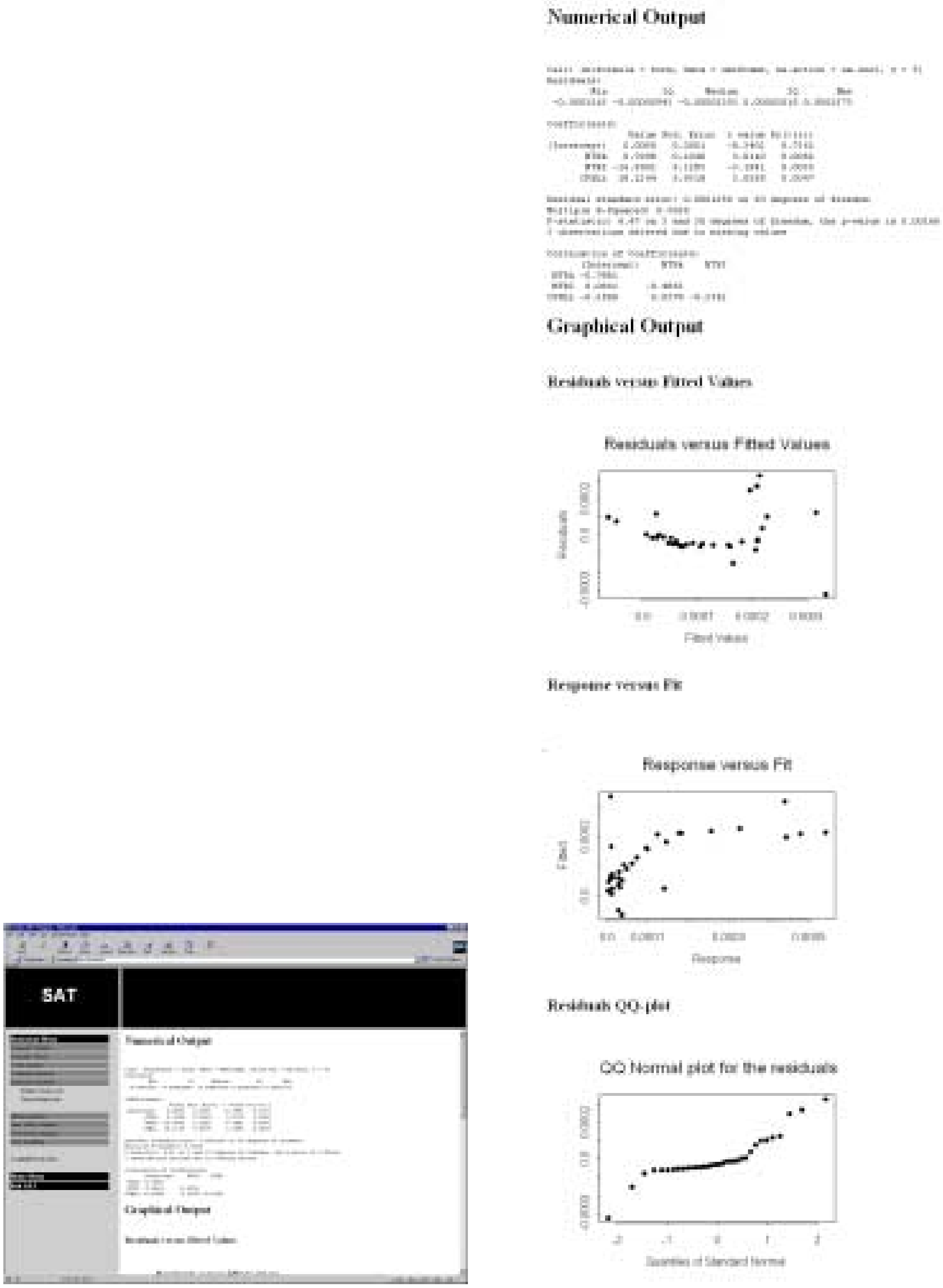
Figure 24: The Multiple Regression Function

The output is split up in a numerical part and a graphical part.

In the numerical part the estimates for the coefficients are shown, with their standard error, the t-value and the p value. The R-square value and the residual standard error are also shown together with the p-value of the model.

In the graphical output, a plot of the residuals versus the fitted values, a plot of the response versus the fitted values and a QQ-normal plot are given.

An example output is shown below:



Numerical Output

```
Call: [illegible]
Residuals:
       Min          1Q      Median          3Q       Max
 [illegible]

Coefficients:
                Value Std. Error  t value Pr(>|t|)
 (Intercept)  [illegible]
 [illegible]

Residual standard error: [illegible] on [illegible] degrees of freedom
Multiple R-Squared: [illegible]
F-statistic: [illegible] on 3 and [illegible] degrees of freedom, the p-value is [illegible]
[illegible] observations deleted due to missing values

Correlation of Coefficients:
       (Intercept)     [illegible]    [illegible]
 [illegible]
```

Graphical Output

Residuals versus Fitted Values

Response versus Fit

Residuals QQ-plot

QQ Normal plot for the residuals
-2
-1
0
1
2
Quantiles of Standard Normal
SAT
Numerical Output
Graphical Output
Residuals versus Fitted Values

Figure 25: Example of the multiple regression function

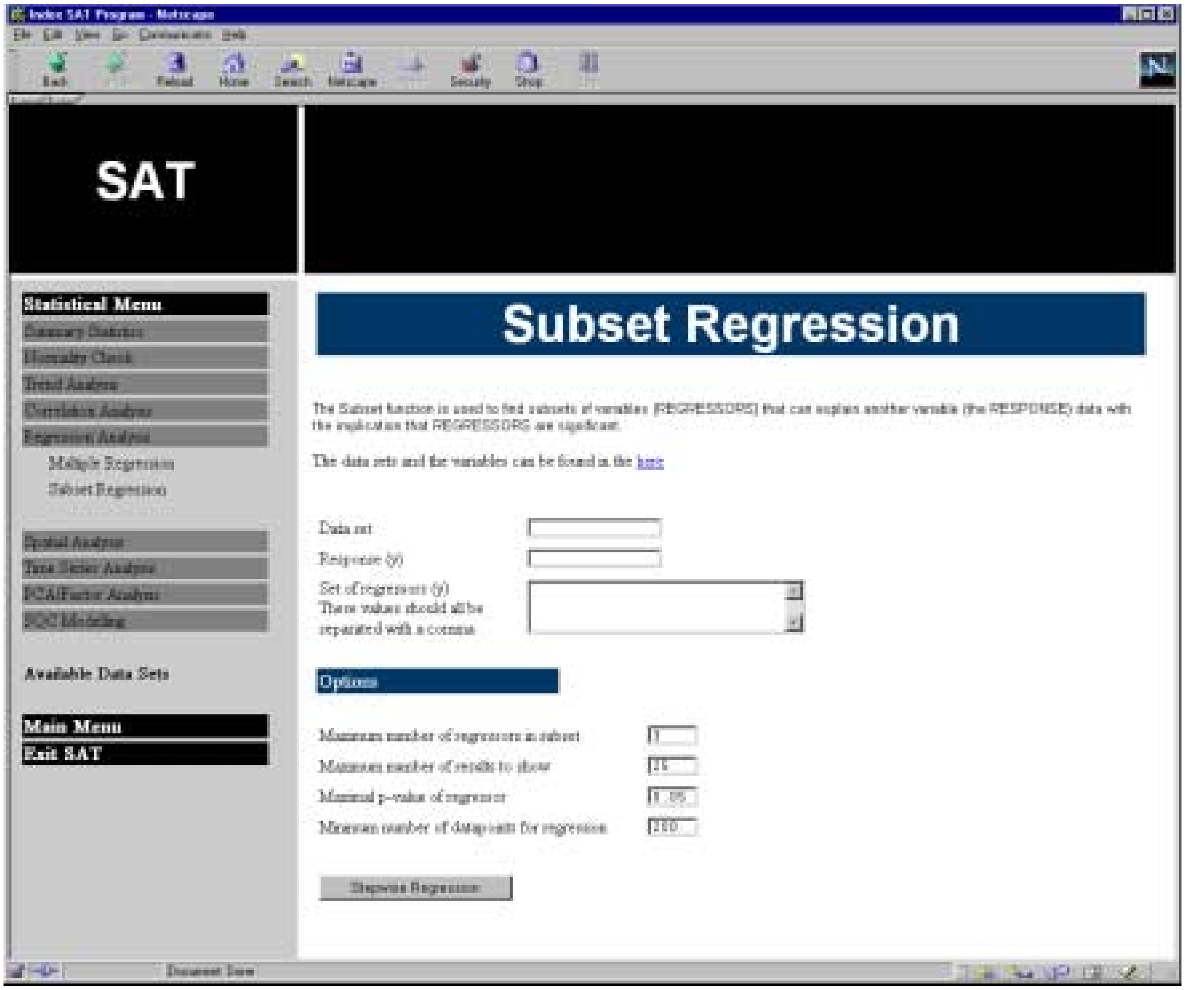
Subset Regression

The subset regression function is used automatically identify the subset of regressor variables that significantly explain the response variable.

The user should specify the data set, the response variable and the list of potential regressor variables.

The user can also choose the maximum size of the subset, the maximum number of subset results that should be derived, the maximum p-value for a regressor to be taken into account in a model and the minimum of data points that should be available to derive results for a subset (i.e. because of missing data, the number of data points can substantially differ depending on which combination of regressor variables considered)

As a result the requested number of subsets are given in order of decreasing R²adjusted value.



Subset Regression

The Subset function is used to find subsets of variables (REGRESSORS) that can explain another variable (the RESPONSE) data with the implication that REGRESSORS are significant.

The data sets and the variables can be found in the here

Data set

Response (y)

Set of regressors (y)
These values should all be separated with a comma

Maximum number of regressors in subset 3

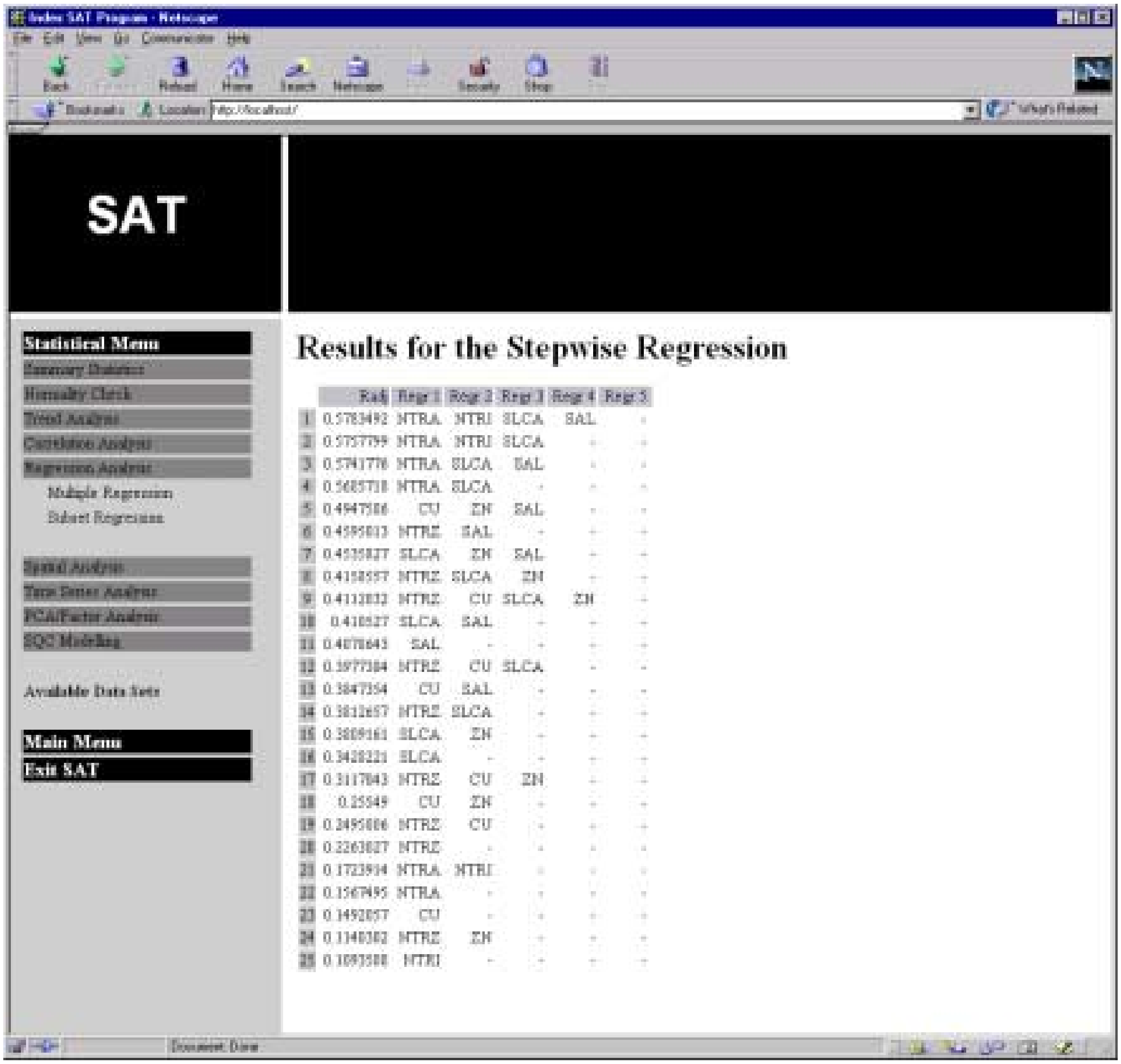
Maximum number of results to show 25

Maximal p-value of regressor 0.05

Minimum number of datapoints for regression 200

Figure 26: The subset regression function

An example of a result of the subset regression is given here



Results for the Stepwise Regression

	Radj	Regr 1	Regr 2	Regr 3	Regr 4	Regr 5
1	0.5783492	NTRA	NTRI	SLCA	SAL	-
2	0.5757799	NTRA	NTRI	SLCA	-	-
3	0.5741776	NTRA	SLCA	SAL	-	-
4	0.5685718	NTRA	SLCA	-	-	-
5	0.4947586	CU	ZN	SAL	-	-
6	0.4595013	NTRZ	SAL	-	-	-
7	0.4535827	SLCA	ZN	SAL	-	-
8	0.4158557	NTRZ	SLCA	ZN	-	-
9	0.4112832	NTRZ	CU	SLCA	ZN	-
10	0.410527	SLCA	SAL	-	-	-
11	0.4078643	SAL	-	-	-	-
12	0.3977384	NTRZ	CU	SLCA	-	-
13	0.3847354	CU	SAL	-	-	-
14	0.3812657	NTRZ	SLCA	-	-	-
15	0.3809161	SLCA	ZN	-	-	-
16	0.3428221	SLCA	-	-	-	-
17	0.3117843	NTRZ	CU	ZN	-	-
18	0.25549	CU	ZN	-	-	-
19	0.2495806	NTRZ	CU	-	-	-
20	0.2263827	NTRZ	-	-	-	-
21	0.1723934	NTRA	NTRI	-	-	-
22	0.1567495	NTRA	-	-	-	-
23	0.1492057	CU	-	-	-	-
24	0.1140302	NTRZ	ZN	-	-	-
25	0.1093588	NTRI	-	-	-	-

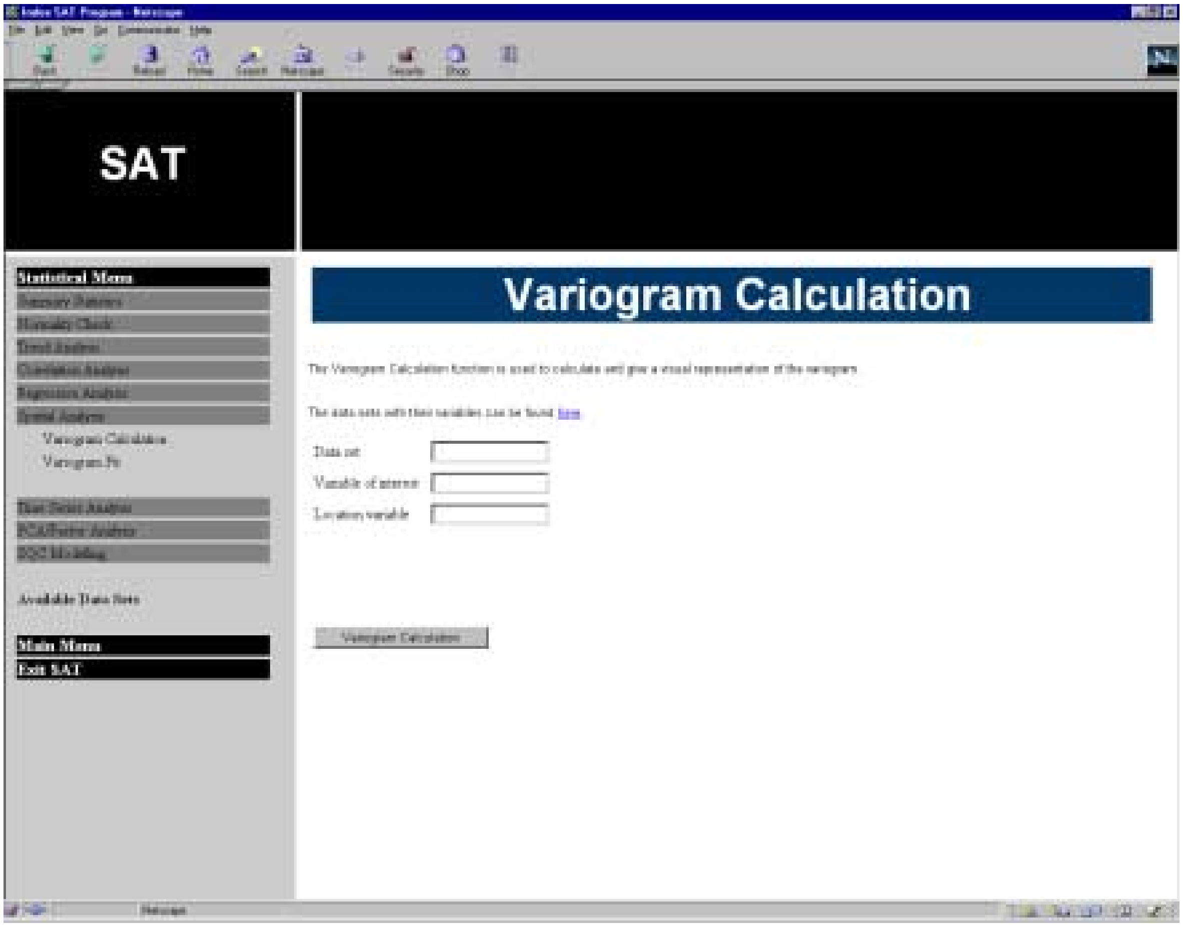
Figure 27: Results of a stepwise regression function

Spatial Analysis

Variogram Calculation

The variogram calculation function is used to calculate a variogram for a variable. The user should specify the data set, the variable of interest and the variable that should be used as location variable.

As result a figure with the variogram and a list of results is presented. The first list is a list with combined results. The second list gives the details for each difference in distance.



Variogram Calculation

The Variogram Calculation function is used to calculate and give a visual representation of the variogram.

The data sets with their variables can be found here.

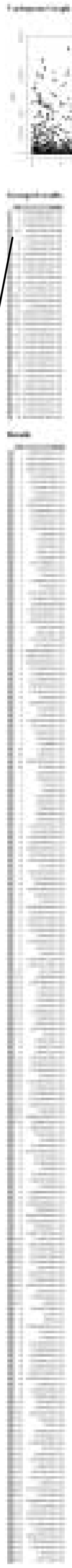
Data set

Variable of interest

Location variable

Figure 28: The Variogram Calculation Function

An example of a result of is given here



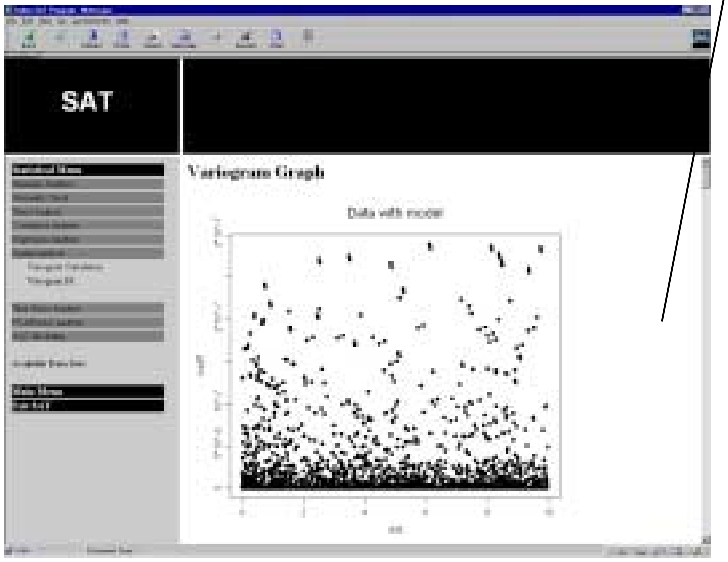


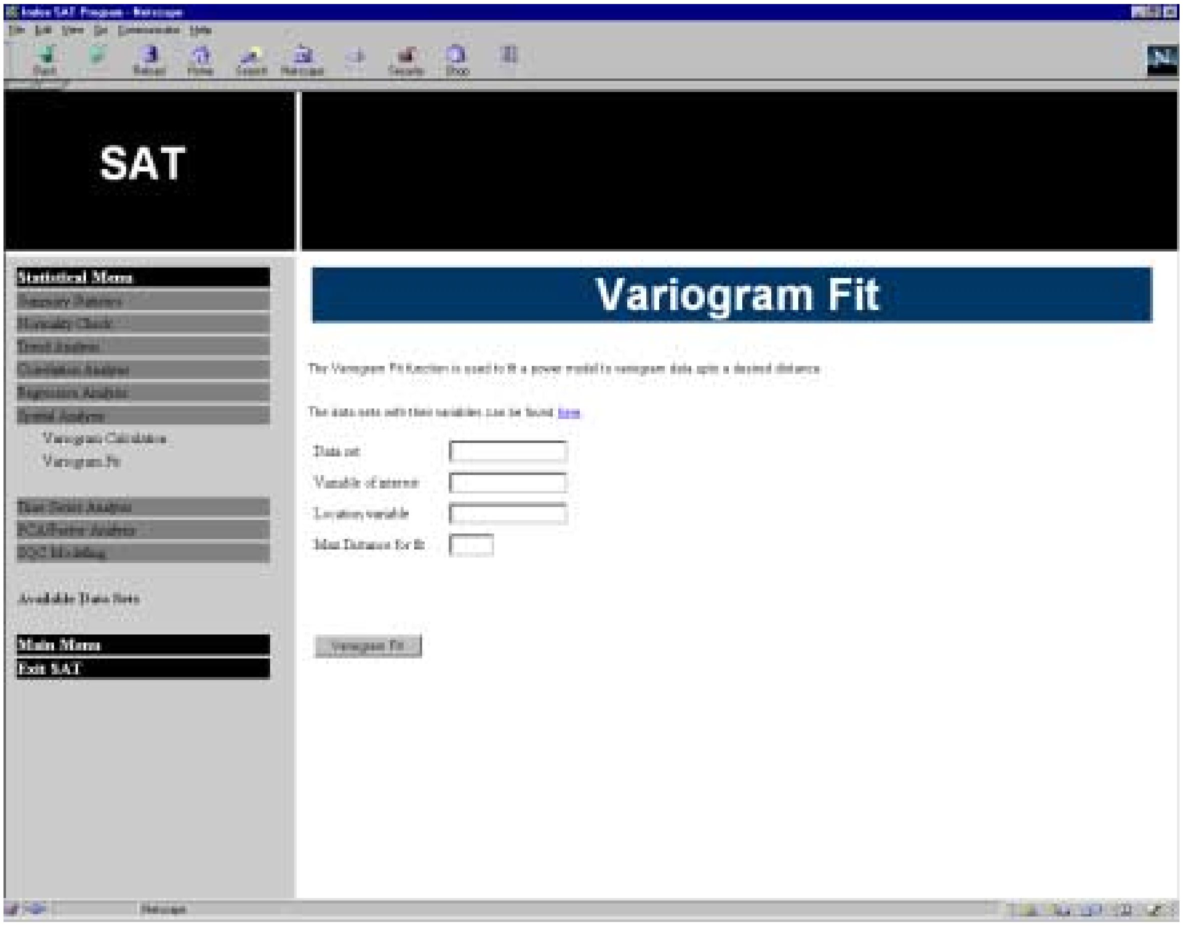
Figure 29:Example of the output of the variogram calculation function

Variogram Fit

The variogram fit function is used to fit a power model $f(x) = a + bx^c$ to the calculated variogram.

The user should specify the data set, the variable of interest and the variable that should be used as location variable. Next to this also the distance up to which the model should be fitted should be given.

As result the variogram graph is given next to the values of the parameters a, b and c.



The Variogram Fit function is used to fit a power model to variogram data upto a desired distance

The data sets with their variables can be found here

Data set
Variable of interest
Location variable
Max Distance for fit

Figure 30: Variogram Fit function

An example of a result of is given here

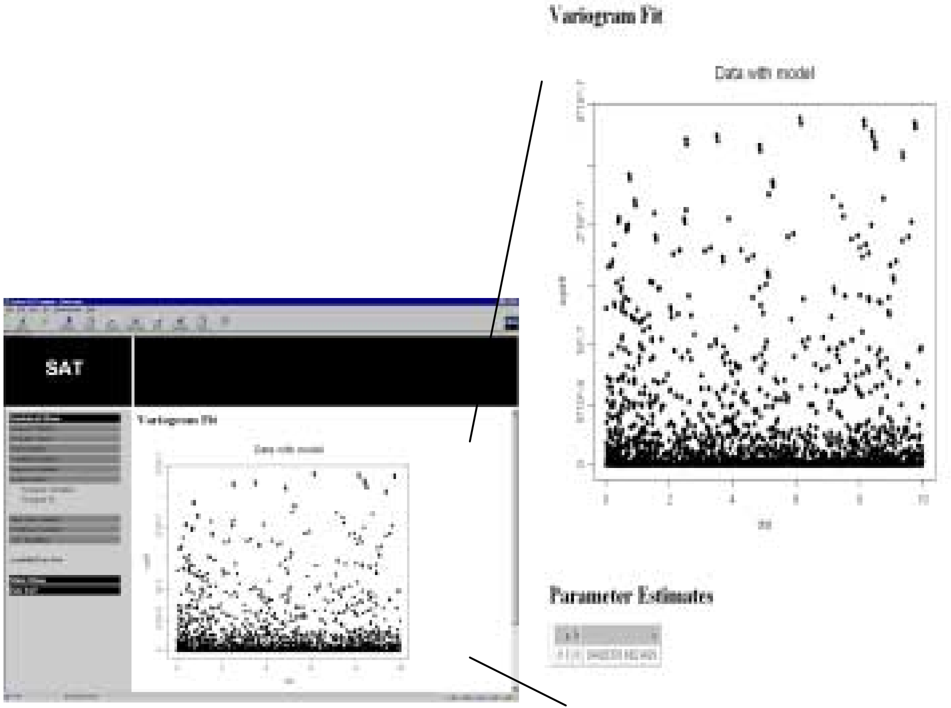


Figure 31: Example of the output of the variogram fit function

Time Series Analysis

Graphical Exploration

This function is not implemented at this moment

ARMA Modeling

This function is not implemented at this moment

PCA/Factor Analysis

Graphical Exploration

This function is not implemented at this moment

Factor Analysis

This function is not implemented at this moment

SQC Modeling

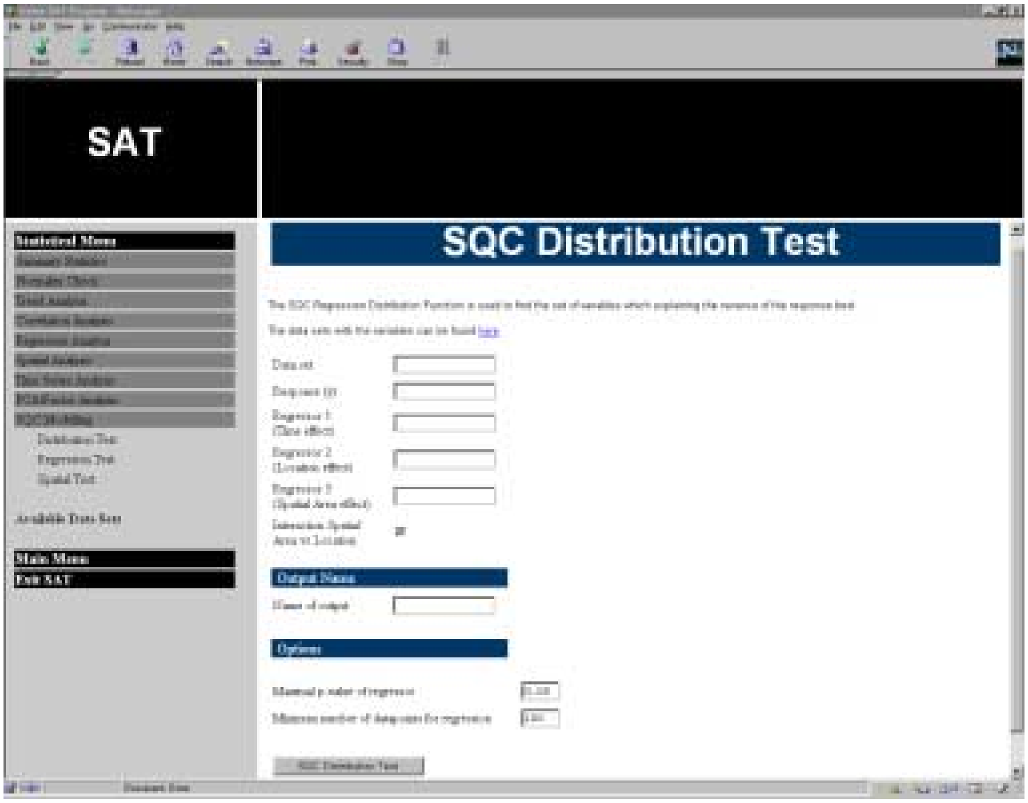
Distribution Test

The SQC Distribution Test function is used to derive a statistical model for the distribution test in the SQC program. This test is used to test the measurement of a variable against some data (meta) data.

The user should provide the name of the data set, the name of the response variable, the name of the variable indicating the time effect, the variable indicating the location effect and the name of a variable indication the spatial area effect (e.g. one could use a variable indicating whether a point is located in sea or in the river Scheldt).
The user can make the model more accurate by looking at the interaction between the spatial area and the location variable.

The function is designed to look for the best possible model where all the regressors in the final model have a p-value lower then the maximum p-value given in the options.
A second option the user can provide is the minimal number of data points that are necessarily to determine the model. In this way the user can prevent to have a final model that is based upon a very low number of data points.

Finally the user (normally the database administrator) should provide the name of the output file. The model is saved under this name and can by means of the Splus2Buffer program be transferred to the Statresults database.



SQC Distribution Test

The SQC Regression Distribution Function is used to find the set of variables which explaining the variance of the response best

The data sets with the variables can be found here

Data set

Response (y)

Regressor 1 (Time effect)

Regressor 2 (Location effect)

Regressor 3 (Spatial Area effect)

Interaction Spatial Area vs Location ☑

Output Name

Name of output

Options

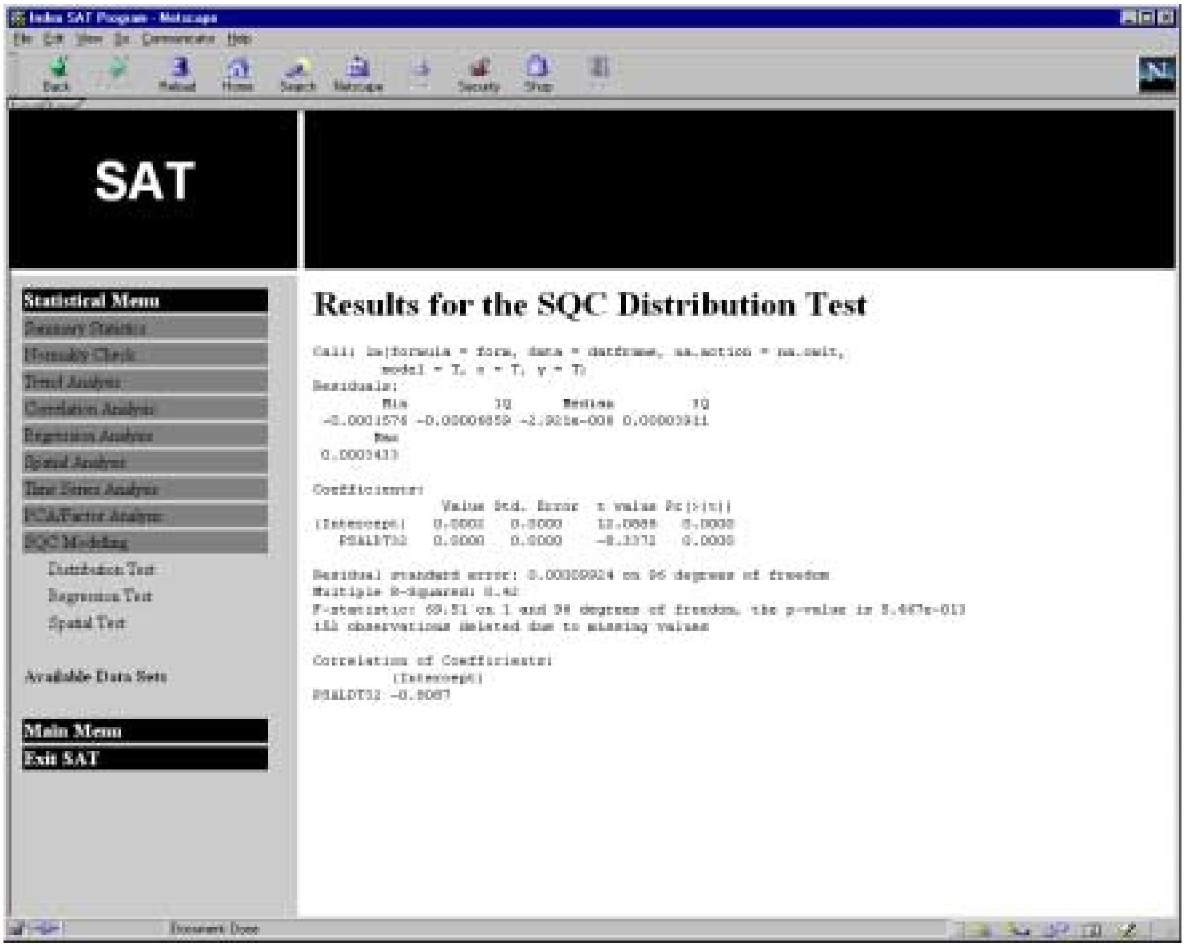
Maximal p-value of regressor

Minimum number of data points for regression

SQC Distribution Test

Figure 32: SQC Distribution Test

An example of a result of this function is given in the following figure.



Results for the SQC Distribution Test

```
Call: lm(formula = form, data = datframe, na.action = na.omit,
        model = T, x = T, y = T)
Residuals:
        Min           1Q       Median         3Q
 -0.0001576 -0.00006859 -2.921e-008 0.00003911
       Max
 0.0003433

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)   0.0002   0.0000    12.0888   0.0000
   PSALDT32   0.0000   0.0000    -8.3372   0.0000

Residual standard error: 0.00009924 on 96 degrees of freedom
Multiple R-Squared: 0.42
F-statistic: 69.51 on 1 and 96 degrees of freedom, the p-value is 5.467e-013
151 observations deleted due to missing values

Correlation of Coefficients:
          (Intercept)
PSALDT32 -0.9087
```

Figure 33: Example of an output of the SQC distribution test

Regression Test

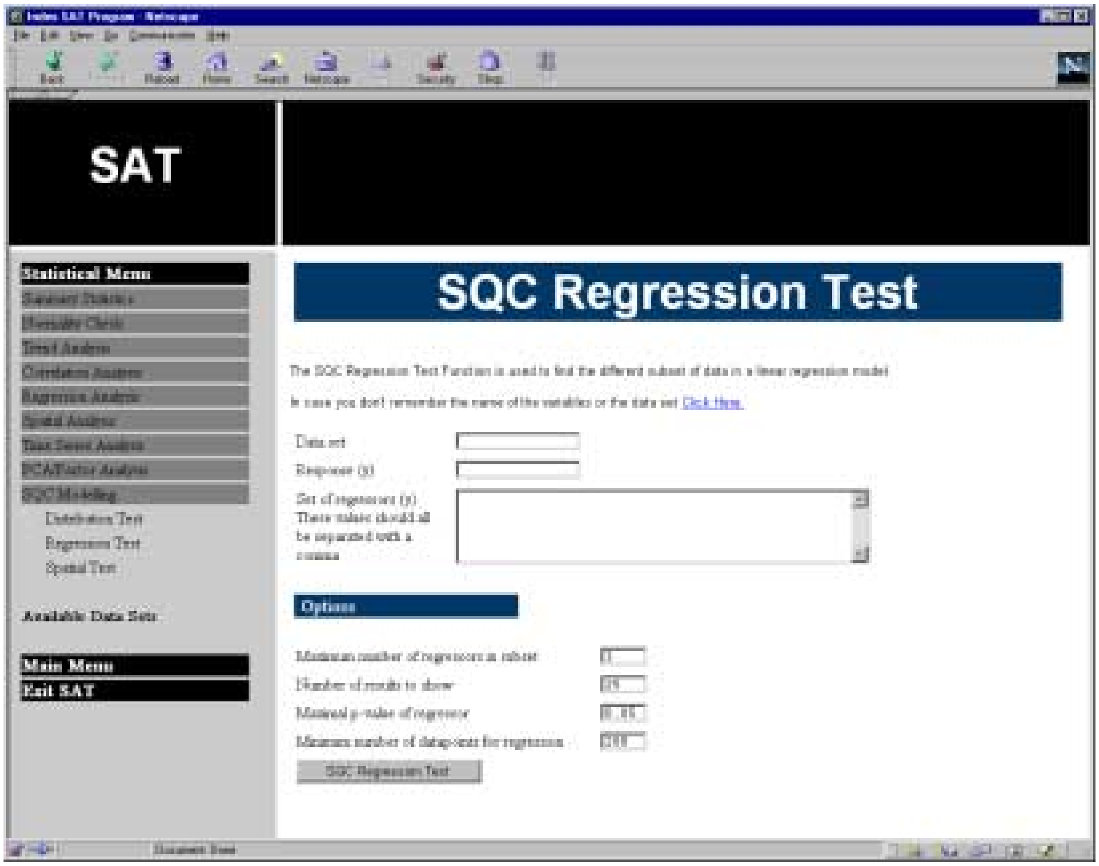
The SQC Regression Test function is used to derive a statistical model for the regression test in the SQC program. This test is used to test the measurement of a variable against the values of other variables measured at the same location and time.

The SQC Regression Test calculates different models from a given subset of regressor variables. These results are ordered based upon their goodness-of-fit.

The user should provide the name of the data set, the name of the response variable and a list of regressor values which should be included to look for the different models. This list must be structured in a specific manner. The different variables should be given with quotes and separated by a comma. The total set should be surrounded by "c(…)". An example is: c("AMON","PSALD","NTRA").

The user can define some options. A first option is the maximum number of regressors in the subset. If for example this option is set to 3, only subsets of 3 or less variables will be retained in the final list of results.
The second option is the maximum number of results which should be saved. Next to this the user can define the maximum p-value for a variable in a model and the minimum number of data points that are necessary to define the model.

Finally the user (normally the database administrator) should provide the name of the output file. The model is saved under this name and can by means of the Splus2Buffer program be transferred to the Statresults database.



The SQC Regression Test Function is used to find the different subset of data in a linear regression model

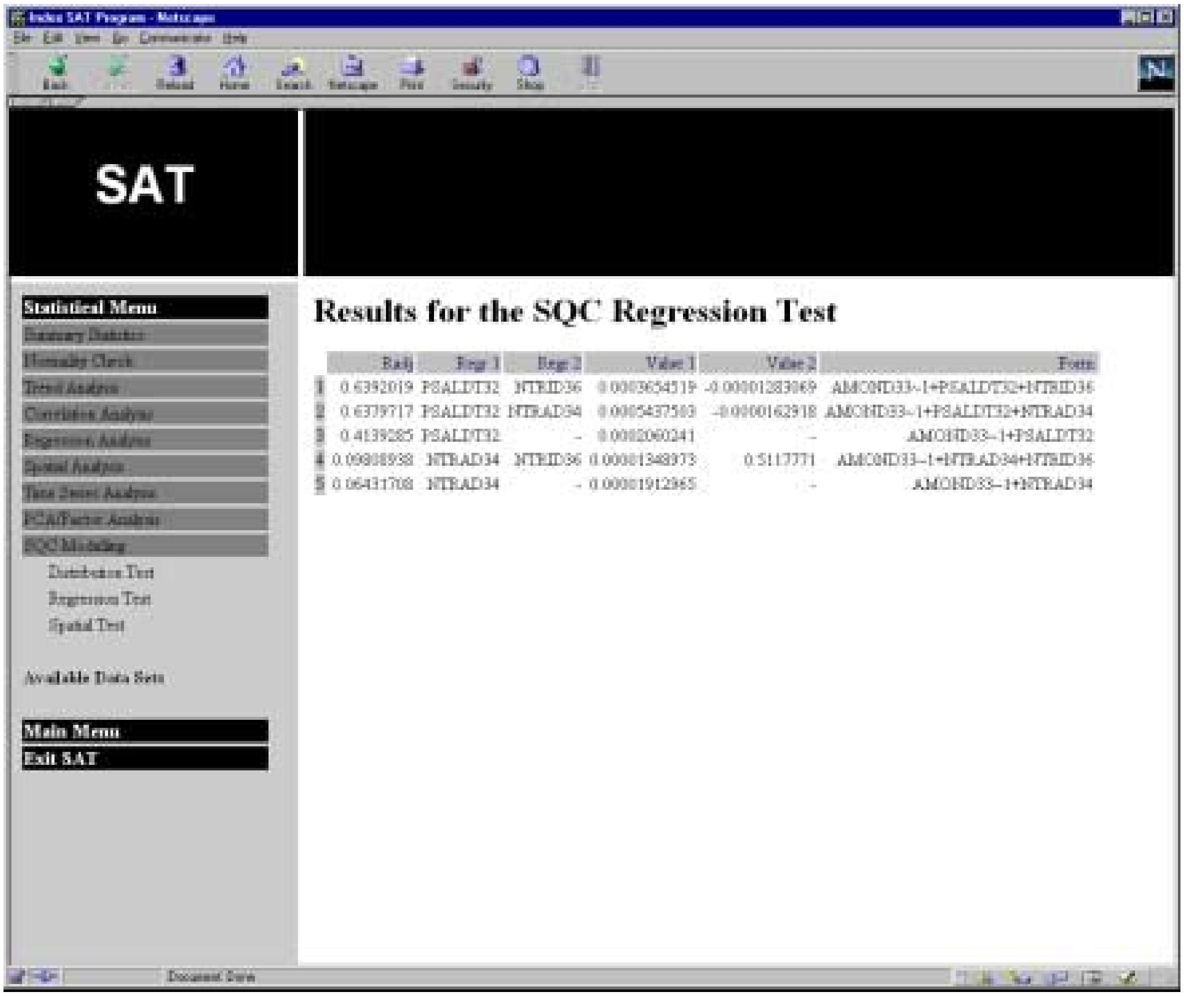
In case you don't remember the name of the variables or the data set Click Here.

Data set
Response (y)
Set of regressors (x)
These values should all be separated with a comma

Maximum number of regressors in subset
Number of results to show
Maximal p-value of regressor
Minimum number of datapoints for regression

Figure 34: The SQC Regression Test

A result of this function is shown in the following figure. A list, ordered by descending Radjusted values is given with the different regressors and the coefficients in the model.



Results for the SQC Regression Test

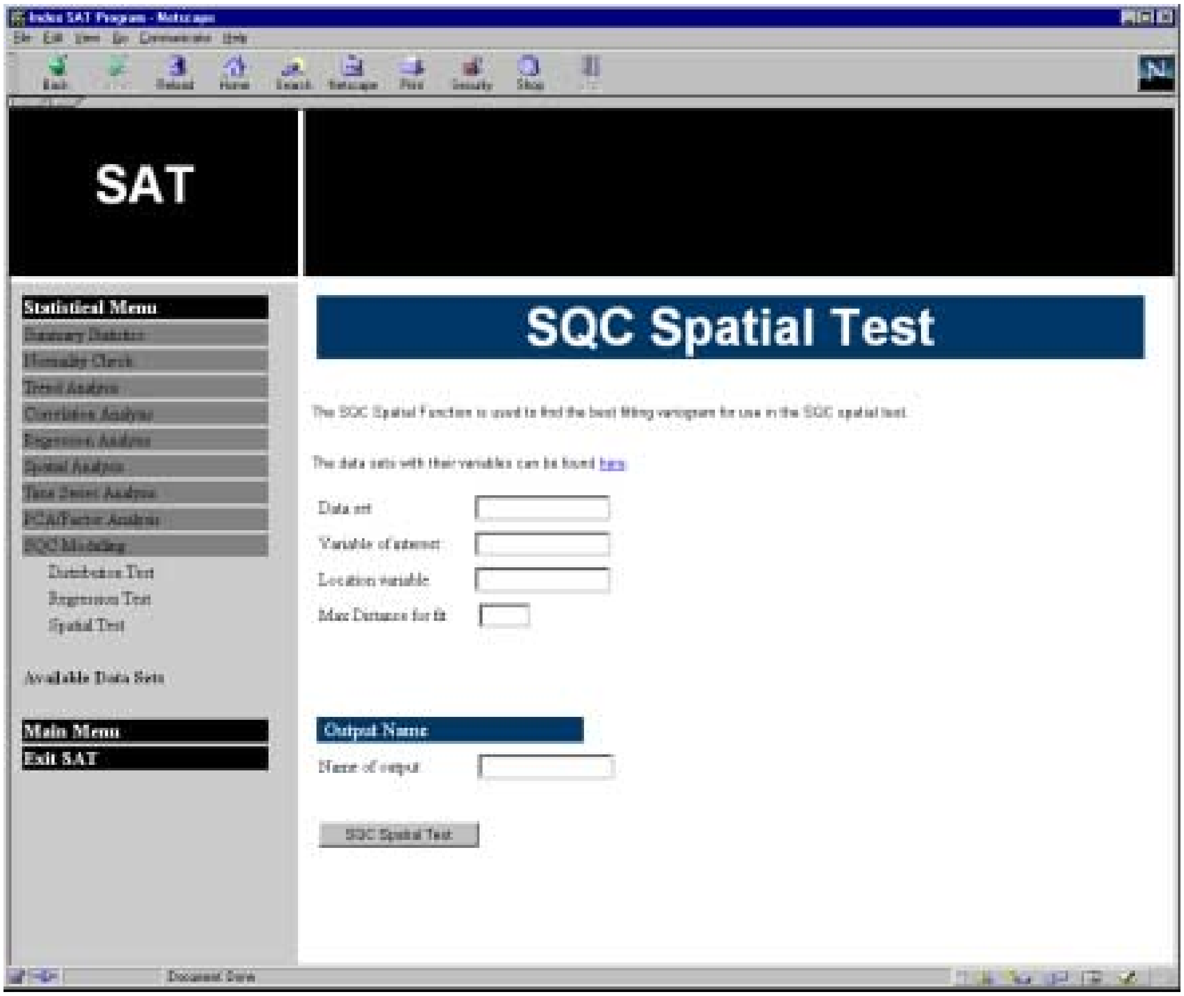
	Radj	Regr 1	Regr 2	Value 1	Value 2	Form
1	0.6392019	PSALDT32	NTRID36	0.00003654519	-0.00001283069	AMCOND33~1+PSALDT32+NTRID36
2	0.6379717	PSALDT32	NTRAD34	0.00005437503	-0.0000162918	AMCOND33~1+PSALDT32+NTRAD34
3	0.4139285	PSALDT32	-	0.00002060241	-	AMCOND33~1+PSALDT32
4	0.09808938	NTRAD34	NTRID36	0.00001348973	0.5117771	AMCOND33~1+NTRAD34+NTRID36
5	0.06431708	NTRAD34	-	0.00001912965	-	AMCOND33~1+NTRAD34

Figure 35: Example of the SQC Regression test

Spatial Test

The SQC Spatial Test is used to derive a model for the spatial test in the SQC program. The user should give the name of the data set, the name of the variable of interest, the location variable and the maximum distance for the fit of a power model $f(x) = a + bx^c$.

The name of the result should also be provided so the result can be transferred later by means of the Splus2Buffer program to the Statresults database.



SQC Spatial Test

The SQC Spatial Function is used to find the best fitting variogram for use in the SQC spatial test.

The data sets with their variables can be found here

Data set

Variable of interest

Location variable

Max Distance for fit

Output Name

Name of output

SQC Spatial Test

Figure 36: The SQC Spatial Test

An example of an output is given below.

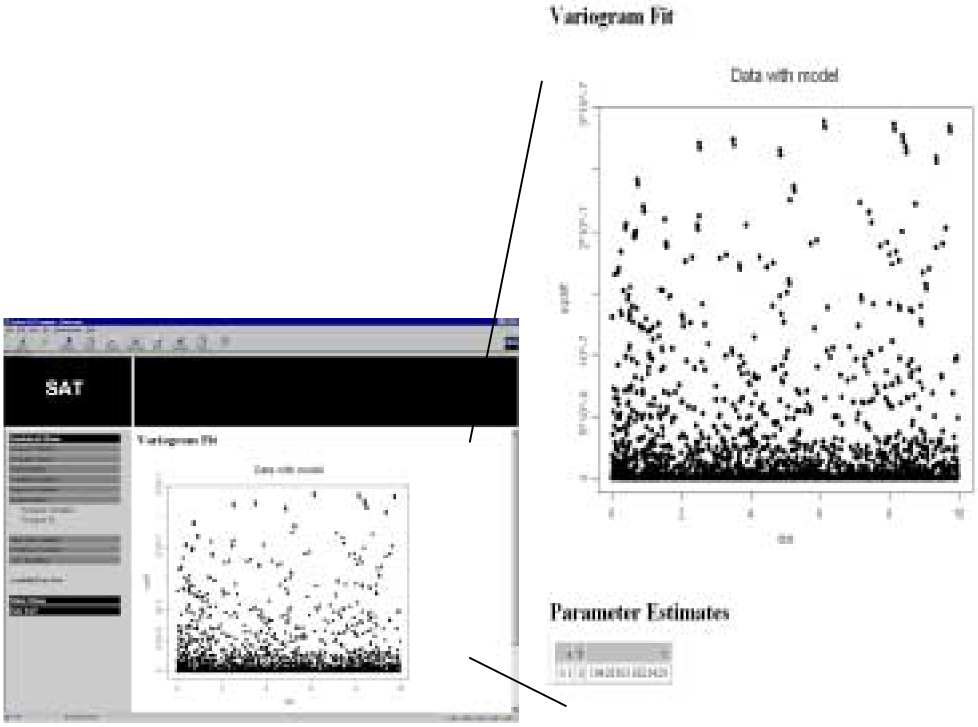


Figure 37: Example of the output of the SQC Spatial Test

Appendix C: Statistical Quality Control Program

Manual

UCS

Universitair Centrum voor Statistiek

Statistical Quality Control Program

Manual

Version 1: Draft

01/11/2000

Table of Contents

Introduction

The statistical quality control program SQC can be used to attach to the different measurements stored in the IDOD database a quality label. This manual describes how to operate the program.

MAIN MENU

Execution of the SQC program is controlled from the main menu (see figure 1) and involves at a minimum following 2 steps: "Load Data & Schemes" and "Perform Quality Control"

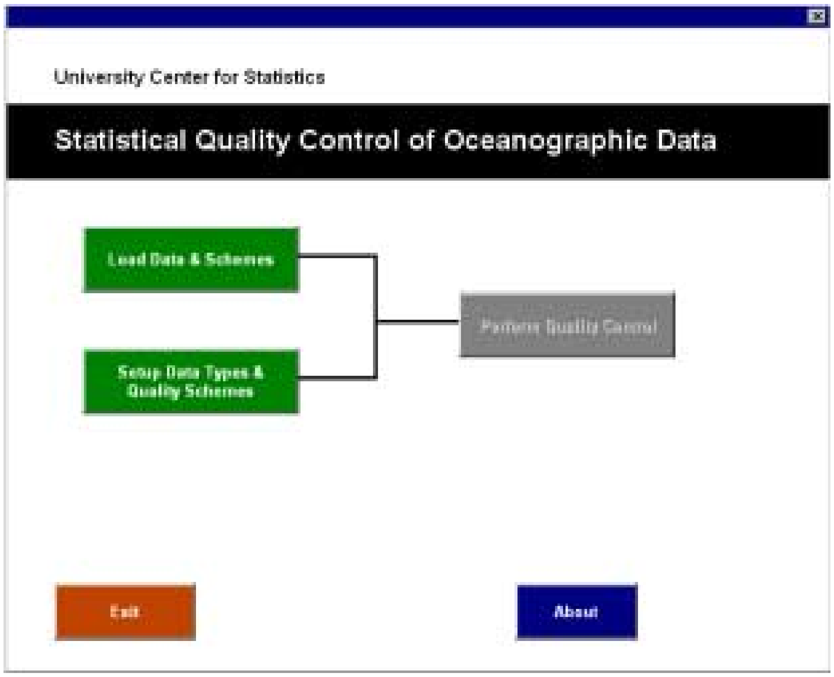

University Center for Statistics
Statistical Quality Control of Oceanographic Data
Load Data & Schemes
Setup Data Types & Quality Schemes
Perform Quality Control
Exit
About

Figure 38: Main Menu (at startup)

Load Data & Schemes

"Load Data & Schemes" prepares the SQC program for the subsequent quality control and involves following 4 steps:

1. import of the data records,
2. indication of the associated data type (a collection of measurements made at the same time and location) and the subset of measurements present in the records,
3. indication of the order of the different measurement types in the imported data records,
4. choice of an appropriate quality control scheme.

Perform Quality Control

"Perform Quality Control" applies the quality control scheme to the imported dataset. (This function is only available after the load data & schemes step is executed.)

Once the quality control has been executed, following 3 additional functionalities are available at the main menu level (see figure 2):

Results
"Results" summarizes on an Excel sheet the results of the quality control for the imported data;

Backtracking
"Backtracking" shows on an Excel sheet the results of the individual tests for measurements that are labeled as "BAD";

Export/Apply
"Export/Apply" exports the quality information to the IDOD database where it is permanently stored.

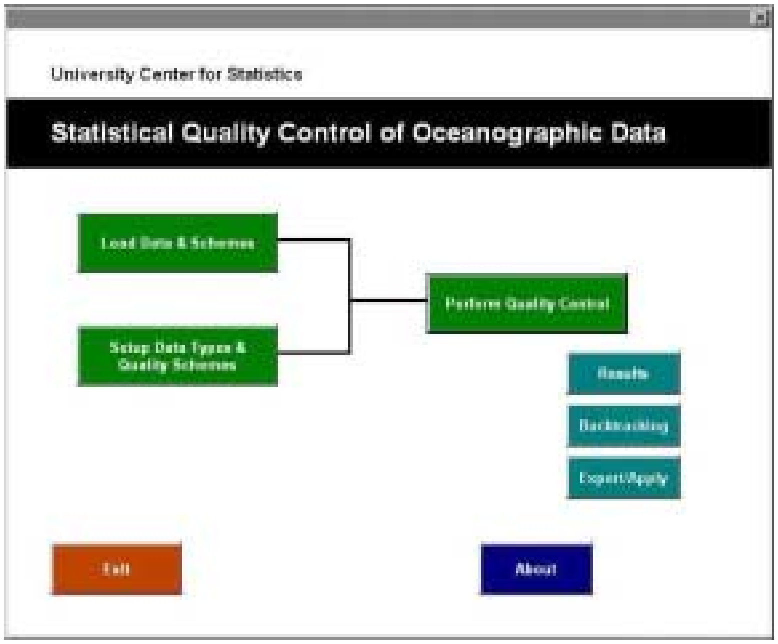

University Center for Statistics
Statistical Quality Control of Oceanographic Data
Load Data & Schemes
Setup Data Types & Quality Schemes
Perform Quality Control
Results
Backtracking
Export/Apply
Exit
About

Figure 39: Main Menu (After Quality Control)

Setup Data Types and Schemes

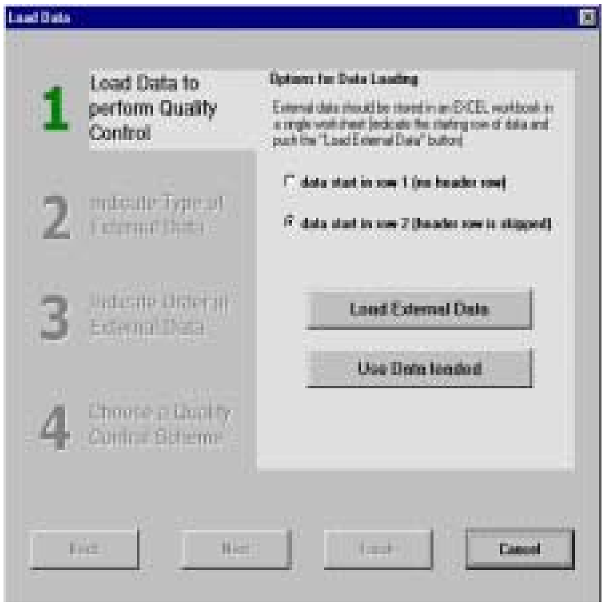
The main menu further includes a password-protected functionality "Setup Data Types and Schemes" that allows the data manager to either define new measurement and data types or define new quality control tests and schemes that later on can be used in the program.

Buttons Exit and About allow respectively to exit the program and to check the version number of the program.

LOAD DATA AND SCHEMES

This button loads a wizard that guides the user through the following 4 steps:

Step 1: Load Data to perform Quality Control (figure 3)



Load Data

1 Load Data to perform Quality Control

2 Indicate Type of External Data

3 Indicate Order of External Data

4 Choose a Quality Control Scheme

Options for Data Loading

External data should be stored in an EXCEL workbook in a single work sheet. Indicate the starting row of data and push the "Load External Data" button

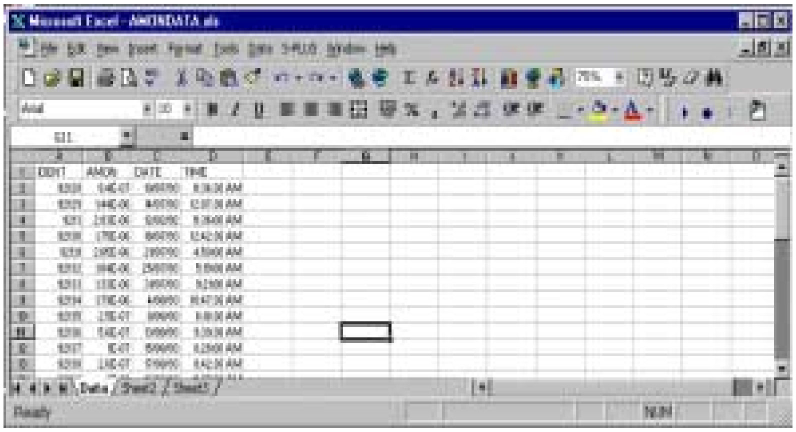
○ data start in row 1 (no header row)

● data start in row 2 (header row is skipped)

Figure 40: Load data: Step 1

This form includes two buttons: "**Use Data loaded**" and "**Load External Data**".

> **"Use Data loaded"** refers to the case where the wizard has been previously activated but a different quality control scheme should be applied. Pushing this button will bring the user directly to Step 4 below.
>
> **"Load External Data"** should be used if a new dataset is to be quality controlled. Prior to pushing this button, the user should indicate whether the data records start on the first or second row. This option is included to remove a header line if it is present. The "Load External Data" button activates the open-file form that is standard to EXCEL. The user should indicate the EXCEL workbook in which the data reside. Within the workbook, the data are assumed to be stored on a sheet by the name of "Data".



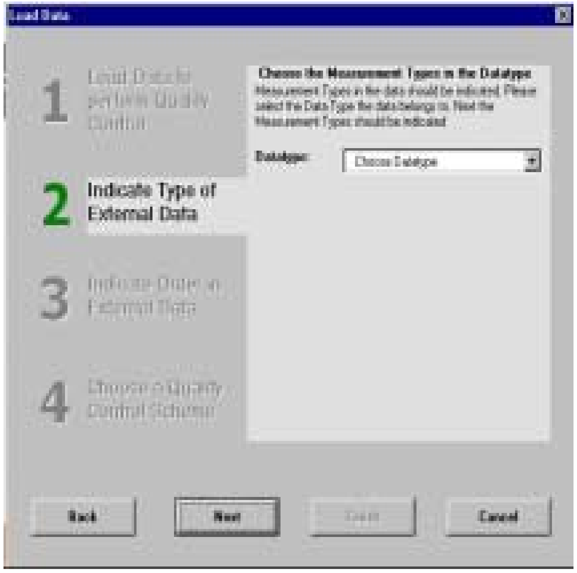
Microsoft Excel - AMONDATA.xls

IDENT	AMON	DATE	TIME
[illegible]	[illegible]	[illegible]	[illegible] AM
[illegible]	[illegible]	[illegible]	[illegible] AM
[illegible]	[illegible]	[illegible]	[illegible] AM
[illegible]	[illegible]	[illegible]	[illegible] AM
[illegible]	[illegible]	[illegible]	[illegible] AM
[illegible]	[illegible]	[illegible]	[illegible] AM
[illegible]	[illegible]	[illegible]	[illegible] AM
[illegible]	[illegible]	[illegible]	[illegible] AM
[illegible]	[illegible]	[illegible]	[illegible] AM
[illegible]	[illegible]	[illegible]	[illegible] AM
[illegible]	[illegible]	[illegible]	[illegible] AM
[illegible]	[illegible]	[illegible]	[illegible] AM

Figure 41: Data organization in external data

Figure 4 shows an example of the data organization that is assumed by SQC (in the example shown a header row is present). The different measurement types should be ordered in consecutive columns; different rows indicate different instances of measurement. SQC assumes that the last row in the worksheet corresponds to the last measured instance. The worksheet should therefore not contain any other information than the header row (optional) and the measurement values. Empty cells, "NA" or any non-numeric value are interpreted as missing values. Once the data have been activated (either existing or new), the second step will appear.

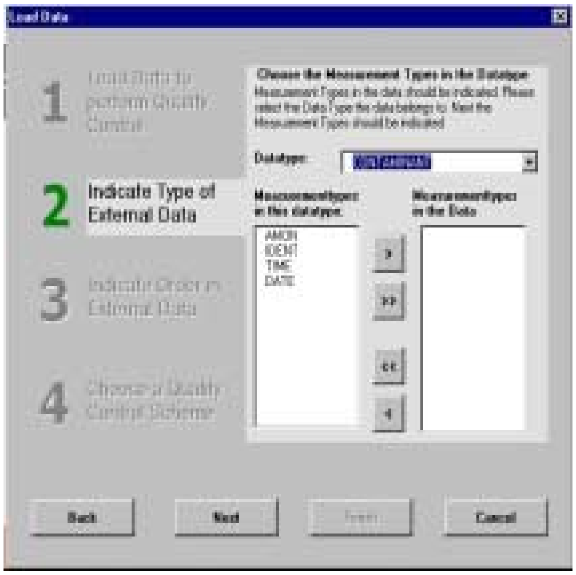
Step 2: Indicate type of external data (fig. 5a and 5b)


Load Data
1 Load Data to perform Quality Control
2 Indicate Type of External Data
3 Indicate Order in External Data
4 Choose a Quality Control Scheme
Back
Next
Finish
Cancel

Choose the Measurement Types in the Datatype
Measurement Types in the data should be indicated. Please select the Data Type the data belongs to. Next the Measurement Types should be indicated.

Datatype: Choose Datatype

Figure 42a: Load Data: Step 2


Load Data
1 Load Data to perform Quality Control
2 Indicate Type of External Data
3 Indicate Order in External Data
4 Choose a Quality Control Scheme
>
>>
<<
<
Back
Next
Finish
Cancel

Choose the Measurement Types in the Datatype
Measurement Types in the data should be indicated. Please select the Data Type the data belongs to. Next the Measurement Types should be indicated.

Datatype: [illegible]

Measurementtypes in this datatype: AMON, IDENT, TIME, DATE

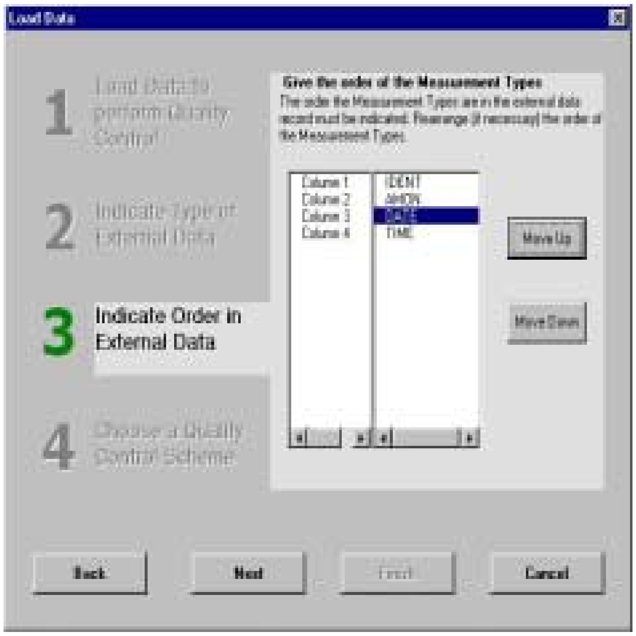
Measurementtypes in the Data:

Figure 5b: Load Data: Step 2

In this step the ‘global’ datatype to which the imported dataset belongs must be indicated (figure 5a). This datatype should include all of the measurement types in the dataset but can be more general.
After the choice of the datatype, a listbox appears with the available measurement types for the chosen datatype (figure 5b). The measurement types that are actually present in the imported data records should then be selected from the available measurement types (button ">>" will copy all available measurement types, button ">" will copy the highlighted available measurement type, button "<<" will remove all selected measurement types and button "<" will remove the highlighted selected measurement type).

Step 3: Indicate order in external data (figure 6)

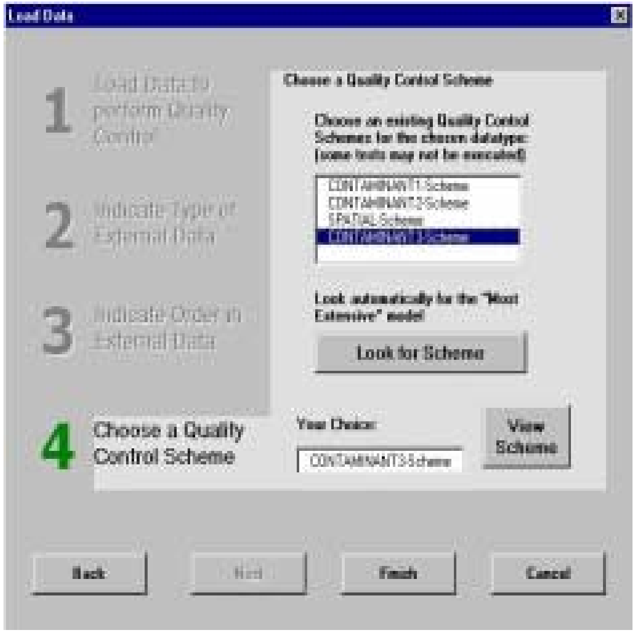
In this step the order in which the different measurement types previously selected appear in the data records must be indicated. "**Move up**" will move the selected measurement type in the ordered list, while "**Move down**" will lower the order of the selected measurement type. For ease of reference, the column number that is associated with each measurement type is indicated in a left window.


Load Data
1 Load Data to perform Quality Control
2 Indicate Type of External Data
3 Indicate Order in External Data
4 Choose a Quality Control Scheme
Move Up
Move Down
Back
Next
Finish
Cancel

Give the order of the Measurement Types

The order the Measurement Types are in the external data record must be indicated. Rearrange (if necessary) the order of the Measurement Types.

Column 1	IDENT
Column 2	AMON
Column 3	DATE
Column 4	TIME


Load Data
1 Load Data to perform Quality Control
2 Indicate Type of External Data
3 Indicate Order in External Data
4 Choose a Quality Control Scheme
Look for Scheme
View Scheme
Back
Next
Finish
Cancel

Choose a Quality Control Scheme

Choose an existing Quality Control Schemes for the chosen datatype: (some tests may not be executed)

CONTAMINANT1-Scheme
CONTAMINANT2-Scheme
SPATIAL-Scheme
CONTAMINANT3-Scheme

Look automatically for the "Most Extensive" model

Your Choice: CONTAMINANT3-Scheme

Figure 43: Load data: Step 3

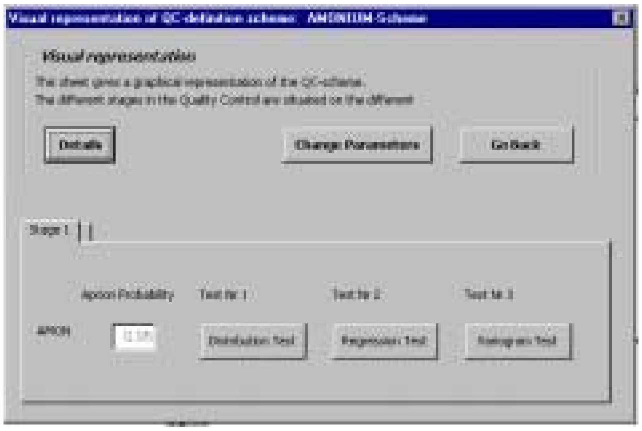
Figure 44: Load data: Step 4

Step 4: Choose a quality control scheme (figure 7)

In this last step the quality control scheme that is to be applied to the data must be chosen.
The possible choices include all quality control schemes that have been previously defined by the data manager for the selected datatype. Possibly some of these quality control schemes include tests that cannot be executed because the test requires a measurement type that is not present in the imported data. This does not cause problems in the application of the test scheme (the test will be simply skipped) but if many such tests are present, it will increase the execution time of the quality procedure considerably.

Alternatively, the button "**Most extensive quality control scheme possible**" can be used to remedy this problem. In this case, a (temporary) quality control scheme is custom build for the imported data. This scheme comprises all tests that can be applied to the different measurement types that are available in the dataset.

Once a scheme has been selected, its definition can be viewed by clicking the **"View Scheme"** button (see figure 8 for an example). The view shows, for each measurement type, the a-priori probability that the measurement is "good" and the sequence of tests that are applied to the different measurement types.



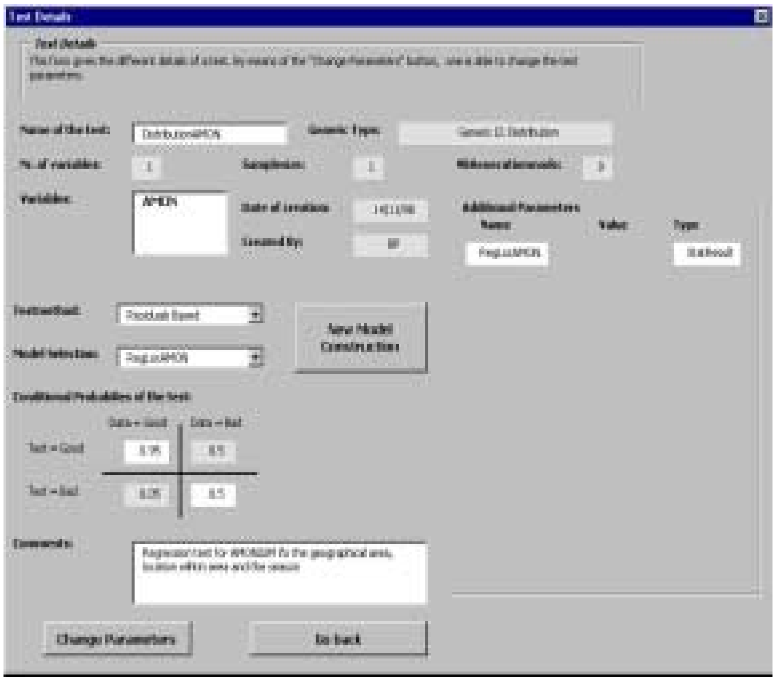
Visual representation of QC-definition scheme: AMONIUM-Scheme

Visual representation

The sheet gives a graphical representation of the QC-scheme.
The different stages in the Quality Control are situated on the different

Figure 45: Visual representation of QC-scheme

In a multi-stage quality control procedure the tests are grouped on separate pages of a multipage control (clicking the page tag shows the tests associated with that stage). The tests themselves are represented by buttons. Clicking a test will show the details of the test (see figure 9 for an example of the distribution test).



Test Details

Test details

This form gives the different details of a test. By means of the "Change Parameters" button, one is able to change the test parameters.

Name of the test: DistributionAMON
Generic Type: Generic 1D Distribution
Nr. of variables: 1
Sample size: 1
[illegible]: 3
Variables: AMON
Date of creation: 14/11/98
Created by: BF

Additional Parameters

Name	Value	Type
RegLinAMON		[illegible]

Test method: [illegible]
Model selection: RegLinAMON

Conditional Probabilities of the test:

	Data = Good	Data = Bad
Test = Good	0.95	0.5
Test = Bad	0.05	0.5

Comments: Regression test for AMONIUM fix the geographical area, location within area and the season

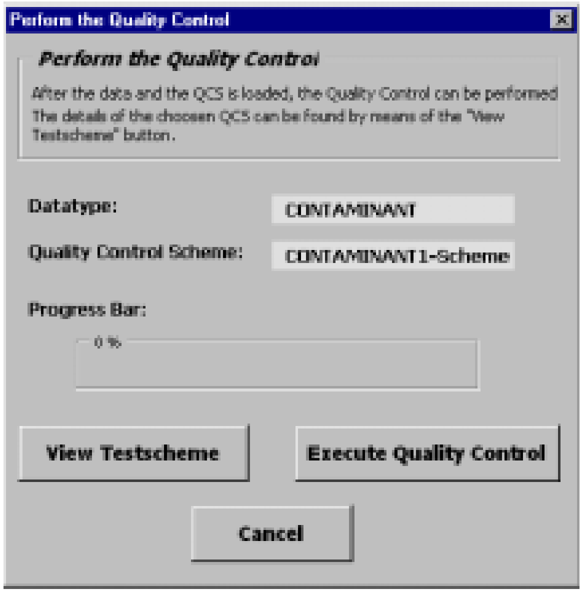
Figure 46: Example of test details

The "**Finish**" button will load the chosen quality control scheme and return the user to the main menu, from which the quality control can be applied by pushing the “**Perform Quality Control**” button (see figure 1).

Perform Quality Control

The "**Perform Quality Control**" button on the main menu is only accessible if the previous initialization process has been successfully finished.

Pushing the button brings forward a form (see figure 10) that shows the selected datatype and quality control scheme and that includes following 3 buttons:



Perform the Quality Control

Perform the Quality Control

After the data and the QCS is loaded, the Quality Control can be performed The details of the choosen QCS can be found by means of the "View Testscheme" button.

Datatype: CONTAMINANT

Quality Control Scheme: CONTAMINANT1-Scheme

Progress Bar: 0 %

Figure 47: Perform the Quality Control

"View Testscheme"

This button allows viewing the definition of the quality control scheme (in the same manner as described earlier in this manual).

"Execute Quality Control"

This button will apply the quality control scheme to the imported data. The execution of the scheme is timed by a progress bar. Once the quality control is finished, the user is returned to the main menu where 3 additional buttons will appear ("Results", "Backtracking" and "Export/Apply", see above in figure 2).

"Cancel"

The button can be used to return to the main menu without execution of the quality control scheme.

Setup Data Types & Quality Schemes

This is a password-protected functionality, which allows the data manager to define new measurement types, data types (sets of measurement types), new tests and quality control schemes (sets of tests).

Figure 11 shows a flowchart of the forms that the data manager may activate. The different forms are discussed next.

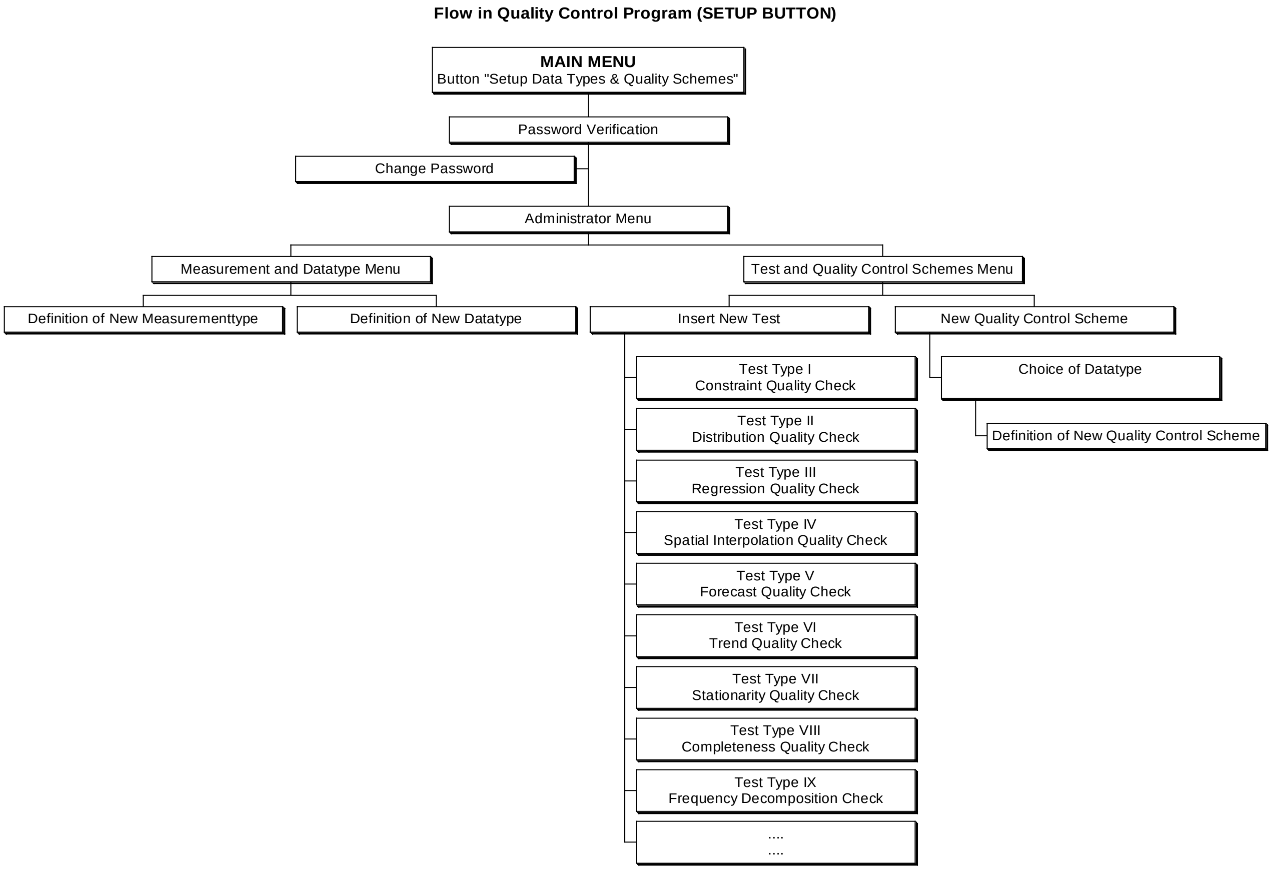
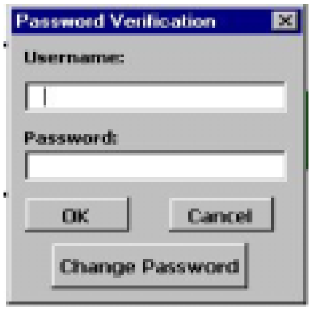
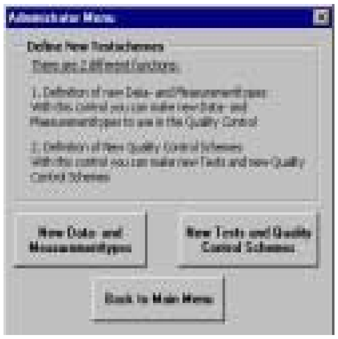

Flow in Quality Control Program (SETUP BUTTON)
MAIN MENU
Button "Setup Data Types & Quality Schemes"
Password Verification
Change Password
Administrator Menu
Measurement and Datatype Menu
Test and Quality Control Schemes Menu
Definition of New Measurementtype
Definition of New Datatype
Insert New Test
New Quality Control Scheme
Test Type I
Constraint Quality Check
Test Type II
Distribution Quality Check
Test Type III
Regression Quality Check
Test Type IV
Spatial Interpolation Quality Check
Test Type V
Forecast Quality Check
Test Type VI
Trend Quality Check
Test Type VII
Stationarity Quality Check
Test Type VIII
Completeness Quality Check
Test Type IX
Frequency Decomposition Check
....
....
Choice of Datatype
Definition of New Quality Control Scheme

Figure 48: Flow chart of userforms activated by "Setup Data Types & Quality Schemes"

In the first step, the user is prompted to give a username and a password (figure 12). Also an option to change the password is available.




Administrator Menu
Define New Testschemes
New Data- and Measurementtypes
New Tests and Quality Control Schemes
Back to Main Menu

There are 2 different functions:

1. Definition of new Data- and Measurementtypes
With this control you can make new Data- and Measurementtypes to use in the Quality Control

2. Definition of New Quality Control Schemes
With this control you can make new Tests and new Quality Control Schemes

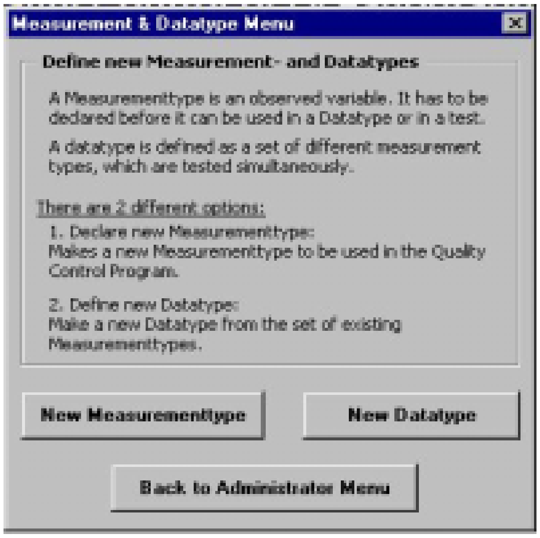
Figure 49: Password verification

Figure 50: Administrator Menu

Next the user enters into the Administrator Menu (figure 13), which presents a choice between either the definition of " **New data- and measurementtype** " or the definition of "**New Tests and Quality Control Schemes**".

Button 1: " New data- and measurementtype "
This button loads a form "Measurement and Data Type Menu" (figure 14). Here the choice can be made between "**New Measurement Type**" or "**New Data Type**".

"**New Measurement Type**"
This button should be used to define a new measurement type. The "Definition of new Measurement type" form is loaded (see figure 15).



Measurement & Datatype Menu

Define new Measurement- and Datatypes

A Measurementtype is an observed variable. It has to be declared before it can be used in a Datatype or in a test.

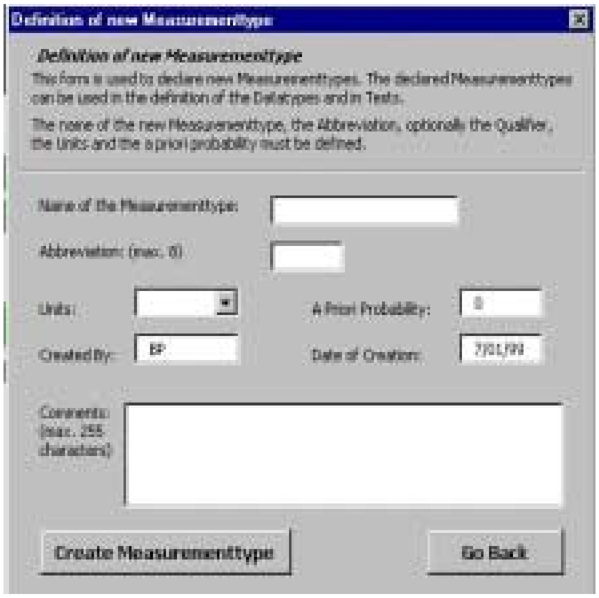
A datatype is defined as a set of different measurement types, which are tested simultaneously.

There are 2 different options:

1. Declare new Measurementtype:
Makes a new Measurementtype to be used in the Quality Control Program.

2. Define new Datatype:
Make a new Datatype from the set of existing Measurementtypes.

Figure 51: Measurement & Datatype Menu



Definition of new Measurementtype

Definition of new Measurementtype

This form is used to declare new Measurementtypes. The declared Measurementtypes can be used in the definition of the Datatypes and in Tests.

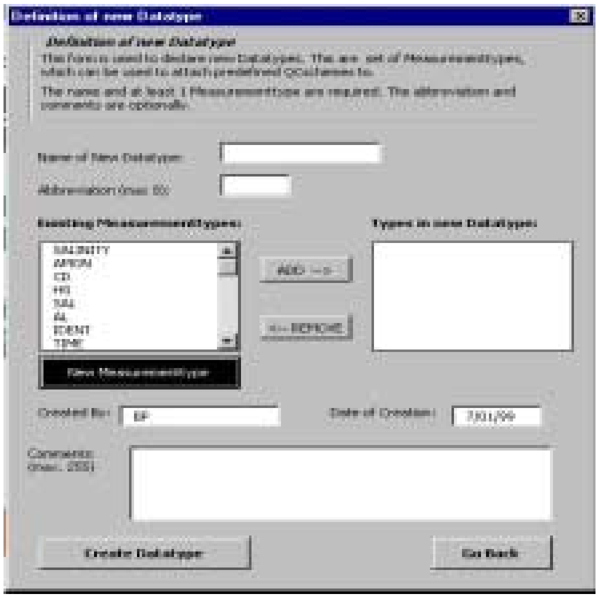
The name of the new Measurementtype, the Abbreviation, optionally the Qualifier, the Units and the a priori probability must be defined.

Name of the Measurementtype:			
Abbreviation: (max. 8)			
Units:		A Priori Probability:	0
Created By:	BP	Date of Creation:	7/01/99
Comments: (max. 255 characters)			

Figure 52: Definition of new Measurement type

The user is requested to assign a name and (optionally) an abbreviation to the measurement type. Comments can be also (optionally) entered. The name of the measurement type must be unique. If not, the user will be prompted to enter another name.
The units that are associated with the measurement type must be chosen from a list of available units. Also the a-priori probability, a value between 0 and 1 that indicates the probability that measurements of this type are “good”, must be indicated. The fields “created by” and “date of creation” are generated automatically, but may be changed if deemed necessary by the user.
The new measurement type is effectively loaded into the database by pushing the "**Create Measurement Type**" button. Afterwards the user can continue to declare another measurement type.

"**New Data Type**"
This button should be used to define a new data type. The form "Definition of New Datatype" is loaded (see figure 16).



Definition of new Datatype

Definition of new Datatype

This Form is used to declare new Datatypes. The are set of Measurementtypes, which can be used to attach predefined QCschemes to.

The name and at least 1 Measurementtype are required. The abbreviation and comments are optionally.

Name of New Datatype:

Abbreviation (max 8):

Existing Measurementtypes:

Types in new Datatype:

SALINITY
AMMON
CD
HG
SAL
AL
IDENT
TIME

Created By: EP

Date of Creation: 7/01/99

Comments (max. 255)

Figure 53: Definition of new datatype

As for a new measurement type, the name and (optionally) an abbreviation must be assigned. A comment field to detail the data type is also available. The name of the data type must be unique. If not, the user will be prompted to enter another name.

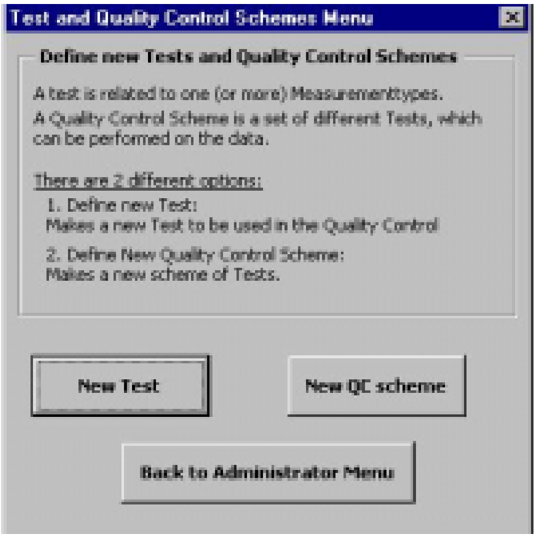
The measurement types that are part of the datatype must be selected from the list of available measurement types. The button "**New Measurementtype**" can be used to activate at this point the "New Measurement Type" form and declare a new measurement type. Once the declaration is finished, the user is returned to the present form, where the new measurement type is added to the list of available measurement types.

The fields "created by" and "date of creation" are generated automatically, but can be changed by the user if necessary.

The new data type is effectively loaded into the database by pushing the "**Create Data Type**" button. Afterwards the user can continue to declare another Data type

Button 2: "New Tests and Quality Control Schemes"
This button loads the "Test and Quality Control Schemes Menu" (see figure 17). In this case a choice must be made between either the declaration of a new test ("**New Test**") or the declaration of a new quality control scheme ("**New Quality Control Scheme**").

"**New Test**"
This button should be used to declare a new test.
A form "Insert New test" (figure 18) is loaded where the user must select the generic type of test that should be created. (see the flowchart for a list of the generic types available).

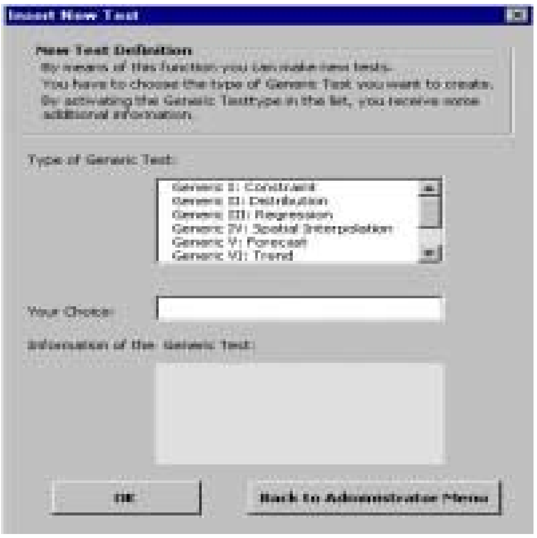


Test and Quality Control Schemes Menu

Define new Tests and Quality Control Schemes

A test is related to one (or more) Measurementtypes.
A Quality Control Scheme is a set of different Tests, which can be performed on the data.

There are 2 different options:
1. Define new Test:
Makes a new Test to be used in the Quality Control
2. Define New Quality Control Scheme:
Makes a new scheme of Tests.



Insert New Test

New Test Definition
By means of this function you can make new tests.
You have to choose the type of Generic Test you want to create.
By activating the Generic Testtype in the list, you receive some additional information.

Type of Generic Test:
Generic I: Constraint
Generic II: Distribution
Generic III: Regression
Generic IV: Spatial Interpolation
Generic V: Forecast
Generic VI: Trend

Your Choice:

Information of the Generic Test:

Figure 54: Test and Quality Schemes Menu

Figure 55: Insert new test

Pushing the **"OK"** button then activates an input form part of which is the same independent of the generic type, part of which is specific to the generic type.
In general, the user must:

- choose the "measurement type" to which the test is to be applied from a list of available measurement types;
- if necessary, modify the "date of creation" and "'created by", two fields that are automatically generated;
- optionally, modify the "name of the test" that is automatically assigned to the test by the program;
- optionally, fill in the "comments" field to clarify in words what is being tested;
- fill in the "conditional probabilities of the test". Specifically, the probability that the test passes for a "good" datapoint, denoted as alfa, and the probability that the test fails for a "bad" datapoint, denoted as beta, must be indicated. While the value of alfa is intrinsically defined by the significance level used in the test and thus is typically well known, the value of beta may be in many instances difficult to determine. This is so, because "erroneous" data may be produced by different causes (reversal of digits, misplacing the decimal point, instrument errors, digital transmission errors, calibration errors, ...) each of which produce a particular error pattern to which the test may be more or less sensitive. We therefore recommend that a guess-estimate is assigned based on

following principles: 1. make a guess-estimate of the frequency of occurrence of different types of errors f_l, 2. make a guess-estimate of the probability that the test would detect each type of error p_i, 3. set the value of beta equal to the sum of the products f_l*p_l. Note that the value of beta should never be less than 1-alfa. This would indicate that the test is not of any use in identifying "bad" measurements. As an example: consider the case where it is known that a measurement should be positive and hence a test is applied that checks whether this is indeed the case. Alfa is in this case equal to 1, since good datapoints will always pass the test. Bad datapoints will however only be identified if they are negative and thus beta should correspond to the frequency with which such errors are expected among "bad" datapoints. If other errors can occur and are more frequent, the value of beta may be relatively low (i.e. 0.05). Consider on the other hand a regression test where the test fails if the residual has a significance value less than 5%. In such a case, alfa corresponds to 0.95. To determine the value of beta, one must consider how many of the "erroneous" measurements would be characterized by a sufficiently large deviation to fail in such a test. If the value is less than 5% it would imply that "bad" datapoints are actually characterized by falling inside the regression limits more frequently than good datapoints (in fact, such a case can occur in practice, for instance when the error consists of replacing missing values with regressed values that do not represent actual measurements) and, using a test in such a case would decrease the a-posteriori probability of a datapoint being "bad", if it is really "bad".
More usual, erroneous data will be typically characterized by larger deviations in a regression plot than good data, so that the proportion of data that fails the test should be larger than 1-alfa. How many of these points do so, may be difficult to determine (without in-depth analysis of the types of errors and their characteristics). In such a case, we suggest that a simple guess-estimate is used, using large values of beta (i.e. 0.95) if deviations are expected to be large for erroneous data and using more moderate values of beta (i.e. 0.5 or even less) if the deviations are expected to be less pronounced or may not show up in such a regression plot (i.e. because also the other measurement may be affected by the same error);

- a series of other fields that depend on the type of generic test. These fields are further detailed in the following section on "Generic Tests and Parameters".

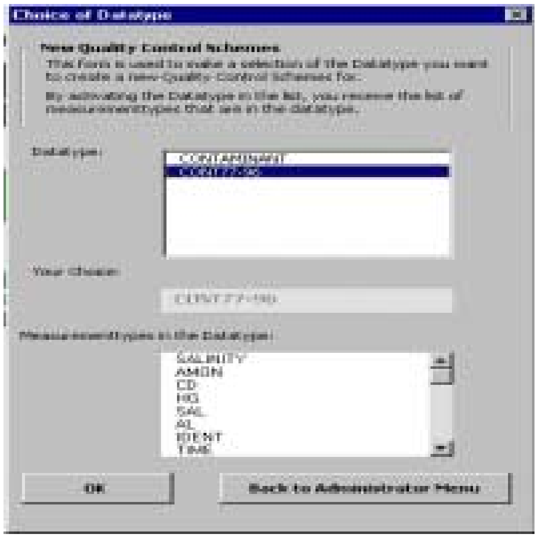
Independent of the type of test, the form to define the test includes following four buttons:

"**Make Test**" generates the test;

"**Another Type x test**" allows to generate another type of the same type x;

"**Go Back**" and "**Test of Another Generic Type**" bring the user back to the form where the generic type of test is selected.

"**New Quality Control Scheme**"
This button loads the "Choice of Data type" form (fig. 19). In this form the user must indicate the data type for which a quality control scheme will be defined. Once the data type is selected, the associated measurement types are shown at the bottom of the form.



Choice of Datatype

New Quality Control Schemes

This form is used to make a selection of the Datatype you want to create a new Quality Control Schemes for.

By activating the Datatype in the list, you receive the list of measurementtypes that are in the datatype.

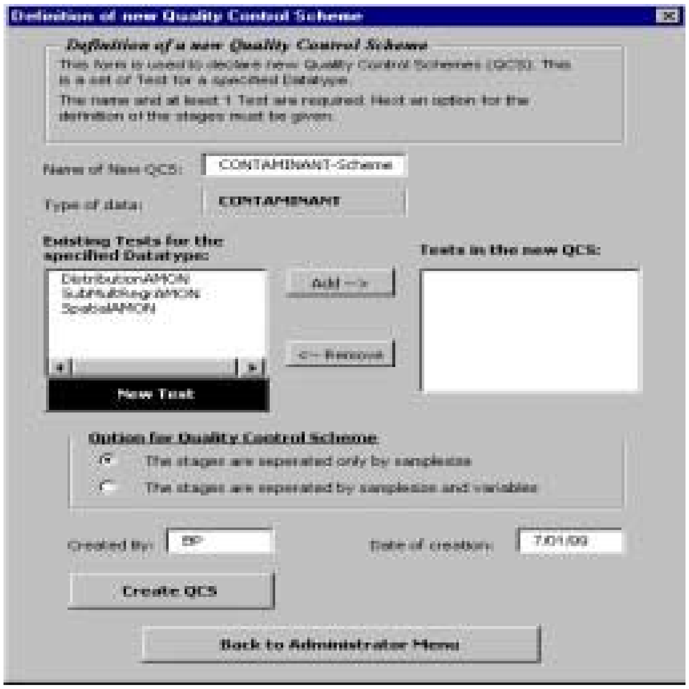
Datatype:
CONTAMINANT
CONT77-96

Your choice: CONT77-96

Measurementtypes in the Datatype:
SALINITY
AMON
CD
HG
SAL
AL
IDENT
TIME

Figure 56: Choice of Datatype

The "**OK**" button in this form brings the user to a form "Definition of new Quality Control Scheme" that allows to construct the quality control scheme for the data type (see figure 20).
In this form, the user must fill in the name of the scheme. Comments (to document the quality control scheme) can be (optionally) added. The name of the quality control scheme must be unique. If not, the user will be prompted for another name.



Definition of new Quality Control Scheme

Definition of a new Quality Control Scheme

This form is used to declare new Quality Control Schemes (QCS). This is a set of Test for a specified Datatype.

The name and at least 1 Test are required. Next an option for the definition of the stages must be given.

Name of New QCS: CONTAMINANT-Scheme

Type of data: CONTAMINANT

Existing Tests for the specified Datatype:
DistributionAMON
SubMultiRegrAMON
SpatialAMON

Tests in the new QCS:

Option for Quality Control Scheme
● The stages are seperated only by samplesize
○ The stages are seperated by samplesize and variables

Created By: EP

Date of creation: 7/01/00

Figure 57: Definition of new Quality Control Scheme

To construct the scheme, the user can select the tests that should be included from a list of existing tests. The button "**New Test**" allows the user to declare at this point a new test, which will (after the declaration) be put into the list of available tests.

The user should also indicate the option he wants to use for organizing the different stages. If the option "Stages are separated only by samplesize" is indicated, all tests using the same samplesize are gathered in 1 stage, independent of which measurement type the test is applied to. If the option "Stages are separated by samplesize and variables" is selected, each stage groups the tests that apply to the same measurement type and that use the same samplesize.

As an example, suppose that 3 tests are used to construct the scheme. The first test applies to the measurement type AMON and uses sample size 1; the second test again applies to AMON but uses sample size 4; the third and final test applies to the measurement type CD and uses samplesize 1.
With the first option ("Stages are separated only by sample size") active, 2 stages will be made, the first one containing the tests on AMON and CD and using sample size 1, and the second stage containing the test on AMON and using sample size 4. With the alternative option ("Stages are separated by samplesize and variables") active, 3 stages are produced and each stage contains a single test.

The fields “created by” and “date of creation” are automatically generated by the program but can be changed by the user.

After the quality control scheme has been created, the user can use the "**View Scheme**" button (which appears after the test is created with the button “**Create QCS**” to graphically display the scheme and examine its characteristics as described earlier.

Generic Tests and Parameters

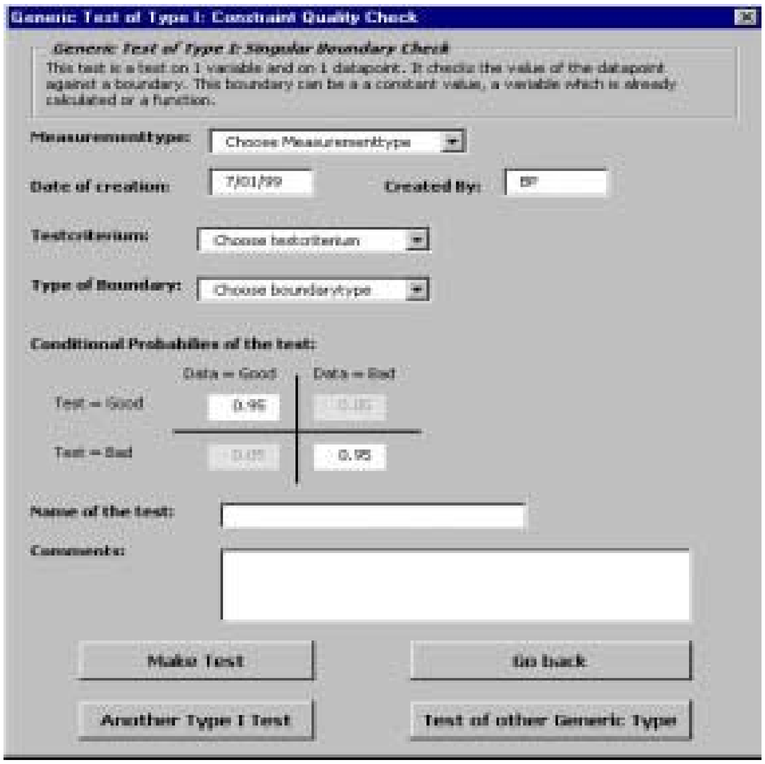
In the SQC program the following list of generic tests is available.

- Constraint QC
- Distribution QC
- Regression QC
- Spatial Interpolation QC
- Forecast QC
- Trend QC
- Stationarity QC
- Completeness QC
- Frequency Decomposition QC

As denoted above each generic test has a series of parameters that depends on the type of generic test. In this section details will be given for each type of generic test.

Generic Test Type I: Constraint Test (fig. 21)

This generic test can be used to test a single measurement type with respect to a given boundary.



Generic Test of Type I: Constraint Quality Check

Generic Test of Type I: Singular Boundary Check

This test is a test on 1 variable and on 1 datapoint. It checks the value of the datapoint against a boundary. This boundary can be a a constant value, a variable which is already calculated or a function.

Measurementtype: Choose Measurementtype

Date of creation: 7/01/99 Created By: BP

Testcriterium: Choose testcriterium

Type of Boundary: Choose boundarytype

Conditional Probabilities of the test:

	Data = Good	Data = Bad
Test = Good	0.95	0.05
Test = Bad	0.05	0.95

Name of the test:

Comments:

Figure 58: Generic Test of Type I

In addition to the general fields described in the previous section following 2 fields need to be declared:

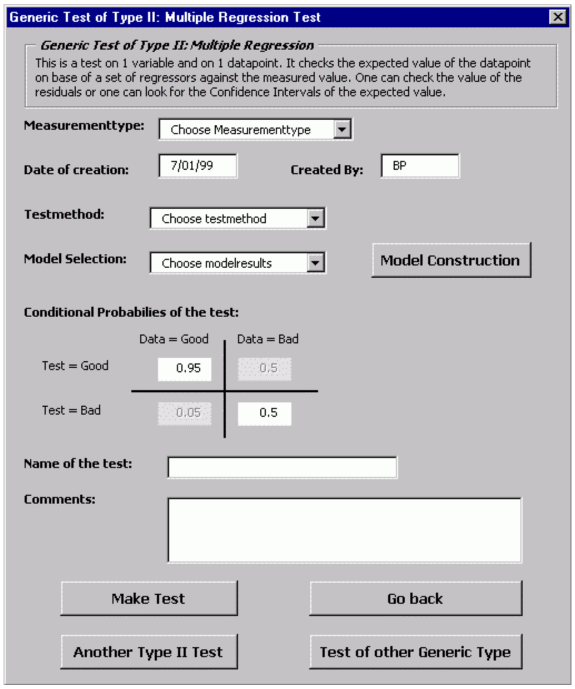
“Test Criterium”
This fields indicates the criterium to be used in the test. Possible choices are <, <=, > , >=, = and !=.

“Type of boundary”
This is a listbox where the user has to select one of following 3 options:

- "constant";
  the value against which the measurement is compared is a constant. If this choice is made, a field is added to the form in which the user should enter the value of the constant;
- "statistic";
  the value against which the measurement is compared is a statistic. If this option is chosen, a listbox is added to the field from which the user can chose the statistic (derived on basis of data in the database) that should be used;
- "function";
  the value against which the measurement is compared is a function combining constants and/or statistics. In this case a field is added in which the user should declare the function to be used. Rules governing the declaration of a function are explained in Section 7 (to be implemented).

Generic Test Type II: Distribution Test (fig. 22)

This test can be used to test a measurement against the univariate distribution of the measurement type. This distribution can be a function of the location of the measurement and/or the time at which the measurement is sampled.



Generic Test of Type II: Multiple Regression Test

Generic Test of Type II: Multiple Regression

This is a test on 1 variable and on 1 datapoint. It checks the expected value of the datapoint on base of a set of regressors against the measured value. One can check the value of the residuals or one can look for the Confidence Intervals of the expected value.

Measurementtype: Choose Measurementtype

Date of creation: 7/01/99 Created By: BP

Testmethod: Choose testmethod

Model Selection: Choose modelresults

Conditional Probabilies of the test:

	Data = Good	Data = Bad
Test = Good	0.95	0.5
Test = Bad	0.05	0.5

Name of the test:

Comments:

Figure 59: Generic Test of Type II

In addition to the general fields, following 2 fields need to be declared:

“Testmethod”
This is a listbox with the 2 possible options:
This is a listbox presenting 2 alternative choices:

- "residual based"
  When this option is selected, the test will be applied as follows. The residual of the measured value with respect to the predicted value by the regression model is calculated and this residual is checked against the lower- and upper-bound value that is exceeded with probability 1-alfa. Statistical uncertainty on the fitted distribution is not considered (i.e. a z-test is used).
- "confidence interval based".
  When this option is selected, the test will be applied as in the previous case, except that a confidence interval is used which accounts for the statistical uncertainty of the fitted distribution (i.e. a t-test is used).

Model selection
This is a listbox that presents the distribution models in the database that can be possibly used. The appropriate model must be indicated by the user.

This form also includes a button "**Model Construction**" which allows the user to use the SAT-program to construct an appropriate model at this time.

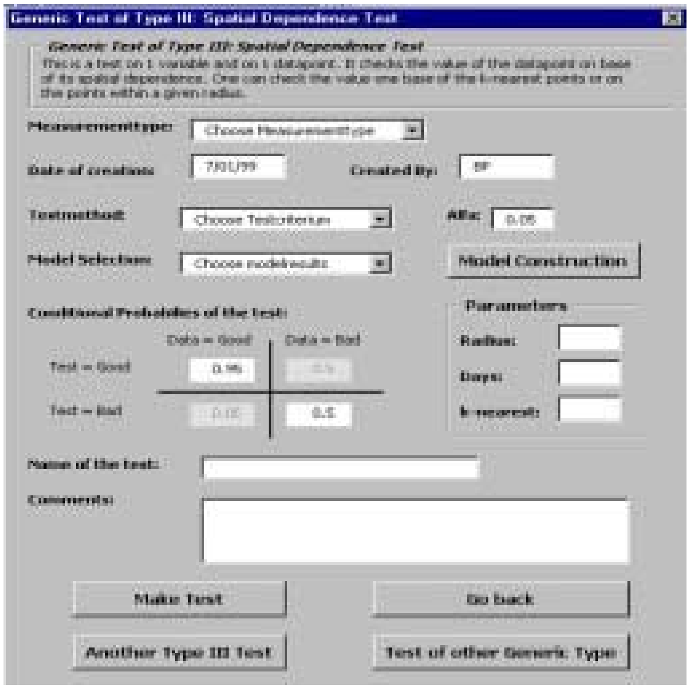
Generic Test Type III: Regression Test (fig. 22)

This type of test can be used to test the value of an observation against other measurements of another type that are obtained at the same time and location.

The form to declare this test is identical to the form used for the distribution test. Again a choice must be made between a “residual-based” test and a “prediction interval” test: in the former case, statistical uncertainty on the fitted regression between the measured physical variable and other variables is neglected, in the second case such uncertainty is accounted for. In the “model selection” listbox the user must choose the appropriate regression model from a list of available regression models stored in the database. Also in this case, the user can construct such a model at the time of defining the test by pushing the “Model Construction” button.

Generic Test Type IV: Spatial interpolation Test (fig. 23)

This type of test is used to test the value of an observation against the value that is obtained by interpolation (Ordinary Kriging) from nearby measurements of the same type.



Generic Test of Type III: Spatial Dependence Test

Generic Test of Type III: Spatial Dependence Test
This is a test on 1 variable and on 1 datapoint. It checks the value of the datapoint on base of its spatial dependence. One can check the value one base of the k-nearest points or on the points within a given radius.

Measurementtype: Choose Measurementtype

Date of creation: 7/01/99 Created By: BP

Testmethod: Choose Testcriterium Alfa: 0.05

Model Selection: Choose modelresults

Conditional Probabilities of the test:

	Data = Good	Data = Bad
Test = Good	0.95	0.5
Test = Bad	0.05	0.5

Parameters
Radius:
Days:
k-nearest:

Name of the test:

Comments:

Figure 60: Spatial Test

In addition to the general fields, following 2 fields need to be declared:

"Test method"
This is a listbox with 2 possible choices:

- "k-nearest points"
  In this case the interpolation uses the k-nearest measurements that fall within a given radius and within a given timelag. The radius (the maximum distance to a neighbor), the timelag (the maximum difference between the measurement times, in number of days) and the value of k (the maximum number of neighbours) have to be filled in by the user. If the number of neighbours within the radius and timelag is less than k, only those measurements are used in the interpolation;
- "within radius".
  In this case the interpolation uses all points within the radius and the timelag that must be specified by the user.

"Model selection"
This is a listbox that presents the available variogram models that can be used to construct the Kriging interpolator. The user must select the appropriate model from this list.

The form also includes a button "**Model Construction**" that allows to construct an appropriate model at this time by using the SAT-program.

Other Generic Test Types

The other generic test types are not yet implemented.

Use of Formulas

In several test procedures functions can be used to declare the parameters. At this moment this option is only used in the Constraint Test. In the future however, a number of parameters in different test procedures could be formulas combining constants and/or statistics.

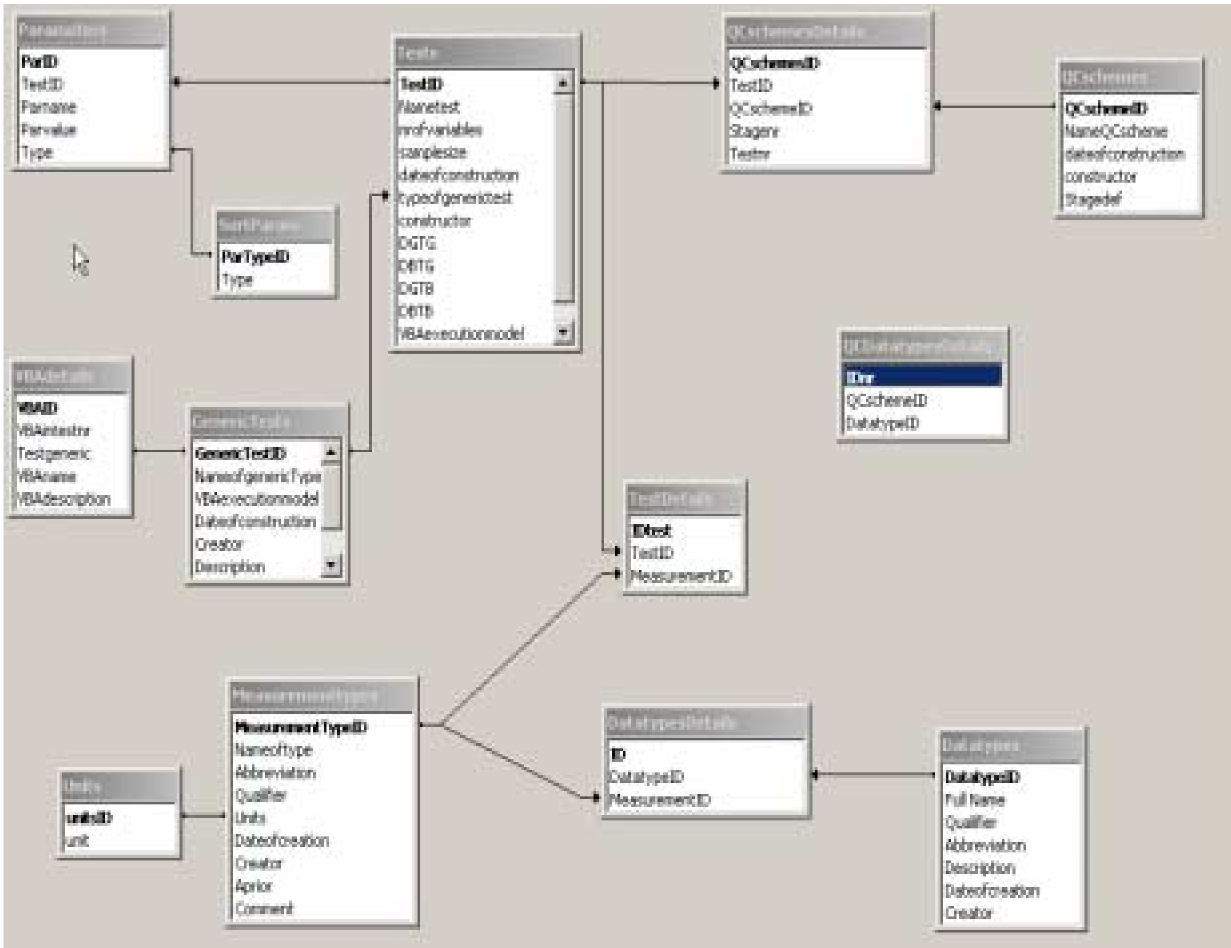
The formulas must be in the same format as Excel formulas, used on an Excel worksheet. All standard build-in functions or user specified functions can be referenced in a formula. Contrary to however the standard Excel formula no cell references can be used. One the other hand, statistics that are stored in the database with statistical results can be referenced by their user-assigned name.

To assist in the construction of a formula, a formula wizard will be developed which lists the most commonly used functions and the names of the statistics that are available in the database.

Appendix D: Structure of QCData.mdb

Structure of the QCData.mdb

QCData.mdb is the database which is used to store the data of the SQC program in. The different datatypes and measurement types defined in the SQC program are stored in this database. The defined Tests and Qcschemes with their details are also kept in this Access database. The structure of the database is given in the following figure.



Parameters
ParID
TestID
Parname
Parvalue
Type

Tests
TestID
Nametest
nrofvariables
samplesize
dateofconstruction
typeofgenerictest
constructor
DGTG
DBTG
DGTB
DBTB
VBAexecutionmodel

QCschemesDetails
QCschemesID
TestID
QCschemeID
Stagenr
Testnr

QCschemes
QCschemeID
NameQCscheme
dateofconstruction
constructor
Stagedef

SortParam
ParTypeID
Type

VBAdetails
VBAID
VBAintestnr
Testgeneric
VBAname
VBAdescription

GenericTests
GenericTestID
NameofgenericType
VBAexecutionmodel
Dateofconstruction
Creator
Description

QCDatatypesDetails
IDnr
QCschemeID
DatatypeID

TestDetails
IDtest
TestID
MeasurementID

MeasurementTypes
MeasurementTypeID
Nameoftype
Abbreviation
Qualifier
Units
Dateofcreation
Creator
Apriori
Comment

Units
unitsID
unit

DatatypesDetails
ID
DatatypeID
MeasurementID

Datatypes
DatatypeID
Full Name
Qualifier
Abbreviation
Description
Dateofcreation
Creator

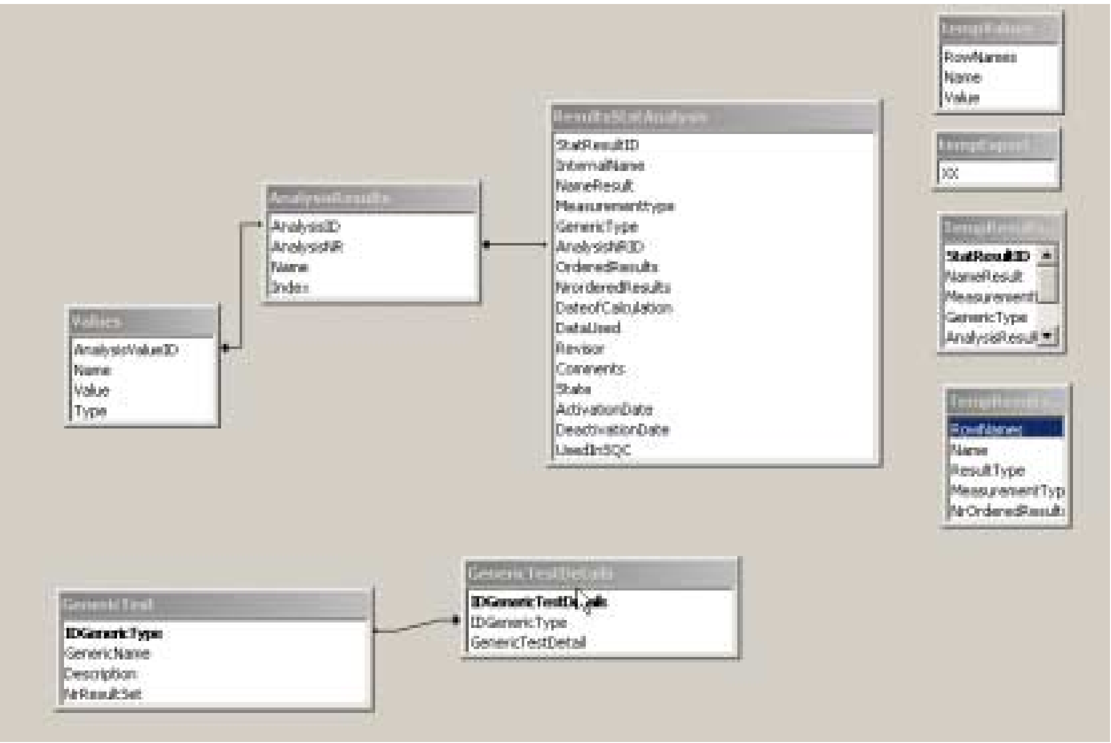
There are 9 tables in this database

Table	Description
Datatypes	Table with the different defined datatypes in SQC.
DatatypesDetails	Intersectiontable giving the relation between the datatype and the measurement type
Measurementtypes	Table with the measurement types defined in SQC
GenericTests	Table with the generic tests that can be used in SQC
VBAdetails	Details for the Generic Test (consist of different calculation methods)
Tests	Table with the different tests created in the IDOD Software
TestDetails	Intersection table with the connection between the test and the measurement type the test is working on
Parameters	Table with the parameters (name of the result in Statresults.mdb) for each test
SortParam	Table with the types of parameters
Units	Table with the different Units
QCschemes	Table with the defined QC schemes
QCschemesDetails	Intersection table giving the tests in a QC-scheme
CurrentUse	Gives the currently used QCscheme in SQC
QCDatatypesDetails	Intersectiontable with the relation between the QCscheme and the datatype it is designed for

Appendix E: Structure of Statresults.mdb

Structure of the StatResults.mdb database

StatResults.mdb is the database with the buffered statistical models created by means of the SAT software. The structure of the database is given in the following figure.



There are 9 tables in this database

Table	Description
ResultsStatAnalysis	Table with the different results which are exported by means of the SPLUS2Buffer software.
AnalysisResults	Detail with the different kind of results exported from SAT.
Values	Table with the details of the exported Statistical Models from the SAT software package
GenericTest	Table with the different Generic Tests
GenericTestDetails	Table giving the type of results in the different Generic Tests
TempResultsStatAnalysis	Table with the details of the exported Statistical Models from the SAT software package
TempResultsSPlusDummy	Table with the details of the exported Statistical Models from the SAT software package
tempExport	Table with the details of the exported Statistical Models from the SAT software package
tempValues	Table with the details of the exported Statistical Models from the SAT software package

IDOD Newsletter

Issue 1 October 1998

Foreword

The purpose of this first issue of the *IDOD Newsletter* is to submit a range of documents describing our present views of some key features of the IDOD database.

These documents are submitted to you as "**Requests For Comments**". You are kindly requested to send any remark to IDOD@mumm.ac.be . If you do not agree with some part(s) of the text, please try to draft an alternate proposal.

Do not wait to long before reading and commenting these documents: the IDOD Committee will meet within 2 weeks to adapt the texts, taking your comments into account…

The IDOD team

MUMM
MANAGEMENT UNIT OF THE NORTH SEA
MATHEMATICAL MODELS

Katholieke Universiteit Leuven
University Center of Statistics

Laboratoire SURFACES
Université de Liège

Inside This Issue

1. IDOD: Goals, Methodology & Planning

by MUMM

The IDOD project ("Integrated and Dynamical Oceanographic Data Management") is funded by the Office for Scientific, Cultural and Technical Affairs in the frame of the "Sustainable management of the North Sea" programme. It officially started in January 1997 and the development and tuning phase of the project will end in December 2001. Afterwards, the tools and expertise built during that period will be kept within the Management Unit of the Mathematical Models of the North Sea, which plays the role of the *Belgian National Oceanographic Data Center*.

> *The purpose of the project is to establish, to manage and to promote a data base of marine environmental data, ensuring a smooth and scientifically sound data flow between the data producers (routine monitoring, field and laboratory experiments, mathematical models,...) and the end users (scientists, sea professionals, policy makers, ...).*

The project can be split into six different –but highly inter-dependent– tasks:

- Inventory of the relevant data sets and data bases,
- Design of the data base
- Development of the procedures pertaining to the incoming flow of data (practical aspects, data quality control, *etc*.),
- Development of data analysis tools (statistical techniques, time– and space analysis, *etc*.),
- Incorporation of validated results of mathematical models (currents, water heights, swell, *etc*.),
- Development of procedures for accessing and diffusing the data and their derived products (maps, tables, reports, *etc*.)

The network supporting this project is made of :

- The Management Unit of the Mathematical Models of the North Sea (Royal Belgian Institute for Natural Sciences)

  URL: http://www.mumm.ac.be/

- The *Universitair Centrum voor Statistieken* (*Katholieke Universiteit Leuven*)

  URL: http://www.kuleuven.ac.be/ucs/

- The *Laboratoire SURFACES* (*Université de Liège*)

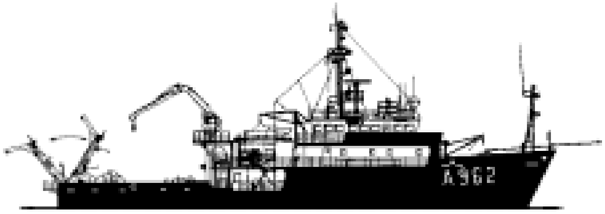
  URL: http://www.geo.ulg.ac.be/Surfaces/

It is not possible to enter here in the details of the methodologies applied to the various parts of the project. Our commitment, however, is to apply wherever possible the state–of–the–art standards and techniques and to confront them to data providers and users. In this issue, for example, the reader will find an overview of the methodologies applied to develop the *Conceptual Model* of the data base and the *Quality Control Scheme*.

An updated planning of the whole project will be made available soon. Please take note, however, of the following major deadlines:

- End of 1998 : a prototype of the data base system is available for demonstration.
- End March 1999 : the first batch of data sets is sent to IDOD by the laboratories participating in the programme.
- Mid-1999 : First release of an operational version of the data base system. ♦

For more information, comments or remarks :
IDOD@mumm.ac.be



2. Conceptual Model of the IDOD data base

by MUMM & ULG

This section presents the conceptual model established for the IDOD database. The text is an explanation to the documents given in annex : the "Synthetic Conceptual Model of the IDOD database", and the "Dictionary of entities and attributes".

2.1. Entity – Attribute - Relation

The scheme (see annex 1) describes in a synthetic manner the conceptual model of the IDOD database. It is attached to the data dictionary in order to better understand the general structure of the data. In this scheme, each rectangle represents an *entity* as described in the dictionary: e.g. *PROJECT*, *SAMPLE*, *PARAMETER*, etc. The entity allows treating a collection of objects of the same nature. It expresses a type, a class, a set of which the elements are described by the same list of attributes. The entity is associated to the notion of table, characterised by its fields, in a database.

Each entity has *attributes* that describe it. The attribute represents elementary information, having a sense on its own, e.g. city code, name, …. For each of the entities, the attributes are fully described in the document "Dictionary of entities and attributes – IDOD database" annexed to the present newsletter.

Entities are linked by *relations*, symbolised here with lines between the entity boxes. For example, a relation links the entities *PROJECT* and *PERSON* because one project involves one or more persons and one person is involved in one project. The characteristics of the relations (e.g. one-to-many, one-to-one, many-to-many,…) are not indicated on the scheme but can be found in the dictionary.

The *QUALITY ANALYSIS* entity is specialised into three sub-entities: *QUASIMEME*, *CONTROL CHART* and *INTER-CALIBRATION* that are the only existing entities, the *QUALITY ANALYSIS* alone cannot exist.

2.2. IDOD Conceptual Model

The scheme (annex 2) represents the storage structure of concentration values measured on individual water samples, and the corresponding meta-information. Similar schemes are being developed for other types of data: continuous measurements of scalar data, sediment cores, biota, …

All values, resulting from analyses and measurements, are stored in a table *NON-CONTINUOUS VALUE*. Together with the value, the date, time, position, as well as quality control and access information is stored. The link with *PARAMETER* provides information on the parameter measured and the unit in which the result is stored. *ANALYSIS METHOD* describes the method used to obtain this value.

One or more values result from one *SUBSAMPLE*, and one or more subsamples result from a *SAMPLE*. For *SAMPLE* and *SUBSAMPLE* a link with respectively *SAMPLING METHOD* and *SAMPLE HANDLING* explains the method used.

Each sample is linked to the *CAMPAIGN* during which it was measured and to the *PROJECT* or programme in the frame of which the sample was collected. *PROJECT* and *CAMPAIGN* contains more descriptive information about the objectives and date, and have links to one or more *SERVICE* and *PERSON* involved. For *CAMPAIGN* the link with *PLATFORM* provides information on the basis used for the measurements. This can be a ship, on foot, a buoy.

QUALITY ANALYSIS and *QUALITY SAMPLE & HANDLING* store the quality control information. The method used, the service and the parameter analysed determine together the quality of the analysis and of sampling and sample handling. ♦

For more information, comments or remarks :
IDOD@mumm.ac.be

This document and its annexes are available on the IDOD anonymous ftp server :

ftp://idod.mumm.ac.be

3. Copyright and access rules

by MUMM

So far, the "Copyright and access rules" of data kept in the IDOD database can only be sketched. It is indeed not straightforward to combine the various rules (International treaties[1], European directives[2], national laws[3], contractual documents[4], *etc.*) that apply to the topic.

MUMM is active in two groups that, at the European level, try to solve the problem:

- EURONODIM, a consortium of 15 oceanographic data centres, funded by the EC–DG XII for a period of three years starting September 1st, 1998.
- EuroGOOS, the European branch of the Global Ocean Observing System (IOC–WMO–UNEP–ICSU)

It is our commitment to inform all parties involved in the setting–up of the IDOD database of the outcomes of the work of these two groups on this topic.

Regarding the "Copyright", it must be remembered that the question involves several "actors" and aspects. As stated in the *Data Policy Handbook* of the U. K. Natural Environment Research Council (about "Ownership of the data"), "*Despite behaviour that might suggest the contrary, data sets frequently do not belong to those who have collected them. They generally belong to the employers of such data collectors or to those who have paid for the data collection.*" (NERC Data Policy Handbook, Version 2.0, February 1998)

Furthermore, it is generally accepted that the individual scientists, principal investigator teams and programmes be permitted a reasonable period of exclusive access to the data sets which they have collected, allowing them to work on them and produce publications.

With all these aspects in mind, one must understand that "Copyright and Access rules" will most of the time be defined on an *ad hoc* basis…

[1] E. g., the Antartic Treaty, for the dissemination of data from the Antartic.
[2] *E.g.*, the European Directive on the protection of data bases (11.03.1996).
[3] *E.g.*, the law on authorship (30.06.1994), the law on the "Freedom of Access to Environmental Information".
[4] The funding contract of the research but also the contract that binds the data collector to his/her institution.

In the frame of the "Sustainable Management of the North Sea" programme, the following access rules have been suggested so far:

- The embargo for scientific use is set to 24 months after data collection.
- There are four categories of users :
  - A. The data collector (usually the team to which the actual data collector belongs),
  - B. The scientific teams involved in the "Sustainable Management of the North Sea" programme,
  - C. The Belgian federal administration (the OSTC, as sponsoring body, and MUMM, as designated federal body for the management of the marine environment),
  - D. The other users.

The basic access scheme can then be sketched as follows:

Users:	A	B	C	D
Data:	Free	Free : ▪ As a source of information for their own research only ▪ With the explicit consent of the data collector (else: after embargo)	Free : ▪ For use in their activities of policy support only (else: after embargo)	After embargo

For more information, comments or remarks : IDOD@mumm.ac.be

4. Quality Control and Statistical Analysis in the IDOD project

by KUL

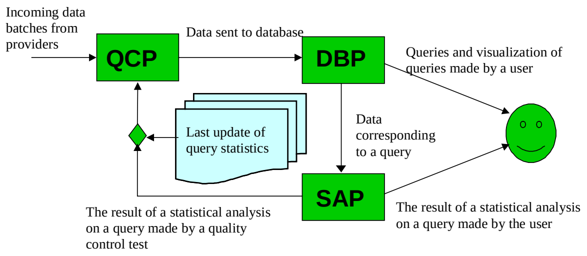
4.1. Introduction

As part of the IDOD project, UCS is responsible for the development of tools for quality control and statistical analysis and their integration in the IDOD database. After several months of familiarisation with the problem and conceptual design, UCS is now starting to look at real data (i.e. monitoring data from the vessel "Belgica"). While we haven't reached yet the level of developing specific quality control schemes or statistical tools, we are rapidly approaching this stage. Because the issue of quality control and statistical analysis is of potential interest to all researchers involved in the broader research programme of "Sustainable Development of the North Sea", this note is meant as a first introduction to our efforts and will be followed up by a presentation of a first example application (hopefully in the last part of the year). We kindly invite all potential data providers and researchers involved in this program to send us their comments and/or suggestions.

4.2. Overall Scheme

The overall organisation of the data flow is presented pictorially in the following figure:

Dataflow within IDOD



Incoming data batches from providers
QCP
Data sent to database
DBP
Queries and visualization of queries made by a user
Last update of query statistics
Data corresponding to a query
SAP
The result of a statistical analysis on a query made by a quality control test
The result of a statistical analysis on a query made by the user

The idea is that data (which satisfy overall quality requirements regarding meta information and quality assurance) are input to a front-end program QCP ("Quality Control Program") that will attempt to verify the "consistency" of the data amongst themselves and with existing data in the database. Within this context, "consistency" should be understood rather loosely as the comparison of the incoming data with known physical relationships and bounds and with statistical limits that are derived from the data already qualified in the database DBP (DBP stands for "Data Base Program"). The type of "consistency" checks will evidently vary from measurement type to measurement type.

Nevertheless, a general applicable framework for the application of these checks and the reporting of the results has been already developed and will be briefly outlined in the following paragraph.

Once the data are entered into DBP, the data can be examined by the user either through direct queries or, indirectly, through SAP ("Statistical Analysis Program") which will support a variety of statistical analyses. SAP is also used by the QCP program to retrieve those statistics that are needed to perform the consistency checks. To limit CPU time, the retrieved statistics are stored and only updated upon explicit request of the data manager (i.e. when the number of available data has significantly increased).

In the following 2 paragraphs, we describe in additional detail how the operation of the QCP and SAP program is envisioned.

4.3. The Quality Control Program

The QCP program is being developed under EXCEL97 using Visual Basic for Applications programming. In this manner, advantage is taken of the fact that EXCEL already supports a wide variety of input formats (e.g. ACCESS databases, EXCEL spreadsheets, ASCII files, ...).

For each type of data (that is a set of measurements of a given type: i.e. latitude, longitude, water temperature, dissolved oxygen, salinity), separate quality control schemes will be developed which share however a common structure. Before describing this structure, it should be emphasised that the results of this quality control are meant to be "informative" and are by no means final. All data that satisfy the more global norms of quality assurance (documentation, qualified lab, etc.) are retained in the final database and evidently every user of the database should judge whether he wants to use or not use the automated quality information. Furthermore, it is envisioned that the data manager and/or data provider can add comments and may assign his own quality labels that may differ or coincide with the previous quality labels.

In a quality control scheme, the data are passed through a series of tests each of which returns a quality result: i.e. "G" (good), "B" (bad) or "M" (missing: meaning that the test cannot be executed, i.e. because of lack of data). The tests are grouped into different stages, where each stage operates on a given window of the data. For instance, a window of size 1 implies that the tests verify only the measurements in a single record. A window of size n means that the tests simultaneously verify the measurements in n subsequent records (thus allowing to check trends, stationarity or non-stationarity of the values, etcetera). For a specific measurement (i.e. salinity) and for a given stage, the end result consists of a qualifier string that combines the outcomes of the different tests in the stage that apply to that measurement type (*i.e.* "BBGMGG"). Knowing the significance level and the power of the tests (i.e. how often would a test qualify by chance a good data point as bad and how often would a bad data point be qualified as good by the test) and knowing the overall a-priori frequency of erroneous measurements, it is then possible to calculate the probability that a measurement is not erroneous. In a loose sense, this probability can be interpreted as a numeric measure of quality, with a 0 indicating that the data point is clearly wrong and a 1 indicating that the data point is with certainty good (this last value will normally never be assigned). A threshold value for the probability allows to separate the data into "qualified" and "disqualified" measurements. Measurements that are "disqualified" in a given stage are no longer taken into consideration for further testing or used in any of the subsequent tests.

The quality control scheme described above is sufficiently general to comprise a wide variety of different test schemes. The advantage of such a general framework is that different schemes can be build and documented in a standardized manner and that problems of missing values, dataflow and combination of test results are appropriately dealed with. Also the "backtracking" (the report which explains why a given measurement is "disqualified") can be organized at this general level.

In its application to real data, specific schemes comprising specific tests will of course need to be developed and the tests should reflect knowledge about the measurements, their accuracy and their relation to other variables. Some of that knowledge may originate directly from the data provider, while other findings may be the result of a statistical analysis of a sample of the data. UCS is currently examining the "Belgica" monitoring data to determine what type of tests could be of use. Results of this analysis and the preliminary quality control scheme that follows will be reported at a later stage for review.

4.4. The Statistical Analysis Program

The statistical analysis program should support both general and more specific statistical analyses of the data. While work on SAP is still in its initial phase, we envision that use will be made of a general statistical package. First evaluations indicate that SPlus would be the best choice and could be integrated into the overall scheme sketched earlier.

On top of this package, the full functionality of which we intend to make available to expert users, more specific applications meeting specific requests (either by the user or the quality control program) will be developed. There is also the possibility to incorporate the most commonly used functions of the SPlus program (i.e. regression analysis, trend analysis) into user interfaces ("wizards") that require no prior knowledge of the statistical package and/or that meet specific requirements of oceanographic research.

After the choice of the package has been consolidated, UCS will first of all work on the integration of the package with the database and with the quality control program. Once this work has been done, UCS plans to demonstrate the technology to the various partners of the program for "Sustainable Development of the North Sea" and we suggest that at that time further thought is given by all partners involved to specific applications of interest and the development of "wizards". ♦

For more information, comments or remarks :

IDOD@mumm.ac.be

5. Guideline for data documentation

Note : This document has already been circulated. It is included here as a reminder. However, it is still open to discussion.

5.1. Introduction

This draft guideline is based on an equivalent document set up by the *MAST Data Committee*. Together with the "Guideline for Project Data Management", it has been written to help teams to implement the "Code on Data Management in MAST Projects".

We found it a good reference to implement the obligation for every laboratory working in the frame of the programme «Sustainable Management of the North Sea» to transfer their data to the «Federal Oceanographic Data Base» being developed in the same context.

The guideline is presented in the form of a checklist of actions which goes into some detail but does not intend to be comprehensive. Appropriate choices have to be made by each project regarding its proper data documentation. Comprehensive documentation for your data may be less simple than expected. Therefore a word of caution. Put yourself in the position of the possible user. It would be a pity if your data are ignored once because nobody trusts their quality due to lacking documentation. The objective of compiling data documentation is to ensure that data are consistently well described, that the quality and limitations of the data are apparent, and that there is sufficient information available to assess the suitability of the data for a particular task. The end result of good data documentation is to contribute to usable data sets of known quality by accompanying them with supporting documentation. For the purpose of this document, data production is considered in two parts.

Data Acquisition - which includes the collection, processing and analysis of raw material or raw data up to the point where data is generated; and

Data Processing - which involves the further manipulation, processing or enhancement of this generated data.

Please note that not all aspects of the guideline will apply for all data types and users are advised to use as appropriate. Please consider that data appear in different forms in different projects -data may be analog readings, numeric, charts, images, samples, specimen, *etc*., and data may generated in the laboratory, from a physical model, or gathered in a field experiment.

5.2. Checklist on data documentation

Data Acquisition

Measurement/Sample Collection

A. Describe the equipment used. Give name, gear code and details of its deployment as appropriate. If possible include references for the method/gear being deployed.

B. Describe the measurement platform and describe the techniques used for positioning the platform and equipment.

C. Give details of calibration of the equipment used, document details on sampling frequency, list experts involved.

D. Describe environmental conditions as appropriate, including special conditions that can affect the representativity of the data (*e.g.* local pollution)

E. Any limitations associated with using the chosen sampling strategy should be documented.

Sample Processing, Sample Analysis

1. Document how the sample was preserved.
2. State whether sample was processed in-situ, shipboard or laboratory. Give the name of the Laboratory where samples were processed .
3. Describe different stages in sample processing (including instrumentation used) from treatment of sample from when it is collected through to final processing of sample to the point where data is generated (details of the methods/instrumentation used should be given with full references).
4. Describe any instrumental error corrections made during the course of processing the sample.
5. Describe the expected precision, accuracy or reproducability of the used methodology and the limits of detection.
6. Give details if the methodologies employed in sample processing have been validated.
7. Comment on any limitations associated with used sample processing techniques.
8. Use references to published literature where ever possible in order to shorten and compact the information.

Data Processing/Analysis

1. Describe clearly the different stages in processing and analysing the data including reduction algorithms, statistical analysis, etc. Details of the methods used should be given with full references.
2. Describe algorithms and computer programmes used.
3. Give details of the degree of precision, accuracy and reproducibility of the data processed/analysed where known.
4. Give details if the methodologies employed in data processing/ analyses have been validated.
5. Please comment on any limitations incurred in using these sample processing techniques
6. Use references as appropriate - give full bibliographic reference.

Quality Control and Assurance Information

1. Quality control and assurance exercises can be carried out at each of the data generation stages as described above. Therefore for each stage, describe and include documentation that will give information on the quality of the data. That is:
2. Describe your cruise properly in a Cruise Report.
3. Describe exercises carried out to standardise the data or to calibrate your data.
4. Describe any data validation exercises carried out in response to the results of calibration and intercalibration as well as comparison with standard methods.
5. Give details of the detection of limitations, gaps or errors in position and time.

Data Formats

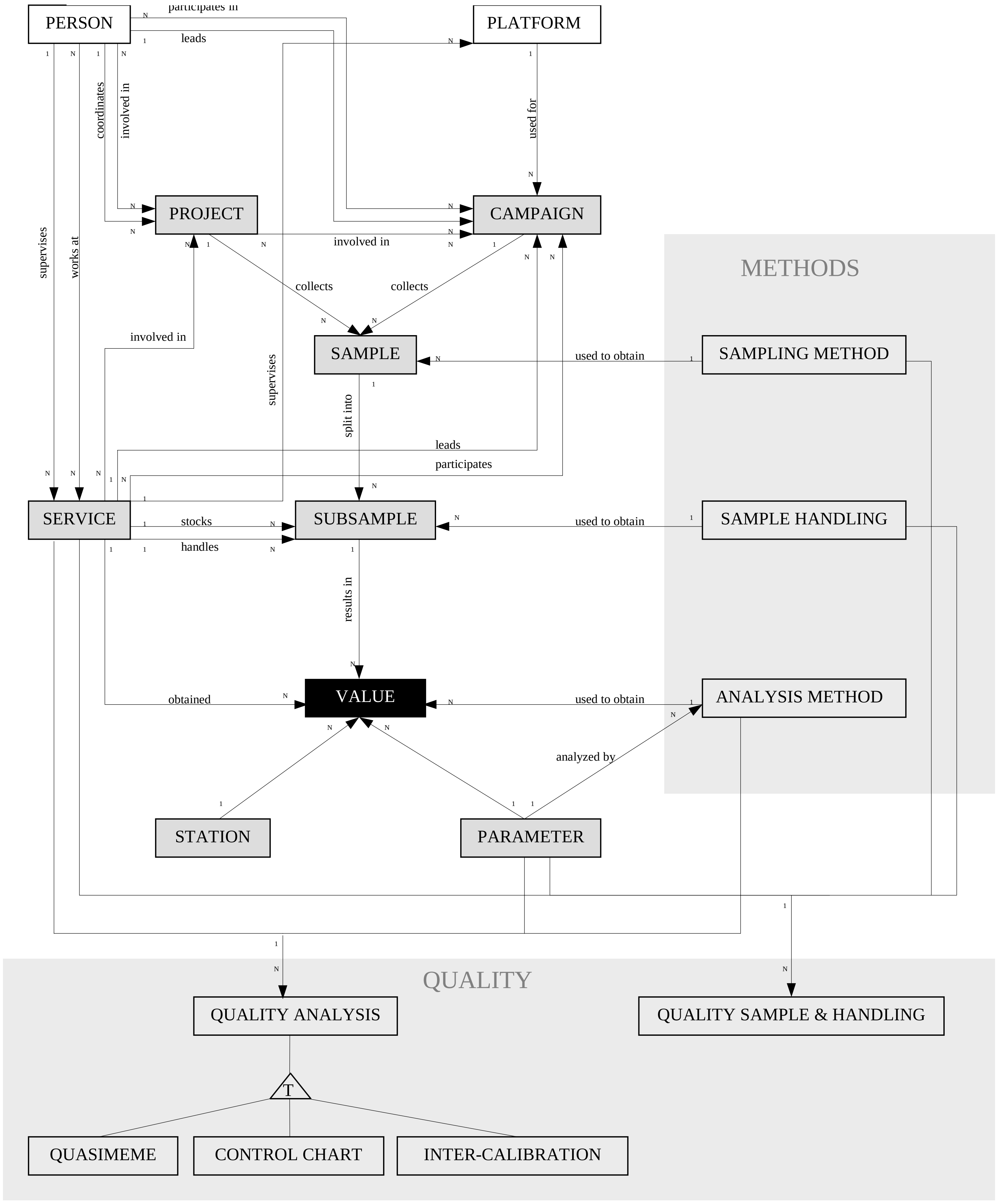
When compiling data files, it is essential that the contents of a data file can be clearly understood by a subsequent user and that the essential information such as station numbering, references to position, time and depth are always clearly identifiable. All data formats must be well described to ensure that there are no misunderstandings and that the data will be usable by the receiver.♦

For more information, comments or remarks :

IDOD@mumm.ac.be

The *IDOD* Newsletter – Issue 1
has been set up by
K. De Cauwer & S. Scory

Annex 1: Synthetic Conceptual Model of the IDOD database for point measurements in seawater



DICTIONARY OF ENTITIES AND ATTRIBUTES

IDOD DATABASE

SURFACES

Université de Liège

MUMM

23/10/98

CONTENTS

REMARKS

For each entity, the attributes, describing it are listed. An explanation, the type of data, the format, an example and comment is given for each field. Besides the attributes proper to the entity, the relations to other entities are described. Those links are symbolized by a ‘@’-sign.

Format « Float »
All variables of format float will be stored as two integers : mantissa and exponent. As such, the original number of significant figures can be represented.

Datatype « List »
For datatype ‘List’, the list of possible entries is given when this list is short. Reference is given to long lists.

DATABASE : IDOD ENTITY : SERVICE

DESCRIPTION :

The SERVICE entity describes a scientific team in charge of collecting and analysing data related to the sea environment.

ATTRIBUTES :

NAME	EXPLANATION	TYPE	FORMAT	EXAMPLE	COMMENT
Service code	A code to identify one occurrence of the SERVICE entity in the database.	Key			Required, unique, system-generated
English name	The service name in English.	String	100	Management Unit of the North Sea Mathematical Models	At least 1 required
French name	The service name in French.	String	100	Unité de gestion du modèle mathématique de la Mer du Nord	
Flemish name	The service name in Dutch.	String	100	Beheerseenheid mathematisch model van de Noordzee	
Address in English	The service address in English.	String	255	100 Gulledelle, B-1200 Brussels	At least 1 required
Address in French	The service address in French.	String	255	Gulledelle 100, B-1200 Bruxelles	
Address in Flemish	The service address in Dutch.	String	255	Gulledelle 100, B-1200 Brussel	
Phone	The main phone number of the service.	String	25	+32 2 773 21 11	Required
Fax	The main fax number of the service.	String	25	+32 2 770 69 72	
Email	The main email address of the service.	String	80	G.Pichot@mumm.ac.be	
Institute name in English	The name in English of the institute to which the service belongs.	String	100	Royal Belgian Institute for Natural Sciences	At least 1 required
Institute name in French	The name in French of the institute to which the service belongs.	String	100	Institut Royal des Sciences Naturelles de Belgique	
Institute name in Dutch	The name in Dutch of the institute to which the service belongs.	String	100	Koninklijk Belgisch Instituut voor Natuurwetenschappen	
Institute code	Code of the institute	List		RBINS	
Laboratory code	Code of the laboratory	List		MUMM	Required
ICES code	Laboratory and institute codes according to ICES	List		MUMM *List ://www.ices.dk/env/repfor/vlabos.html*	
ISO country code	country code from ISO 3166	List		056	Required
Description	Reference to documents describing the service procedures and organisation to provide a more complete description	Memo			
Service website	URL of the service website	String		//www.mumm.ac.be	
@ Head of department	One link to one person who is head of the service.	Link			Required
@ Members	One or more links to the persons who are working in the service; excluding the head of department.	Link			
@ Projects	One or more links to the projects in which the service is involved.	Link			
@ Campaigns heading	One or more links to the campaigns leaded by the service.	Link			
@ Campaigns participation	One or more links to the campaigns in which the services participates.	Link			

@ **Values**	One or more links to the values obtained from measurements realised by the service.	Link			
@ **Quality samples & handling**	One or more links to quality of samples and sample handling operated by the service.	Link			
@ **Quality analyses quasimeme**	One or more links to the quasimeme exercises (quality of analysis) realised by the service.	Link			
@ **Quality analyses control chart**	One or more links to the information about control charts (quality of analysis) kept by the service.	Link			
@ **Quality analyses inter-calibration**	One or more links to the inter-calibration exercises (quality of analysis) realised by the service.	Link			
@ **Platforms**	One or more links to the platforms supervised by the service.	Link			
@ **Subsamples**	One or more links to the subsamples kept by the service.	Link			

DATABASE: IDOD **ENTITY : PERSON**

DESCRIPTION :

The PERSON entity describes a physical person is involved in a project, campaign or service related to maritime topics.

ATTRIBUTES :

NAME	EXPLANATION	TYPE	FORMAT	EXAMPLE	COMMENT
Person code	A code to identify one occurrence of the PERSON entity in the database.	Key			Required, unique, system-generated
First name	The first name of the person	String	15	Georges	Required
Last name	The last name of the person	String	15	Pichot	Required
Position	The main function of the person in his service	List		*List :* - *Head of department* - *Project co-ordinator* - *Scientist* - *Laboratory worker* - *Informatician* - *Laboratory responsible* - *Secretary*	
Personal phone	The personal phone number of the person.	String	25	+32 2 773 21 22	
Personal email	The personal email address of the person.	String	80	G.Pichot@mumm.ac.be	
@ **Projects heading**	One or more links to the projects supervised by the person.	Link			
@ **Projects participation**	One or more links to the projects in which the person is involved.	Link			
@ **Campaigns heading**	One or more links to the campaigns supervised by the person.	Link			
@ **Campaigns participation**	One or more links to the campaigns in which the person is involved.	Link			
@ **Service heading**	One link to the service supervised by the person.	Link			
@ **Services participation**	One or more links to the services in which the person is engaged.	Link			Required

DATABASE: IDOD **ENTITY : PROJECT**

DESCRIPTION :

The PROJECT entity describes a scientific project concerning maritime topics.

ATTRIBUTES :

NAME	EXPLANATION	TYPE	FORMAT	EXAMPLE	COMMENT
Project code	A code to identify one occurrence of the PROJECT entity in the database.	Key			Required, unique, system-generated
Name	The project name.	String	256	Monitoring of the quality of the marine environment	Required
Project acronym	Project acronym	String	15	Monit	
Start date	The start date of the project.	Date	8 : ddmmyyyy	01/01/1977	
End date	The end date of the project.	Date	8 : ddmmyyyy		
Sponsor	The name of the sponsoring body.	String	100	Belgian state	
Research programme	The research programme to which the project belongs.	String	255	Structural programme	
Theme	The theme of the project	List		*List :* - *Monitoring* - *Geology* - *Ecosystem* - *Fisheries* - *Data Management* - *Modelling*	Required
Keywords	A list of keywords describing the project	String	100		
Abstract	A text description of the project.	Memo		The North Sea environment is monitored for international purposes : Oslo and Paris Commission, the Joint Assessment …	Required
Study area	A text describing the geographical coverage of the project.	Memo		Belgian Continental Shelf and Western Scheldt estuary	Required
Objectives	A text describing all the objectives of the project.	Memo		Monitoring and evaluation of the quality of the marine environment of the Belgian Continental Shelf en the Scheldt estuary	Required
Project website	URL of the project website	String	100		Automatically
Reference	Reference to project proposal for more information about the project.	Memo			
@Parameters	A link to the parameters analysed in the frame of the project.	Link			
@ Co-ordinating organisation	One link to the service co-ordinating the project.				Required
@ Co-ordinator	One link to the person who is in charge of the project supervision.	Link			
@ Members	One or more links to the persons who are involved in the project.	Link			Required

@ **Services**	One or more links to the partner services	Link			
@ **Samples**	One or more links to the samples collected in the scope of the project.	Link			
@ **Campaigns**	One or more links to the campaigns in which the project is involved.	Link			

DATABASE: IDOD **ENTITY : CAMPAIGN**

DESCRIPTION :

The CAMPAIGN entity describes a well-defined period of measurements related to the maritime domain on/using a certain platform (e.g. a Belgica cruise, continuous series of measurements with buoy).

ATTRIBUTES :

NAME	EXPLANATION	TYPE	FORMAT	EXAMPLE	COMMENT
Campaign code	A code to identify one occurrence of the CAMPAIGN entity in the database.	Key			Required, unique, system-generated
Name	The name of the campaign.	String	50	Belgica 98/8	
Start date	The start date of the campaign.	Date	8 : ddmmyyyy	15/04/1998	Required
End date	The end date of the campaign.	Date	8 : ddmmyyyy	17/04/1998	Required
Port of departure	The name of the departure harbour + country code (in case of ship platform).	String	25	Zeebrugge, BE	
Port of arrival	The name of the return harbour + country code (in case of ship platform).	String	25	Zeebrugge, BE	
Objectives	A text to describe the objectives of the campaign.	Memo		(see campaign reports)	
Area description	A text to describe the geographical area covered by the campaign.	Memo		Belgian and Dutch (till 3°20’ EL & 52°N) continental shelf and Scheldt estuary	Required
Marsden square	Code referring to 10° x 10° squares to locate the campaign area.	String	3	216	
IHB area	International Hydrographic Bureau area	String		North Sea	
@ Projects	One or more links to the projects involved in the campaign.	Link			
@ Responsible laboratory	One link to the service that supervises or leads the campaign.	Link			Required
@ Chief scientist	One link to the person who supervises the campaign.	Link			
@ Participating laboratories	One or more links to the services participating in the campaign.	Link			
@ Principal investigators	One or more links to the persons responsible for the data collected on the cruises and who may be contacted for further information about the data.	Link			Required
@ Collected samples	One or more links to the samples collected during the campaign.	Link			
@ Continuous values	One or more links to continuous values measured during the campaign.	Link			
@ Platform	One link to the platform used during the data collection	Link			Required

DATABASE: IDOD	ENTITY : SAMPLE

DESCRIPTION :

The SAMPLE entity describes a physical sample of matter collected during a campaign on a certain location (e.g. Niskin bottle of water).

ATTRIBUTES :

NAME	EXPLANATION	TYPE	FORMAT	EXAMPLE	COMMENT
Sample code	A code to identify one occurrence of the SAMPLE entity in the database.	Key			Required, unique, system-generated
Water depth	Depth of water at sampling time and position.	Float		16 meter	meters
Meta wind speed	Average wind speed at sampling time and position	Float		12.03 m/s	m/s
Meta wind direction	Average wind direction at sampling time and position.	Float		303 degrees	Degrees
Meta sea state	Average sea state at sampling time and position.	Integer		6 Beaufort	Beaufort
Meta atmospheric	Average atmospheric pressure at sampling time and position.	Float		1012 mbar	mbar
Meta air temperature	Average air temperature at sampling time and position.	Float		13 °C	°C
@ **Campaign**	One link to the campaign during which the sample was collected.	Link			Required
@ **Project**	A link to the project related to the sample.	Link			Optional
@ **Subsamples**	One or more links to the subsamples obtained by subdivision of the sample.	Link			Required
@ **Sampling method**	One link to a sampling method describing the technique used for the collection of the sample.	Link			Required

DATABASE: IDOD **ENTITY : SUBSAMPLE**

DESCRIPTION :

The SUBSAMPLE entity describes a subdivision of a sample obtained after separation, prior to analysis.

ATTRIBUTES :

NAME	EXPLANATION	TYPE	FORMAT	EXAMPLE	COMMENT
Subsample code	A code to identify one occurrence of the SUBSAMPLE entity in the database.	Key			Required, unique, system-generated
@ **Subsample service**	A link to the service who handles the sample	Link		*IHE*	Required
@ **Stock area**	A link to the service where the subsample is kept stock.	Link		*MUMM*	Optional
@ **Sample**	One link to the sample from which the subsample is derived.	Link			Required
@ **Sample handling**	One link to a sample handling method that describes how the sample is pre-treated and separated in order to obtain the subsample and how the subsample is preserved.	Link			Required
@ **Value**	A link to the value obtained by the analysis of the subsample.	Link			Optional

DATABASE: IDOD **ENTITY : SAMPLING METHOD**

DESCRIPTION :

The SAMPLING METHOD entity describes how a sample is collected.

ATTRIBUTES :

NAME	EXPLANATION	TYPE	FORMAT	EXAMPLE	COMMENT
Sampling method code	A code to identify one occurrence of the SAMPLING METHOD entity in the database.	Key			Required, unique, system-generated
Sampler	To describe the recipient with which the sample is taken	List		*List (water) :* – *Go-Flo* – *Niskin bottle or analogous* – *pump* – *bucket* – *detector/probe* – *meter* – *other*	Required
Sampler deployment	To describe how the recipient is handled	List		*List :* – *winch* – *teflon tube* – *nothing/rope* – *nothing/RIB* – *other*	Required
Description	Description of sampling method				Required if sampler or sampler deployment ="other"
Sampler example	A picture of the described sampler.	Image			
Sampler handling example	A picture showing sampler handling.	Image			
@ Samples	One or more links to the samples collected using the sampling method.	Link			Required
@ Quality samples & handling	One or more links to the quality of samples & sample handling.	Link			

DATABASE: IDOD	ENTITY : SAMPLE HANDLING

DESCRIPTION :

The SAMPLE HANDLING entity describes the methods used for pre-treatment and separation/division of the sample in subsamples, and preservation of the subsample.

ATTRIBUTES :

NAME	EXPLANATION	TYPE	FORMAT	EXAMPLE	COMMENT
Sample handling code	A code to identify one occurrence of the SAMPLE HANDLING entity in the database.	Key			Required, unique, system-generated
Sample preservation	Method of storage/preservation	List		*List :* – *determined at once/on the spot* – *kept at room temperature* – *kept in fridge* – *kept in deep-freeze* – *other*	Required
Sample pre-treatment	Method of pre-treatment	List		*List :* – *none* – *acidification* – *other reagent*	Required
Sample separation	Method of separation	List		*List (water) :* – *unfiltered* – *filtered on membrane of 0,2 µm* – *filtered on membrane of 0,45 µm* – *filtered on membrane of 0,8 µm* – *filtered on glassfiber GF/C 1,2 µm* – *other*	Required
Procedure description	A synthetic and chronological description of the methods used. Detailed information if one of three previous attributes is 'other'.	Memo		After filtration, the sample is kept in the freezer.	Required
@ Subsamples	One or more links to the subsamples obtained by this handling	Link			
@ Quality samples & handling	One or more links to quality of sample handling.	Link			

DATABASE: IDOD

ENTITY : ANALYSIS METHOD

DESCRIPTION :

The ANALYSIS METHOD entity describes the method used for the analysis of the subsample.

ATTRIBUTES :

NAME	EXPLANATION	TYPE	FORMAT	EXAMPLE	COMMENT
Analysis method code	A code to identify one occurrence of the ANALYSIS METHOD entity in the database.	Key			Required, unique, system-generated
Description	The complete description of the method used for the analysis and how the corresponding measured value is obtained.	Memo		Autoanalyzer adaptation of the Wood et al. (1967) method of cd cu reduction to nitrite, followed by a diazoreaction with sulfanilamide	Required
Detection limit	The limit of detection (lowest value that can be measured) with this analysis method.	Float		0.1	
Detection unit	The unit in which the detection limit is expressed.	String	8	µmol/l	Required if previous attribute exists.
@ Values	One or more links to the values obtained by this analysis method.	Link			
@ Parameter	One link to the parameter analysed by this method.	Link			Required
@ Quality analyses quasimeme	One or more links to the quasimeme exercises (quality of analysis) realised by the service.	Link			
@ Quality analyses control chart	One or more links to the information on control charts (quality of analysis) kept by the service.	Link			
@ Quality analyses inter-calibration	One or more links to the inter-calibration exercises (quality of analysis) realised by the service.	Link			

DATABASE: IDOD ENTITY : QUALITY ANALYSIS - QUASIMEME

DESCRIPTION :

The QUALITY ANALYSIS – QUASIMEME describes the result of the ICES or QUASIMEME intercalibration/intercomparison exercises for a parameter determined with a particular analysis method by a service.

ATTRIBUTES :

NAME	EXPLANATION	TYPE	FORMAT		COMMENT
Quasimeme code	A code to identify one occurrence of the QUALITY ANALYSIS - QUASIMEME entity in the database.	Key			Required, unique, system-generated
Intercomparison exercise code	Code of the ICES or Quasimeme inter-calibration/inter-comparison exercises to provide reference to the report.	List		Z1 *List ://www.ices.dk/env/repfor/viccoi.html* *//www.ices.dk/env/repfor/viccoq.html*	Required
Assigned value	Value assigned to the sample analysed.	Float		1.02	
Robust mean	The robust mean as obtained in the Quasimeme programme for the particular inter-calibration exercise.	Float		0.97	Required
Z score	Z score obtained in the Quasimeme programme for the particular inter-calibration exercise.	Float		-0.47	
P score	P score obtained in the Quasimeme programme for the particular inter-calibration exercise.	Float		0.49	
Intercalibration basis	The type of ecosystem or artificial basis subject to analysis.	List		*List (water) :* – *estuarine water* – *seawater* – *standard solutions*	Required
Start date	The start date of the exercise.	Date	6 : mmyyyy	06/1997	Required
End Date	The end date of the exercise.	Date	6 : mmyyyy	10/1997	Required
@ **Analysis method**	One link to an analysis method.	Link			Required
@ **Parameter**	One link to a parameter.	Link			Required
@ **Service**	One link to the service who realised the exercise of quality analysis.	Link			Required

DATABASE: IDOD **ENTITY : QUALITY ANALYSIS – CONTROL CHART**

DESCRIPTION :

The QUALITY ANALYSIS – CONTROL CHART entity describes the information concerning the control chart kept by a service for a particular parameter determined with a certain analysis method.

ATTRIBUTES :

NAME	EXPLANATION	TYPE	FORMAT		COMMENT
Control chart code	A code to identify one occurrence of the QUALITY ANALYSIS – CONTROL CHART entity in the database.	Key			Required, unique, system-generated
Control chart basis	This describes the basis of repeated measurements on which the laboratory control chart is constructed	List		*List :* *(a) internationally available Certified Reference Material (CRM)* *(b) internal reference material* (c) *internal reference material tested against a CRM*	Required
CRM code	Code for internationally available Certified Reference Material used.	List		SD-N-2 *List : //www.ices.dk/env/repfor/vcrmco.html*	Required if control chart basis is (a) or (c)
Certified concentration	Concentration of CRM	Float		11 E-6	Required if control chart basis is (a)
Description	Description of internal reference material	Memo		MERCK- control standards with known concentrations.	Required if control chart basis is (c)
Control chart mean value	Mean value of the results of the analyses of the reference material used to construct the control chart.	Float		120 E-7	Required
Control chart standard deviation	Control chart reference material: standard deviation.	Float		275 E-7	Required
Number of measurements	Number of measurements of the reference material extracted from the control chart and used to calculate the mean value.	Integer		26	Required
Period	Period over which the measurements of the reference material extracted from the control chart and used to calculate the mean value were made.	Integer	Weeks	52	Required
Start date	The start date of the measurements for the construction of the control chart.	Date	8 : ddmmyyyy	05/05/1996	Required
@ Analysis method	One link to an analysis method.	Link			Required
@ Parameter	One link to a parameter.	Link			Required
@ Service	One link to the service who realised the exercise of quality analysis.	Link			Required

DATABASE: IDOD **ENTITY : QUALITY ANALYSIS – INTERCALIBRATION**

DESCRIPTION :

The QUALITY ANALYSIS – INTERCALIBRATION entity provides information on intercalibration exercises carried out by two or more services.

ATTRIBUTES :

NAME	EXPLANATION	TYPE	FORMAT		COMMENT
Inter-calibration code	A code to identify one occurrence of the QUALITY ANALYSIS – INTER-CALIBRATION entity in the database.	Key			Required, unique, system-generated
Inter-calibration exercise reference	Reference to the report of inter-calibration exercises other than ICES or QUASIMEME, e.g. bilateral inter-calibration exercise.			Schreurs et al. (1998). The effect of storage by freezing on dissolved nutrients for samples from estuarine waters. Werkdocument RIKZ/IT-98.830A.	Required
Intercalibration basis	The type of ecosystem or artificial basis subject to analysis.	List		*List (water) :* - *estuarine water* - *seawater* - *standard solutions*	Required
Start date	The start date of the inter-calibration exercise.	Date	8 : ddmmyyyy	02/05/1996	
@ **Analysis method**	One link to an analysis method.	Link			Required
@ **Parameter**	links to the parameters analysed.	Link			Required
@ **Service**	Links to the services that realised the exercise of quality analysis.	Link			Required

DATABASE: IDOD **ENTITY : PARAMETER**

DESCRIPTION :

The PARAMETER entity describes the parameters relative to the measured values.

ATTRIBUTES :

NAME	EXPLANATION	TYPE	FORMAT		COMMENT
Parameter code	A code to identify one occurrence of the PARAMETER entity in the database.	Key			Required, unique, system-generated
Name	The name of the parameter.	List		nitrate	Required
ICES parameter code	Parameter/contaminants codes according to ICES.	List	5	NTRA *List ://www.ices.dk/env/repfor/vparam.html*	
Category	The category to which the parameter belongs, e.g. metals, nutrients, etc.	List		Nutrients	Required
Measurement unit	The unit in which the parameter is expressed.	String		µmol/l	Required
Matrix	Actual matrix analysed	List		*List (water) :* - *Total* - *dissolved* - *particulate*	Required
Substrate	Basic sampling medium	List		*List :* - *water* - *sediment* - *biota*	
Formula	The chemical formula of the parameter	String		NO_3^{2-}—N	
@ Values	One or more links to the values of type parameter.	Link			
@ Analysis method	One or more links to the analysis methods used for this parameter.				
@ Quality sample and handling	One or more link to the quality of samples and sample handling.	Link			
@ Quality analyses – quasimeme	One or more links to the quasimeme exercises (quality of analysis) realised by the service.	Link			
@ Quality analyses - control chart	One or more links to the information on control charts (quality of analysis) kept by the service.	Link			
@ Quality analyses - inter-calibration	One or more links to the inter-calibration exercises (quality of analysis) realised by the service.	Link			

DATABASE: IDOD **ENTITY : NON-CONTINUOUS VALUE**

DESCRIPTION :

The NON-CONTINUOUS VALUE entity describes the value obtained after analysis or measurement by a service. The NON-CONTINUOUS VALUE is not part of a series of automated measurements at regular time interval.

ATTRIBUTES :

NAME	EXPLANATION	TYPE	FORMAT		COMMENT
Value code	A code to identify one occurrence of the NON-CONTINUOUS VALUE entity in the database.	Key			Required, unique, system-generated
Value – mantissa	The mantissa of the measured value in scientific notation.	Integer		1342	
Value – exponent	The exponent of the measured value in scientific notation.	Integer		-6	
Latitude	The geographical latitude of the location where the datum was acquired.	Float		51.1833	Required Decimal degrees
Longitude	The geographical longitude of the location where the datum was acquired.	Float		02.475	Required Decimal degrees
Date	The date of the sample acquisition.	Date	8 : ddmmyyyy	31/01/1996	Required
Time	The time of the sample acquisition.	Time	4 : hhmm	12.05	Required
Sampling depth	The depth at which the analysed sample was collected.	Float		-3	Required Meters
Replicate number	The number of replicate measurements.	Integer		1 (Default value)	Required
Quality flag	Qualifying character to indicate when reported value should be qualified as *less than*, or *greater than*.	List		*List :* - *less than* - *greater than* - *equal (default)*	
Validity flag	Validation code of measured value by data provider or manager	List		*List :* - *acceptable* - *suspect* - *replaced* - *missing* - *questionable*	
Ecosystem code	The code of the ecosystem in which the sample is taken	List		*List* - *estuary : Scheldt* - *Loire* - *…* - *coastal water* - *open sea*	
Comment	Some comments on the measured value or the assigned quality label.	Memo			
QC scheme number	The number of the QC scheme used.	Integer		5	

Classification chart	Refers to the final classification assigned by the program	List		*List :* - *good* - *bad* - *missing*	
A posteriori probability	A posteriori probability of a measurement being good.	Real		0.8	
Backtracking info	The backtracking information allows the identification of the tests for which the value scored badly.	String	30	GBG	
@ **Subsample**	One link to the subsample used for the measurement.	Link			Required
@ **Analysis method**	One link to the analysis method used to obtain the value.	Link			Required
@ **Service**	One link to the service that operates the measurement.	Link			Required
@ **Parameter**	One link to the parameter that describes the value.	Link			Required
@ **Station**	A link to the station where the sample was collected.	Link			Optional

DATABASE: IDOD **ENTITY : QUALITY SAMPLE & HANDLING**

DESCRIPTION :

The QUALITY SAMPLE & HANDLING entity describes the quality of sampling and sample handling

ATTRIBUTES :

NAME	EXPLANATION	TYPE	FORMAT		COMMENT
Quality sample and handling code	A code to identify one occurrence of the QUALITY SAMPLE & HANDLING entity in the database.	Key			Required, unique, system-generated
Quality sample and handling reference	Reference to a report describing the results of the quality of sample and sample handling	Memo			
@ **Sample handling**	One link to the sample handling involved in the quality of sampling and sample handling procedure.	Link			Required
@ **Sampling method**	One link to the sampling method involved in the quality of sampling and sample handling procedure.	Link			Required
@ **Service**	One link to the service that operates the sampling and sample handling.	Link			Required
@ **Parameter**	One link to the parameter analysed during the quality of sampling and sample handling exercise.	Link			Required

DATABASE: IDOD **ENTITY : PLATFORM**

DESCRIPTION :

The PLATFORM entity describes the physical basis used for measurements (e.g. Belgica, buoy).

ATTRIBUTES :

NAME	EXPLANATION	TYPE	FORMAT	EXAMPLE	COMMENT
Platform code	A code to identify one occurrence of the PLATFORM entity in the database.	Key			Required, unique, system-generated
Name	The name of the platform.	String	50	rv Belgica	
Type	The type of the platform	List		Research ship	Required
Description	Description of the equipment installed on the platform.	Memo		Oceanographic winches: number: 2 Winch for bottom sampling, …	
Call sign	The international radio call sign.	String		ORGQ	
IOC ship codes	Ship code according to the NODC/IOC.	List		BE *List: //ftp.ices.dk/dist/ocean/codes/ship98.txt*	Required if type="ship"
Platform picture	A picture of the platform.	Image			
@ Campaigns	One or more links to the campaigns that use the platform.	Link			
@ Responsible	One link to the service who is responsible of the platform.	Link			Required

DATABASE: IDOD **ENTITY : STATION**

DESCRIPTION :

The STATION entity describes a geographical location sampled regularly.

ATTRIBUTES :

NAME	EXPLANATION	TYPE	FORMAT	EXAMPLE	COMMENT
Station code	A code to identify one occurrence of the STATION entity in the database.	Key			Required, unique, system-generated
Name	The name of the station.	List	10	330a	Required
Start date	Date of first sampling	Date	8 : ddmmyyyy	21/11/85	
End date	Date of last sampling	Date	8 : ddmmyyyy	07/12/95	
Reference latitude	The latitude of the station location.	Float		51.43333	Required Decimal degrees
Reference longitude	The longitude of the station location.	Float		2.80833	Required Decimal degrees
@ Values	One or more links to the non-continuous values collected at this station.	Link			

IDOD Newsletter

Issue 2 March 1999

Foreword

«*What's the average temperature near the Westhinder during the winter season?*»

«*We are currently launching a project to map the marine resources in the Dover Strait. Could you provide us with useful data?*»

«*For my PhD thesis I need salinity plots in the Scheldt estuary. Do you have that information?*»

These are the kind of questions the IDOD team is regularly receiving. In Section 2 of this second *IDOD Newsletter*, the basic queries to the database are described. As they form the foundation of our service to our «customers» we really would appreciate your remarks.

The other Sections deal with the database design, the analysis of data, and the international and national co-operation. They are intended to give our readers some insight into the progresses made in the IDOD project. Feel free to send us your comments on these topics too!

The IDOD team

MUMM
MANAGEMENT UNIT OF THE NORTH SEA
MATHEMATICAL MODELS

UCS
Katholieke **U**niversiteit **L**euven
University **C**enter of **S**tatistics

Laboratoire SURFACES
Université de Liège

INSIDE THIS ISSUE

1. The database prototype

by MUMM & SURFACES

Since the first issue of the *IDOD Newsletter* (November 1998), a prototype of the IDOD database was completed. At present time, it only includes the parameters for the quality of seawater measured in 1996 in the frame of the monitoring programme. These data result from non-continuous measurements. The prototype will shortly be extended with the parameters relative to sediments and biota.

The IDOD database prototype is currently implemented in *Microsoft Access97* but the choice of the final DBMS (Database Management System) is not yet fixed. A benchmark for the purchase of a CASE tool, a database design tool, is presently in progress. With such a tool implementing all the different data models from the conceptual to the physical levels, time will be saved when modifications will be required and applied to the database scheme or structure.

The conceptual model of the database and the data dictionary included in the previous Newsletter were modified to satisfy the database normalisation rules. The conceptual data model, on which the present prototype is based, now satisfies the three first normal forms. The physical level was obtained after transformation of the conceptual data model into the logical data model and then by implementing it into the DBMS as described in Figure 1.1.

Feel free to send your comments to :

IDOD@mumm.ac.be

For more information,
consult the IDOD web site at

http://idod.mumm.ac.be/

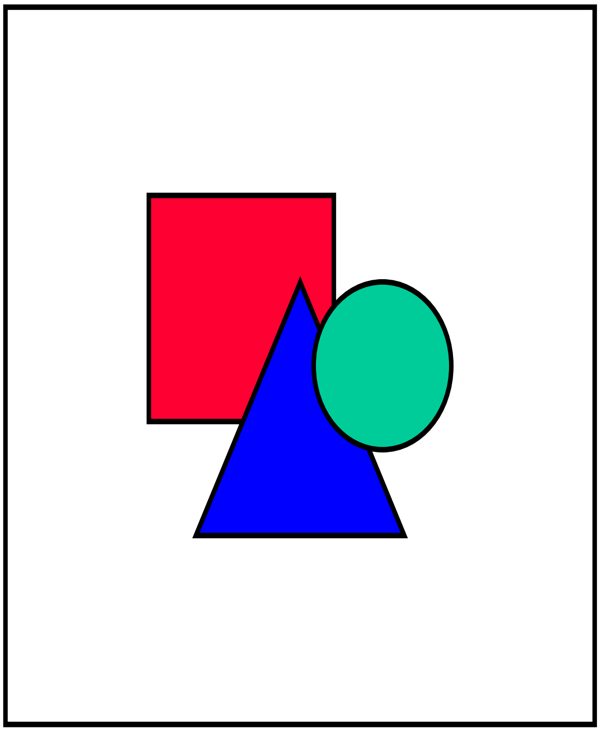


Figure 1.1 – Steps to build the database.

The corresponding data dictionary was adapted with respect to the changes required in the conceptual datamodel. Some entities were split into two or more new entities. In the logical datamodel, new tables were introduced to establish the many-to-many relations.

People interested in a complete description of the database datamodel are invited to visit the website of the IDOD project where the *User's Guide* of the database and the corresponding *Data dictionary* are available in PDF format :

http://idod.mumm.ac.be/documents.html ♦

The IDOD project ("Integrated and Dynamical Oceanographic Data Management") is funded by the Office for Scientific, Technical and Cultural Affairs in the frame of the programme "Sustainable management of the North Sea".

OSTC

2. Requests to the IDOD database

by UCS & MUMM

A wide range of users will request information stored in the database, and accordingly the type of question will differ from user to user. In order to optimise the database functionality, it is important to have a good idea of the possible data requests. Therefore we present a (temporary) list of selection criteria that will be foreseen for the extraction of data and a brief overview of the possible treatments on this dataset.

We would like to know if those criteria cover your general needs. If you think of some other, possibly more specific, query that would be interesting for your work, please let us know. Besides a selection on certain items (like place, time, parameter and campaign), some calculations, like averaging and retrieval of maximum values, or other operations including ways of visualisation, could be important for you.

Note : The examples are given for data measured in seawater. Of course the same type of general criteria will be foreseen for biota and sediment. Your typical requests related to those substrates are also welcome.

Which criteria can we use to extract data ?

- PROJECT and TIME PERIOD
  e.g. MUMM monitoring data for 1996
- a CRUISE and a PLATFORM
  e.g. values collected during BELGICA cruise 97/3
- ANALYSIS METHOD and PARAMETER
  e.g. values for chlorophyll measured with the method of Lorenzen
- ANALYTICAL LABORATORY
  e.g. values measured by MUMM
- PARAMETERS and/or PARAMETER CATEGORIES, SEASON and TIME PERIOD, with extra criteria (SAMPLING DEPTH and/or RESTRICTIONS on the values) :
  - at a geographical location
    e.g. Nutrients and general inorganic parameters at station 330 measured during spring and autumn 1995-1997.
  - within a geographical area
    e.g. Organic contaminants in seawater of the Westerscheldt during the last 5 years.
  - at a given distance

e.g. Nutrients, pigments and phytoplankton in seawater within 2 km of the Belgian coast measured in autumn in the period 1980-1997.

- k-nearest measurements
  e.g. 10 nearest measurements of nitrate within a 6-days period relative to the value obtained at station 120 on 15/5/96

Besides the retrieval of values, non-numerical information like inventories can be extracted, for instance :

- the frequency of sampling of certain parameters or certain sampling stations, or
- the list of parameters measured by service and year

Which operations can be performed on the extracted data ?

- BASIC CALCULATIONS

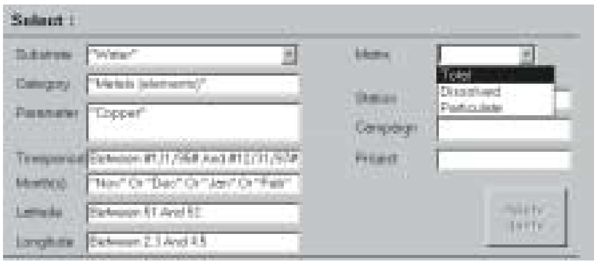
e.g. Monthly or seasonal averages of temperature and salinity for given stations since 1977;
e.g. Seasonal or yearly overviews with averages, minimum and maximum values and number of measurements of suspended solids in the entrance routes to a harbour;
e.g. 30 minutes averaged data for navigation and meteorological parameters for a given campaign;
e.g. Monthly averages of temperature, suspended particles and mean water depth in squares of 5'.

- STATISTICAL ANALYSIS AND GRAPHICAL REPRESENTATION

e.g. regression and graphical representation of total inorganic nitrogen versus salinity;
e.g. time series plots.

- SPATIAL ANALYSIS

e.g. spatial interpolation and cartographic visualisation .



Select :

Substance "Water"
Category "Metals (elements)"
Parameter "Copper"
Temporal Between #1/1/1968 And #12/31/93#
Month(s) "Nov" Or "Dec" Or "Jan" Or "Feb"
Latitude Between 51 And 52
Longitude Between 2.3 And 4.5
Matrix
Status
Campaign
Project

Figure 2.1 : A simple interface for the queries to the database.

3. Statistical analysis of the 1977–1996 monitoring data for seawater

by UCS

Introduction

As explained in the first IDOD newsletter the quality control will compromise a series of tests in which the consistency of new measurements is checked against the statistical characteristics that are exhibited by the measurements that are already present in the IDOD database. To determine the type of tests that could be used, UCS has performed a statistical analysis of the contaminant data that have been gathered from 1977 until 1996 as part of the Belgian monitoring program. The results of this statistical analysis and the conclusions that were reached with respect to the quality control are summarised in the following paragraph.

For the purpose of this study, only those variables have been retained that have been frequently measured over the monitoring period of 19 years. The retained variables include for instance measurements of ammonium, chlorophyll, dissolved oxygen, nitrate, phosphate, salinity, silicate, suspended solids as well as a range of metal concentrations (cadmium, copper, mercury, lead and zinc). The different variables have been sampled (with different degrees of completeness) at the fixed locations shown in Figures 3.1a and 3.1b.

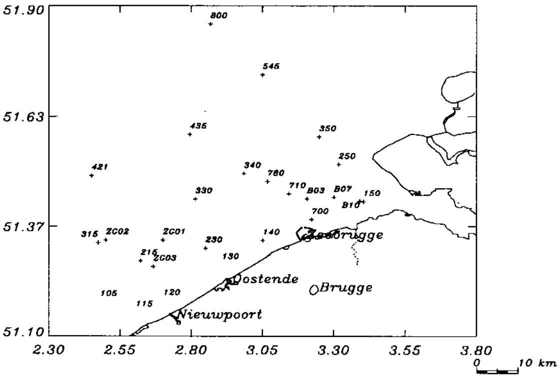


Figure 3.1a – Indicative map of the sampling stations off the Belgian coast.

In the study three statistical characteristics have been examined: 1. the univariate distribution, 2. the spatial continuity of the data, and 3. the correlation between different variables.

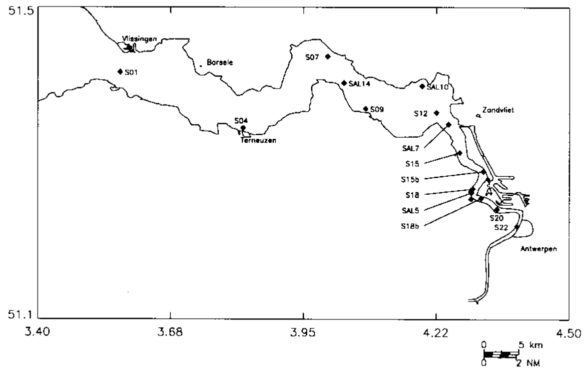


Figure 3.1b – Indicative map of the sampling stations in the Scheldt estuary.

In this newsletter, the results of the study of the univariate distribution of the variables are discussed. In the next newsletter, the conclusions reached from a study of the spatial continuity of the data and the correlation between different variables will be detailed.

Univariate distribution

A first, straightforward technique to validate a measurement compares the measured value against the distribution of the different measurements that have been previously obtained. Doing so, involves two sub-problems: 1. fitting an appropriate distribution, 2. allowing for dependency of the distribution parameters (*i.e.* mean value and standard deviation) as a function of location and time.

The problems associated with such an analysis are illustrated here for the case of the ammonium (AMON) measurements.

Figure 3.2 shows the distribution of the ammonium measurements without accounting for dependency on time or location. The plots at the top refer to the original measured value, while the plots at the bottom refer to the log-transformed values. At the left side, the histogram is shown, while at the right side a normal Q-Q plot is shown. In the latter figure the ranked measurements are plotted against the corresponding quantile value under the assumption that the data follow a normal distribution. If this assumption is satisfied then a straight line should be found.

It is immediately clear that in the case of AMON, a log-transformation is absolutely necessary. However, even with such a transformation the assumption of normality does not appear to be satisfied, in particular for the observations with a lower value.

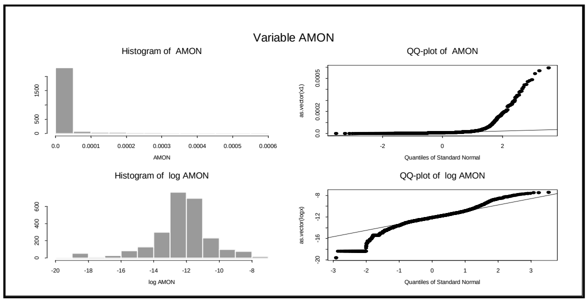


Figure 3.2 - Histogram and Q-Q plot for Ammonium before and after transformation

The histogram for log(AMON) shows that some of the data are characterised by a very small value. Presumably these data correspond to measurement values that fall below the threshold value of the measurement technique. Alternatively, but less plausibly, the data may correspond to particular locations (or times of measurements) where the value is lower than usual.

If the data show a significant spatial or temporal trend, then the univariate distribution of all data will be very wide and will be of little use to verify the data. In such a case, the dependence of the distribution parameters on location and time should be estimated before modelling the distribution of the residuals.

Figure 3.3 shows the result is obtained for the dependence of the average value of the ammonium measurement (after log transformation) on location and time. The distance along the x-axis is measured relative to station S01, located in the mouth of the river Scheldt. The value is defined to be negative for measurements located in the river Scheldt and correspond in this case to the distance measured along the river banks. Clearly, the ammonium value decreases rapidly going from Antwerp to the mouth of the river Scheldt and then continues to decrease as one goes further into the open sea, but this decrease has a smaller gradient. The seasonal influence is accounted for in this analysis assuming an approximate linear trend with time of the year and appears to be of less importance. Using a non-parametric or cyclic seasonal effect might however lead to a more substantial seasonal influence.

The model that is shown in Figure 3.3 and its estimate are obtained using a multiple regression analysis that allows to both include categorical and continuous explanatory variables and allows to specify interaction effects (for instance, the model includes a categorical “SEA” variable that indicates whether the measurement is taken in the open sea or in the river Scheldt; in the model the dependence on the continuous “DIST” variable is then declared to interact with the “SEA” variable). The definition of the possible explanatory location and time variables and the choice of their interaction terms used in this study is merely an exploratory one and in a final analysis other definitions may need to be considered (*i.e.* instead of assuming a linear trend during the year, it may be more appropriate to consider a season-by-season variation; one may wish to consider the possibility that the dependence with DISTANCE in the river Scheldt is a function of the SEASON variable; other spatial location variables may need to be considered; *etc.*).

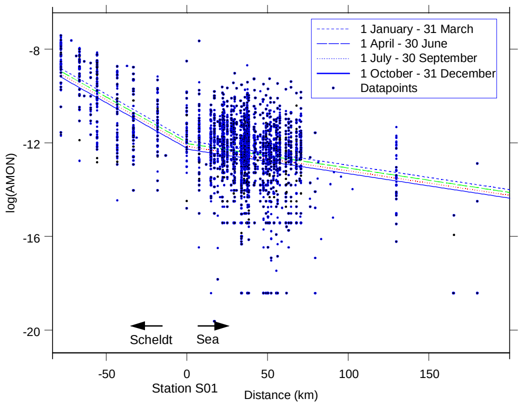


Figure 3.3 – Mean value of log-ammonium as a function of distance to station S01 and for different seasons.

Regardless of the definition of the model, one should consider when estimating the model that not each of the possible dependencies needs to be significant. This can be accounted for, as is done in this study, by using a stepwise regression method which includes only additional terms when the goodness-of-fit improvement is a significant one.

The same method has been applied also to the other variables and it is found that, as is the case for the AMON variable, the distance variable is in practically all cases retained as significant indicating that there is a significant spatial effect. The automatic choice of the best model is however not always found to be satisfactory, presumably because of the particular choice of the goodness-of-fit statistic used in the stepwise method (in Mallows C_p statistic is used and, in some cases, terms are retained of which the significance is debatable). Investigation of detailed results furthermore shows that in some cases not only the mean value but also the standard deviation may vary as a function of location and time. Finally, after fitting such a model, the normality or log-normality of the residuals should again be investigated through the use of Q-Q plots.

Based on this exploratory analysis, UCS plans to implement into the SAT program a more general estimation technique for the modelling of the univariate distribution that would include following functionalities:

- optional transformation of the original variables in order to obtain a more Gaussian distributed dataset;
- optional declaration of a lower threshold value such that values below this threshold value are considered as indications that the measured value is less than the threshold value, rather than actual measurements;
- the possibility to include in a model both categorical and/or continuous spatial and time indicators and the possibility to declare interaction terms, such that the variation with one variable differs depending on the value of another variable (interaction terms);
- automatic selection and estimation of the “best” model accounting for the significance of the different terms. For this purpose, the stepwise linear regression will be used, but further study is required to select a good criterion for the selection of the “best” model;
- the possibility to fit, as is done for the mean value, a spatially and temporally varying model to the squared residuals to account for the variation of the standard deviation with time and spatial location. This procedure will be nearly identical to the previous one, except that the non-normality of the data must be explicitly accounted for (since the squared residuals are chi-squared, rather than normally distributed). This problem will be

solved through the application of general linear models, a technique that will be further detailed when discussing the problem of modelling the spatial continuity of the data in the second part of this publication. ♦

4. Sea Search

by MUMM

Introduction

MUMM is participating in the Sea Search initiative: a European co-operative network of 14 Data Centres working together to improve the access to and the management of marine data and information in Europe. This co-operation, known as Sea Search, is building on the experience of the MAST Data Management Committee. The Sea Search initiative started in August 1998 and is undertaken as a concerted action within the EC-MAST programme.

The Sea Search website

A Gateway to European Marine Data, Information, Products and Services...

The Sea Search website, www.sea-search.net, has been set up and will be further developed to provide an effective navigation tool to oceanographic data and information sources in Europe, and to European centres with expertise and skills in oceanographic and marine data & information management.

The aim is to support researchers, engineers and other users of marine data to identify and take benefit of available data resources. The access to oceanographic and marine data in Europe will be improved by :

- Online tracking of Datasets held within Europe (EDMED catalogue)

At present EDMED covers *ca.* 2300 datasets from *ca.* 500 data holding centres. Activities are underway by the 16 European partners to update the EDMED directory and to develop an innovative infrastructure for updating, searching and presenting the EDMED database by means of the Internet. Also an interlink will be made with the Information Service for Earth Observation data (INFEO) of the Centre of Earth Observation (CEO). Furthermore co-operation is underway with the IOC-IODE to harmonise formats in order to establish a uniform global metadata format, that will be the basis for further EDMED developments.

- Online tracking of cruises carried out by European institutes – Cruise Summary Reports (ROSCOPS)

ROSCOP (Reports of Oceanographic Cruises and Oceanographic Programmes) was conceived by the Intergovernmental Oceanographic Commission (IOC) of UNESCO in the late 1960s in order to provide a low level inventory for tracking oceanographic data collected on Research Vessels. Nowadays, most marine disciplines are represented in ROSCOP. Within the Sea Search initiative, activities are ongoing to install an infrastructure for updating, searching and presenting the ROSCOP database by means of the Internet. Also co-operation is underway with the ICES Marine Data Management group to evaluate the ROSCOP format and to establish an extended metadata format, that will be adopted by all partners.

- Online tracking of European research projects

Within the Sea-Search initiative the idea of a European compilation of Marine Research Projects descriptions has been adopted and activities are underway to develop an infrastructure for updating, searching and presenting the Research Projects database by means of the Internet. Also co-operation is underway with the EC-MAST secretariat to include all MAST Projects in the database. The Research Projects database will give very useful information on ongoing projects, data collection activities, involved organisations and scientists, and resulting products (datasets, publications, knowledge, ..). Useful features of the database are *e.g.* that also coastal and estuarine research projects are identified, including their data collection activities, and that grey literature is identified.

Other objectives are :

- to support and stimulate researchers improving their data management
- to provide researchers with information on expert data management organisations in Europe

- to stimulate co-operation between researchers and expert data management organisations in Europe

Sea Search framework

Keys to the success of community research projects and to an effective support of international environmental policy and marine economic activities are the speed and the ease with which users can locate and get access to oceanographic and marine data & information. Further these data must carry a certain overall data quality level.

Data and Information Management plays a vital role in achieving these goals. It assists science, it safeguards scientific data for future use by a wider community and moreover it enables to combine these scientific data with other available oceanographic and marine data resources for a wide range of applications.

Given the large and still expanding size of diversity in of data types, organisations engaged in data acquisition, volume of data acquired, and offer of computer technologies for data processing, storing, retrieving and distribution, Data and Information Management has become a professional discipline and service. It requires Data Management organisations or units that collaborate with and give service to scientists and that are skilled in processing, quality controlling, archiving, producing added-value dataproducts and disseminating data & dataproducts.

A closer collaboration on a European scale is essential to achieve a more cost effective approach to ocean and marine data and information management and to fulfil the growing demand for data & information from scientists and other communities.

The Sea Search partners

The Sea Search website, services and infrastructure are developed and operated by 15 institutes/centres from 14 different European coastal states and EU.

Each of these centres is operating in their country as national data centre or focal point for oceanographic and marine data & information. They are representative nodes in their countries with links to other organisations active in marine research and marine environmental management, thus monitoring and overseeing national marine research activities and data flows. The partnership is complemented with the Space Applications Institute (SAI/ME) of the CEC Joint Research Centre that is highly skilled in management and processing of marine remote sensing data and that is especially promoting integrated approaches combining the use of RS data, in situ measurements and modelling. It is the objective of Sea Search to achieve a further integration of marine RS data and in situ data.

Belgium : Management Unit of the Mathematical Models of the North Sea (MUMM)

European Commission : CEC-Joint Research Centre - Space Applications Institute (CEC-JRC-SAI/ME)

Finland : Finnish Institute of Marine Research (FIMR)

France : Systèmes d'Informations Scientifiques pour la Mer (IFREMER-SISMER)

Germany : Deutsches Ozeanographisches Datenzentrum (BSH-DOD)

Greece : Hellenic National Oceanographic Data Centre (HNODC-NCMR)

Iceland : Marine Research Institute (MRI)

Ireland : Irish Marine Data Centre - Marine Institute

Italy : Marine Environmental Research Centre (ENEA-CRAM)

The Netherlands : Marine Information Service (MARIS) (Sea Search Co-ordinator)

Norway : Norwegian Marine Data Centre - Institute of Marine Research (IMR)

Portugal : Instituto Hidrográfico (IHPT)

Spain : Instituto Español de Oceanografia (IEO)

Sweden : Swedish Meteorological and Hydrological Institute (SMHI)

United Kingdom : British Oceanographic Data Centre (BODC) ♦

The *IDOD* Newsletter – Issue 2
has been set up by
K. De Cauwer & S. Scory

To assure an efficient exchange of information between IDOD and the other projects of the programme “Sustainable Development of the North Sea”, Project Data Managers were designated within each of the research projects. The IDOD team will be in contact with the persons listed below for the practical aspects of data exchange.

Furthermore, additional information on specific data types, *e.g.* methodology, are of the utmost importance to verify the conceptual model of the database. PDM's are expected to be able to provide IDOD with such information.

We feel it useful to list these people here, as they will be concerned with the same type of questions and could find it useful to directly exchange information between them.

The structural and functional biodiversity of North Sea ecosystems – Species and their habitats as indicators for the sustainable management of the Belgian coastal shelf

Pr. M. Vincx, Pr. E. Kuijken, Pr. F. Ollevier

Project PDM : *André Cattrijsse*

UG - Marine Biology Section
tel. +32–9–264.52.30
fax +32–9–264.53.44
Andre.Cattrijsse@rug.ac.be

Birds and marine mammals of the North Sea: Pathology and Ecotoxicology

Pr. J.–M. Bouquegneau, Pr. Fr. Coignoul,
Pr. Cl. Joiris, Pr. P. Meire

Project PDM : *Virginie Debacker*

ULg - Laboratoire d'Océanologie
tel. +32–4–366.48.44
fax +32–4–366.33.25
Virginie.Debacker@ulg.ac.be

AMORE Advanced MOdelling and Research on Eutrophication

Pr. Ch. Lancelot, Dr. M. Tackx, Dr. K. Ruddick

Project PDM : *Véronique Rousseau*
ULB - GMMA
tel. +32–2–650.59.90
fax +32–2–650.59.93
vrousso@ulb.ac.be

The biogeochemistry of nutrients, metals and organic micropollutants in the North Sea

Pr. R. van Grieken, Pr. W. Baeyens,
Pr. H. Van Langenhove, Pr. R. Wollast

Coordinator, Sub–project PDM : *Kurt Eyckmans*

UIA - Centrum voor Micro- en Sporenanalyse
tel. +32–3–820.23.57
fax +32-3-820.23.76
eyckmans@uia.ua.ac.be

Satellite 1, Sub–project PDM : *Koen Parmentier*

VUB - Laboratorium voor Analytische Chemie
tel. +32–2–629.32.66
fax +32–2–629.32.74
kparment@vub.ac.be

Satellite 2, Sub–project PDM : *Tom Huybrechts*

UG - Laboratorium voor Organische Scheikunde
tel. +32–9–264.59.98
fax +32–9–264.62.43
Tom.Huybrechts@rug.ac.be

Satellite 3, Sub–project PDM : *Nathalie Roevros*

ULB - Laboratoire d'Océanographie Chimique
tel. +32–2–650.52.33
fax +32–2–646.34.92
natroev@ulb.ac.be

ICAS The Impact on North Sea organisms of pollutants Associated with Sediments

Dr. Ph. Dubois, Pr. M. Jangoux,
Pr. R. Flammang

Project PDM : *Pol Gosselin*

Laboratoire de Biologie marine
Université de Mons-Hainaut
Pol.Gosselin@umh.ac.be

MARE-DASM MArine REsources Damage Assessment and Sustainable Management of the North Sea

Pr. E. Somers, Pr. C. Janssen, Dr. G. Pichot, Pr. H. Bocken

Project PDM : *Serge Scory*

MUMM
tel. +32–2–773.21.11
fax +32–2–770.69.72
Serge Scory@mumm.ac.be

IDOD Newsletter

Issue 3 December 1999

Foreword

The third issue of a newsletter is always the most difficult to bring into reality. We did it!

We hope you will enjoy this new delivery of fresh news about the IDOD project : *How evolves the database itself ? What do the data mean ? How can they be analysed in their spatial dimensions ? At what rate are new datasets received ? ...*

But, first of all, we wouldn't like to miss the opportunity to wish you and your colleagues a happy new year!

The IDOD team

MUMM
MANAGEMENT UNIT OF THE NORTH SEA
MATHEMATICAL MODELS

Katholieke **U**niversiteit **L**euven
University **C**enter of **S**tatistics

Laboratoire SURFACES
Université de Liège

INSIDE THIS ISSUE

Feel free to send your comments to :

IDOD@mumm.ac.be

For more information,
consult the IDOD web site at

http://idod.mumm.ac.be/

1. Changes to the structure of the IDOD–prototype

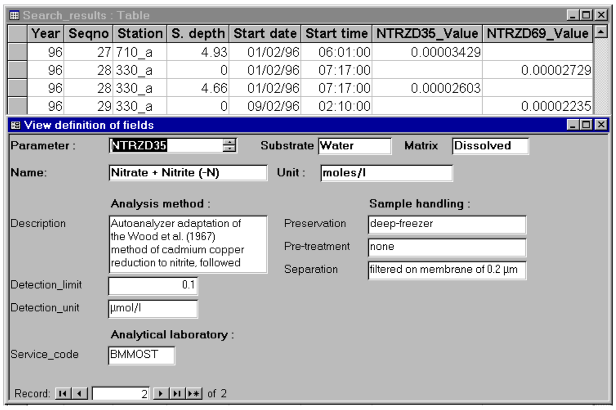
by MUMM

Since the implementation of the prototype, the database structure (see a description of the previous version in the *IDOD Newsletter #1*) has been changed to solve some problems encountered during data entry and retrieval.

Parameter and methodology

One of the major changes is the way information on parameters and methodologies is organised. During one sampling event, the same parameter is often measured by different services and/or analysed with different methods. A clear display and correct statistical analysis of these results appeared to be very complicated. More particularly, it was difficult to distinguish clearly values obtained with completely different methods. As certain services measure a parameter for monitoring purposes with a precise, quality-controlled analytical method, and others measure that parameter for relative comparison with other data, the users should only mix these data in further analysis with care.

This problem was solved by adding a reference of the analytical method and responsible service to the parameter name instead of to the value.



Search_results : Table

Year	Seqno	Station	S. depth	Start date	Start time	NTRZD35_Value	NTRZD69_Value
96	27	710_a	4.93	01/02/96	06:01:00	0.00003429	
96	28	330_a	0	01/02/96	07:17:00		0.00002729
96	28	330_a	4.66	01/02/96	07:17:00	0.00002603	
96	29	330_a	0	09/02/96	02:10:00		0.00002235

View definition of fields

Parameter : NTRZD35 Substrate Water Matrix Dissolved

Name: Nitrate + Nitrite (-N) Unit : moles/l

Analysis method :

Description: Autoanalyzer adaptation of the Wood et al. (1967) method of cadmium copper reduction to nitrite, followed

Detection_limit: 0.1

Detection_unit: µmol/l

Sample handling :

Preservation: deep-freezer

Pre-treatment: none

Separation: filtered on membrane of 0.2 µm

Analytical laboratory :

Service_code: BMMOST

Record: 2 of 2

Figure 1.1 – Example of data display

The data display of the search for the parameter 'Nitrate + Nitrite' is shown as example in Figure 1.1. This involves that the user should always take into consideration the applied methodology before proceeding with the data analysis.

Display and analysis of replicate samples and replicate values

A similar problem for display and statistical analysis occurs with data for one parameter obtained on several samples at one sampling station. Such 'simultaneous' values provide information on sampling reproducibility or the homogeneity of the substrate sampled.

Replicate analyses of a given parameter in a particular sample provide information on the analytical reproducibility.

For both problems it was decided to only display and use the average value in the standard procedures of the statistical analysis. Of course for specific applications, the individual values can still be retrieved.

The sampling event

Other needs for changes became apparent when investigating the data related to seawater and air received so far in the frame of the Programme.

- Sampling position has changed into start and end sampling position, and similarly for sampling time, to enable the storage of data resulting from sampling tracks in the same structure (e.g. collection of air particles, continuous collection of suspended particles).
- The attribute 'exceptional circumstances' is added as a memo-field to store the notes of the laboratory on the sampling conditions.
- Besides the actual sampling depth, a reference sampling depth is included to enable retrieval and analysis based on depth levels *e.g.* "surface", "bottom", "middle", "-3 m". ♦

2. Statistical analysis of the 1977–1996 monitoring data for seawater (Part 2)

by UCS

Introduction

In the previous IDOD newsletter the univariate analysis of the 1977-1996 monitoring data for seawater has been detailed. In this newsletter, the method used to model the correlation between different variables is described. As explained in the first IDOD newsletter, the univariate distribution (previous newsletter), the correlation between the variables (this newsletter) and the quantification of spatial continuity is used to set up quality checks of the data. The tools developed as part of this research will be also incorporated into the SAT program (statistical analysis tool) that will allow users of the IDOD database to examine the data in a similar manner.

Correlation between variables

One would expect that different variables, if measured at the same location and time, are not entirely uncorrelated. This is evidently true for variables that are by definition related (*i.e.* NTRZ=NTRI+NTRA), but it may also apply to other variables that are indirectly related. Correlation of this type, if it exists, is obviously of interest in a quality control, since it would allow to predict the measured value (with a given accuracy) on the basis of other measurements. In this paragraph the extent to which such correlations exist for the 1977-1996 monitoring data is investigated and a method to model these correlations such that they can be used in the quality control is developed.

A-priori one may expect that a measurement variable could be related to each of the other variables or a combination of such variables. For practical reasons, the analysis is limited to consider only relationships of the measurement variable of interest (the response variable) to 1, 2 or 3 of the other measurement variables (the predictor variables). The possible relationships that can be fitted are further constrained by the fact that a sufficient number of simultaneous readings of the potential subset of measurement variables should be available. For the purpose of this study "a sufficient number" has been defined to correspond to approximately 10% of the number of observations of the response variables. Typically, the number of records is either close to 2500 or 1000 records, so that the estimates are based on at least either 250 or 100 records.

For each measurement variable and for each subset of predictor variables that satisfies the previous condition, a linear regression model is fitted using least-squares. The significance of the estimated regression coefficients is then considered (using a sequential t-test at a 5% significance level) and if one of the terms is not significant the model is rejected (the simpler model with the non-significant term removed is then tested further possibly adding or removing other variables). The models that thus remain are then ordered according to decreasing adjusted R^2 goodness-of-fit statistic. This statistic indicates the variance that is explained by the model (a high value corresponds to a good fit), but adjusts this estimate for the number of parameters used to explain that variance (i.e. if the same variance is explained by a model with 1 measurement variable and another model using 3 measurement variables, the first model will be ranked higher according to the adjusted R^2 statistic).

As an example the results that are thus obtained for the AMON variable (ammonium) are listed below.

R²adj	Regressor 1	Regressor 2	Regressor 3
0.734	CPHLL	NTRA	SLCA
0.713	NTRA	PHATL	SLCA
0.576	NTRA	NTRI	SLCA
0.574	NTRA	SLCA	SAL
0.569	NTRA	SLCA	
0.554	NTRI	NTRZ	SLCA
0.540	NTRZ	PHATL	SAL
0.496	CPHLL	SLCA	SAL
0.485	PHATL	SLCA	SUSP
0.481	NTRZ	SUSP	SAL

Ranked list of predictor variables for ammonium

In the case of the AMON variable the 'best' model uses the predictor variables CPHLL, NTRA and SLCA. The adjusted R^2 statistic corresponds to 73%. An alternative model of nearly equal performance uses the predictor variables NTRA, PHATL and SLCA.

In a quality control check, this list can be used as follows: when checking the ammonium value, all simultaneously measured variables are gathered from the database (*e.g.* NTRA, PHATL, SLCA and SAL). Then the regression list is searched to find the highest ranked set of predictor variables that forms a subset of these variables (in this case, NTRA+PHATL+SLCA) and using these variables the expected value and the standard deviation of the AMON measurement can be derived. If the measured ammonium value falls outside the expected value ± 3 standard deviations, then the test would fail and, as explained in the first newsletter, this failure together with the results of the other tests (univariate distribution and spatial continuity) is used to assign a quality label to the datapoint (varying from "suspect" to "normal").

As an illustration of the correlation that exists between the different variables, the table below shows the sets of predictor variables (limited to maximum 3) that best explain a given variable.

Response	R²adj	Regressors		
AMON	0.73	CPHLL	NTRA	SLCA
CD	0.54	CU	DOXY	SUSP
CPHLL	0.98	CPHLC	CU	PHATL
CPHLH	0.88	CPHLC	CPHLL	DEPH
CPHLC	0.99	CPHLL	CU	PHATL
CU	0.34	AMON	CD	CPHLC
DOXY	0.57	CPHLC	TEMP	SAL
HG	0.42	AMON	CPHLL	PHATL
NTRA	0.997	DEPH	NTRI	NTRZ
NTRI	0.62	NTRA	NTRZ	PHATL
NTRZ	0.997	DEPH	NTRA	NTRI
PB	0.49	CD	CPHLC	ZNP
PHATL	0.86	CPHLC	CPHLL	SAL
PHOS	0.66	DOXY	SLCA	SAL
SAL	0.87	DOXY	NTRZ	PHATL
SUSP	0.87	AMON	CPHLL	NTRA
ZN	0.51	DOXY	PHATL	TEMP

The previous results show that in most cases very high values of the adjusted R^2 statistic are obtained. Only for overall metal concentrations

(CD, CU, HG, Pb, ZN) , the goodness-of-fit statistic is relatively low.

The results for NTRA, NTRI and NTRZ are of particular interest since in this case the relationship is known : NTRZ=NTRI+NTRA. The table above shows that indeed a model is selected that includes the correct predictor variables but in addition another variable (for NTRA and NTRZ depth DEPH, for NTRI phaeophytin-a PHATL) is included. Detailed results show that the models without these variables are ranked as second best and have an adjusted R2-statistic that is very close to the optimal. On the other hand, the additional variable included is found to be statistically significant (albeit of little practical importance). Whether these variables may in fact have a physical influence (i.e. depth may indirectly have an influence through the measurement procedure?) remains to be seen. In general, it stresses the fact that when using such lists in a quality control, the composition of such a list should not be entirely automatic, but should be at least verified by an expert user that is knowledgeable about the physo-chemical background of the relationships. ♦

3. Spatial data analysis

by SURFACES

The data already included in the IDOD database prototype were used to produce spatio-maps for different parameters such as temperature and salinity. The spatial representation of oceanographic measurements requires the use of interpolation procedures. Many techniques can be applied to achieve the interpolation : nearest neighbour, inverse distance to a power, triangulation with linear interpolation, kriging, and so on. Unfortunately, the obtained results are often strongly linked to the data type and the applied procedures. For these reasons, particular attention must be paid to the preliminary study of the data set statistics. The most elaborated gridding technique for interpolation is the kriging technique that can produce better results but that requires a good understanding of the data set statistics. The preliminary analysis is done using the variogram or covariogram of the studied data sets. This variogram helps to chose the most appropriated kriging model : exponential, Gaussian, quadratic, rational quadratic, power, linear, spherical, logarithmic, *etc*. models.

The following examples of spatial data analysis use the temperature measurements collected at the sea stations during the campaign BE96/1 (30/01/1996–02/02/1996).

The spatio-map presented in Figure 3.2 is computed with the "inverse distance to power two" interpolation technique and shows the tendency of the method to generate "bull's-eye" patterns of concentric contours around the data points. Moreover, this technique does not extrapolate values beyond the range of data.

The spatio-map presented in figure 3.3 is calculated by kriging. This interpolator involves several steps : exploratory statistical analysis of the data, variogram modelling, then creating surface. The variogram is shown in figure 3.1 where the curve represents the best fit function.

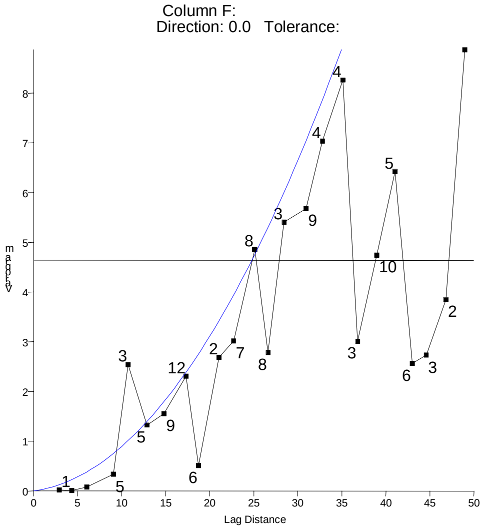


Figure 3.1. Variogram of the sea temperatures.

For all spatio-maps produced by kriging, a complete statistical report can be joined to the map to complete it. Kriging is a complex process and requires a large part of tuning in trying to obtain the best result. Therefore, it is very difficult, or even meaningless, to envisage an implementation that runs it automatically.

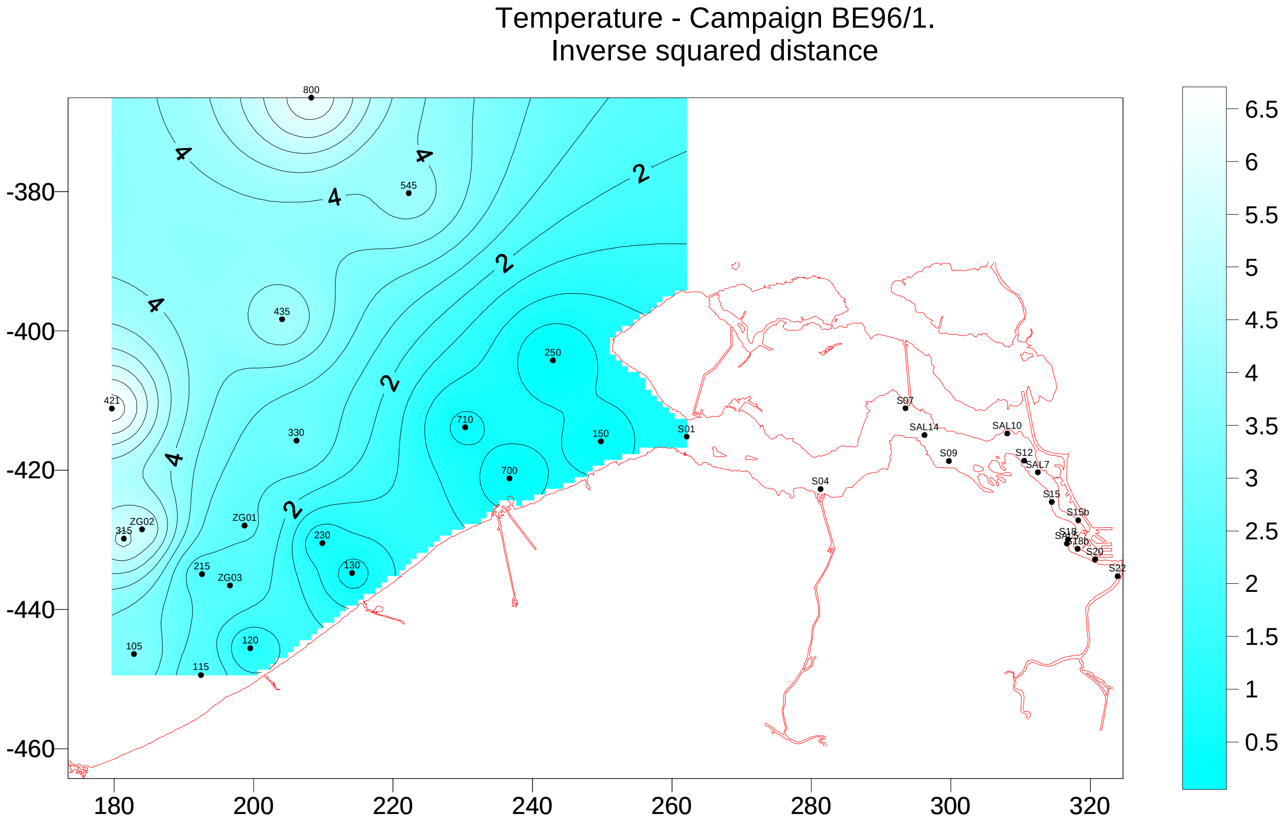


Figure 3.2. Map of sea temperatures by "inverse distance to power two".

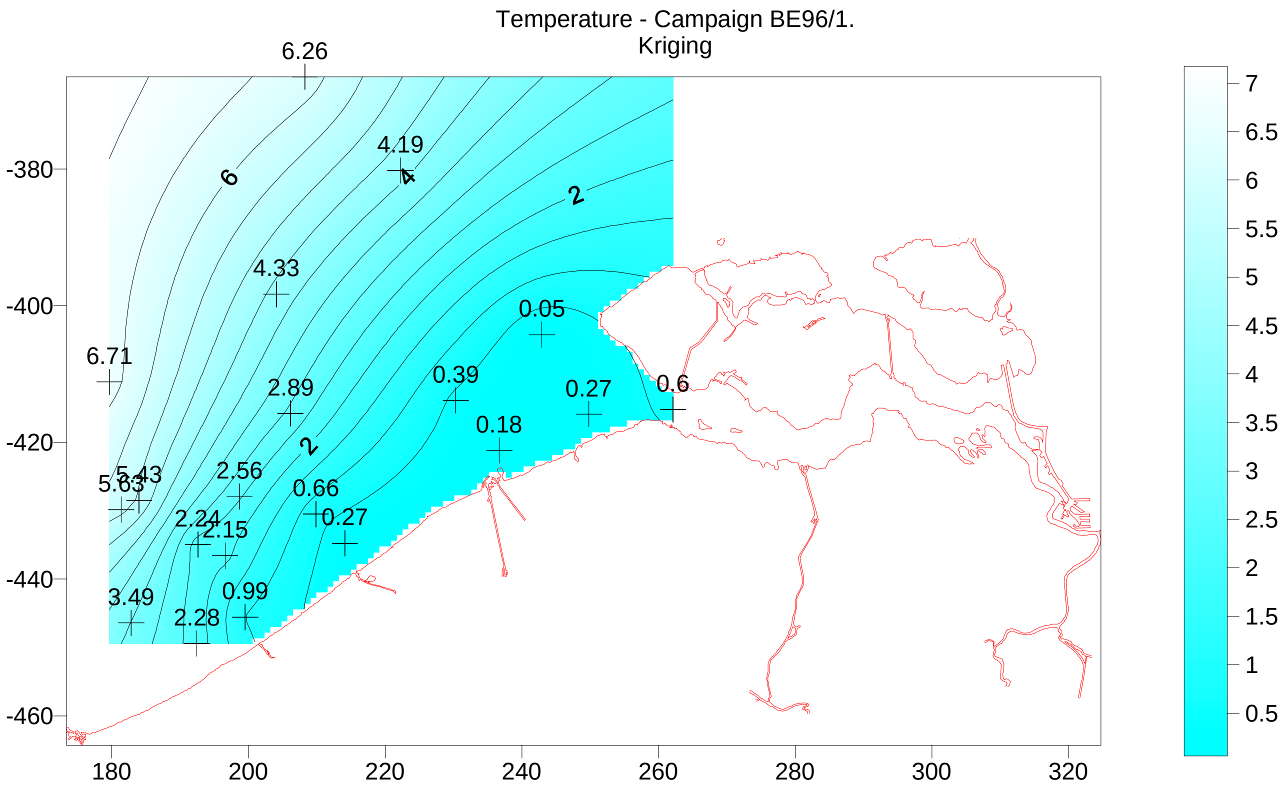


Figure 3.3. Map of sea temperatures by kriging.

Stations proximity

A map of proximity for all the sampling stations was produced by use of Thiesen's polygons (Figure 3.4). These polygons are calculated from a Delaunay's triangulation. In such proximity mapping, each point of the surface area is assigned to the station that is nearest to. This map shows the "influence area" for every station on basis of the Euclidian distance all the way through the sea. On the map, red points locate the sampling stations in the sea and Scheldt estuary. ♦

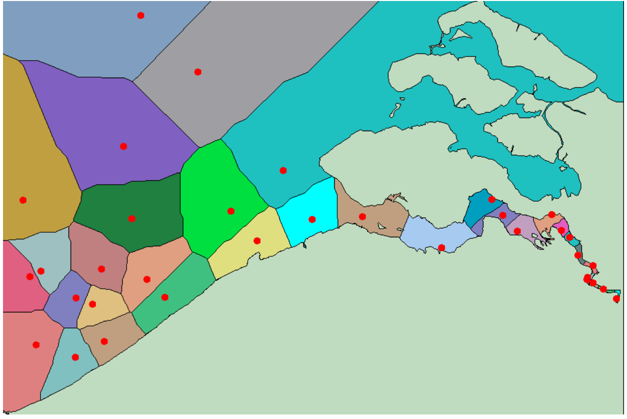


Figure 3.4. Stations proximities map by Thiesen's polygons.

4. Data sets input

by MUMM

The *IDOD* data base is first of all fed by the measurements made in the frame of the Programme "Sustainable Management of the North Sea".

The technical specifications of the research projects mention that the data sets have to be sent to the *IDOD* center "at the end of March following the year where the corresponding samples were taken".

During the start-up phase, it was decided that, by March 1998, all the teams should have sent to *IDOD* a list of the parameters they intend to measure and, by the end of March 1999, the data sets corresponding to campaigns performed in 1997 and 1998.

The actual "input flow rate" is shown on Figure 4.1.

As the figure shows, the input flow rate is not as satisfactory as it could be. Several reasons are identified (lack of awareness about the interest of having the data banked at *IDOD*, delays due to laboratory practices, *etc.*).

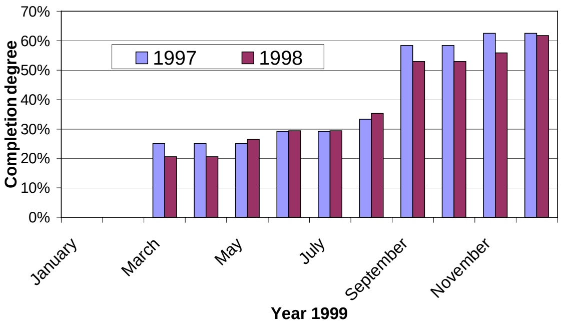


Figure 4.1 : Actual flow rate of data sets to the IDOD centre during the year 1999. 12 data sets were expected for 1997 and 17 for 1998.

The *IDOD* team is currently working to find solutions to improve this flow rate, since any delay in data transfer significantly impacts the agenda of the project. ♦

5. Products

by MUMM

The MUMM monitoring data for 1998, available in the IDOD-database, were used to fulfil some international obligations of Belgium. This resulted in two reports.

National comments

Each year, the results of the monitoring activities of Belgium in the frame of the Joint Assessment and Monitoring Programme (JAMP) of the Oslo and Paris Commission (OSPAR) are reported to the International Council for the Exploration of the Sea (ICES). National commentaries on the monitoring data are provided together with the data submission. These include information on measurements, compliance with analytical guidelines/procedures and information on quality assurance. The information on measurements provides a description of the results, including maps and tables, explains the underlying processes and compares data with results from previous years.

In the Belgian national comments for the monitoring data of 1998, winter loads at the zero-salinity boundary have been assessed for nutrients. Figure 5.1 shows that the nitrogen loading at the freshwater boundary was 471 µmol/l in 1998, which is lower than previous years (500 µmol/l, 610 µmol/l and 580 µmol/l in 1995, 1996, 1997 respectively).

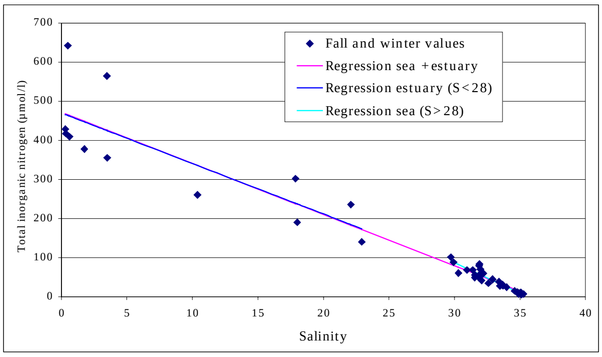


Figure 5.1. Regression of total inorganic nitrogen (Winter 1998)

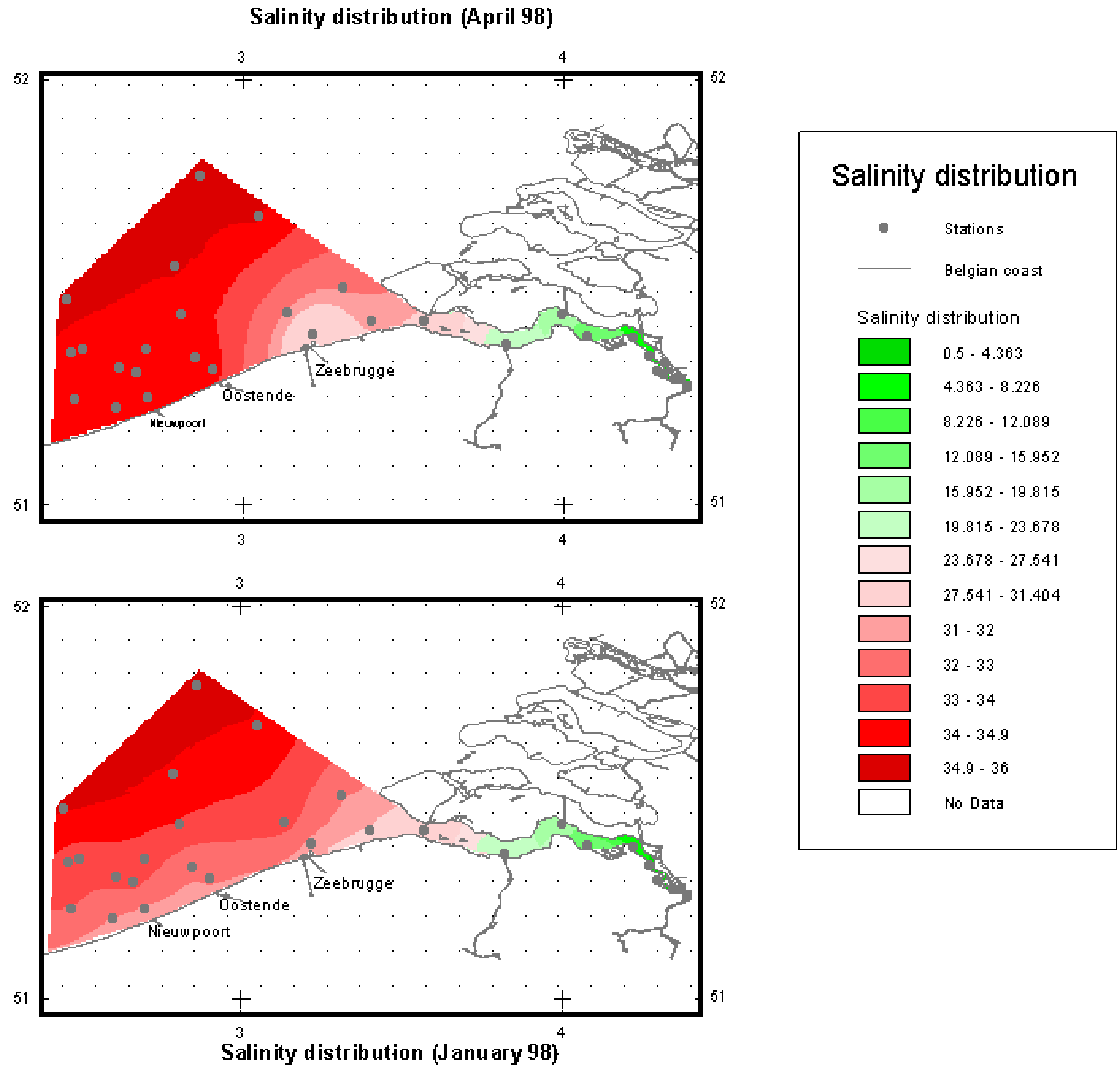


Figure 5.2. Salinity distribution (1998)

The spring phytoplankton bloom in 1998, indicated by high chlorophyll-a concentrations, is well visible at sea and more particularly close to the mouth of the Scheldt estuary. The phytoplankton development at sea began mid-March and reached a maximum in mid-April as shown by a chlorophyll-a concentration of 18 µg/l. This maximum is less than half the maximum concentration measured in 1997 but tends to be close to the years before.

The cartographic software purchased by MUMM, ArcView3.1, was used to produce spatial distribution maps. Figure 5.2 shows the spatial distribution of salinity for winter and spring 1998.

Greater North Sea Quality Status Report

During the last meeting of the Environmental Assessment and Monitoring Committee (ASMO) of OSPAR that took place in Nanterre (29 November – 3 December 1999), the Greater North Sea Quality Status Report (QSR) was approved.

This report presents an assessment of marine environmental conditions or the quality status of the Greater North Sea and of its changes. Representatives from relevant countries within the Greater North Sea region were assigned for the production of the report. The QSR is based upon the most recent information available from national and international sources. These sources include products based upon the results of the JAMP and the ICES database that includes the monitoring data of all contracting parties of OSPAR (see ‘National Comments’).

The report contains a description of the physical geography, hydrography and climate of the Greater North Sea, examines human activities that directly or indirectly impinge on marine areas and summarises existing knowledge on chemical and biological features. In the last chapter, the Overall Assessment, conclusions were drawn to identify where improvements in the quality of the North Sea have been achieved, the major causes of any environmental degradation and recommendations for the managerial and scientific action needed to correct these.

An important conclusion is that generally, significant improvements can be reported in connection with the inputs of heavy metals, oil and the nutrient phosphorus. These improvements are predominantly reflected in the reduced pressure on the marine environment at a local or regional level. At the same time, an increasing number of man-made compounds are being detected in the North Sea for which the ecological effects are largely unknown. The general improvements that have been made until now are reassuring and the OSPAR strategies provide a framework in which measures are developed for continuing these improvements.

6. Short news

by MUMM, UCS & SURFACES

The design of the *IDOD* database for the seawater data was presented during the last CoastGIS'99 symposium at Brest, France (9 to 11 of September 1999). The paper, entitled "Design of an oceanographic database" is available on request at idod@mumm.ac.be. This paper is selected for publication in the CoastGIS'99 book shortly co-edited by Ifremer and SHOM. ♦

The IDOD team prepared an extensive description of the project, to be shown as a "demonstration" during the workshop entitled "À la recherche d'un dialogue durable entre science et politique – Op zoek naar een duurzame dialoog tussen onderzoek en beleid" (Brussels, 24-25 November 1999). This slideshow is published at http://idod.mumm.ac.be/slideshow/index.htm. ♦

The IDOD project ("Integrated and Dynamical Oceanographic Data Management") is funded by the Office for Scientific, Technical and Cultural Affairs in the frame of the programme “Sustainable management of the North Sea”.

OSTC

The *IDOD* Newsletter – Issue 3
has been edited and set up by S. Scory.
Contributions by :
K. De Cauwer, M. Devolder, S. Jans,
F. Muller, B. Plevoets & S. Scory

1. Annex : List of parameters measured in the frame of the Programme

		COORDINATOR	Vincx			Van Grieken				Lancelot			Dubois			Bouquegneau			
		PARAMETER	UG-Vincx	IN-Kuijken	KUL-Ollevier	VUB-Goeyens	UG-V.Langenhove	UIA-Van Grieken	ULB-Wollast	ULB-GMMA	VUB-ECOL	MUMM	ULB-Dubois	UMH-Jangoux	UMH-Flammang	ULg-Bouquegneau	Ulg-Coignoul	VUB-Joiris	IN-Meire
M	**General**																		
E		Date	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
T		Time	x	x	x	x	x	x	x	x	x	x				x			
A		Position	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
D	**Methods**					?		?									?	?	?
A		Sampling method	x	x	x		x	x	x	x	x	x	x	x	x				?
T		Sample handling	x	x	x		x	x	x	x	x	x	x	x	x				?
A		Analysis method	x	x	x		x	x	x	x	x	x	x	x	x	x	x		?
		Statistical analysis	x	x	x								x	x	x				
		Programmes used											x	x	x				
	Quality control							?			?	?					?	?	?
		Quasimeme scores				x	x		x							x			
		Certified references											x		x				
		Internal references								x			x		x				
		Intercalibration				x			x										
	Meteorology					?		?	?										
		PAR										x							
		Wind speed	x				x					x							
		Wind direction	x				x					x							
		Solar radiation	x				x												
		Atmospheric pressure	x				x												
		Temperature	x				x												
W	**Physical**	temperature	x (v)				x			x		x (v,c)							
A		suspended matter							x	x		x							
T		depth	x (v)				x						x	x	x				
E		secchi depth										x							
R		PAR	x (v)																
		profile of the pore							x										
		specific surface							x										
		average pore							x										

	radius																	
Major inorganic	salinity	x (v)				x			x		x (v,c)	x	x	x				
	dissolved oxygen																	
	pH																	
	carbon								?									
Nutrients	nitrate	x			x				x									
	nitrite	x			x				x									
	phosphate	x			x				x									
	silicate				x				x									
	ammonia	x			x				x									
	urea				x													
Metals	Cu				x (s,p)			x (p)										
	Cd				x (s,p)			x (p)										
	Ni				x (s,p)			x (p)										
	Zn				x (s,p)			x (p)										
	Pb				x (s,p)			x (p)										
	Mn							x (p)										
	Co							x (p)										
	Cr							x (p)										
	Al							x (p)										
	Ca							x (p)										
	Fe							x (p)										
	Si							x (p)										
	As				x (s,p)													
	Hg°				x													
	monomethyl Hg				x (s,p)													
	total Hg				x (s,p)													
Major organic	organic nitrogen				x (s,p)													
	total nitrogen							x (p)										
	organic carbon				x (p)			x (p)	?									
	total carbon							x(p)	?									
Pigments	Chlorophyl-a	x						x	x	x	x							
	Chlorophyl-c	x																
	fuccoxanthine	x								x								
	19'hexanoxanthine									x								
	diadinoxanthine									x								
	alloxanthine									x								
	peridinine									x								
	phaeopigment							?			x							

	Aromatic hydrocarbons	benzene					x												
		toluene					x												
		m/p/o-xylene					x												
		ethylbenzene					x												
	Organochlorines	chloroform					x												
		tetrachloromethane					x												
		1,1-dichloroethane					x												
		1,2-dichloroethane					x												
		1,1,1-trichloroethane					x												
		trichloroethylene					x												
		tetrachloroethylene					x												
	Optical	downwelling PAR irradiance										x (v)							
		optical backscatter	x(v)									x (v)							
		scalar PAR irradiance										x (v)							
		upwelling fluoresc. radiance										x (v)							
		PAR attenuation coefficient										x							
		upwelling radiance spectra										x							
		downwelling irradiance spectra										x							
		sub-surface irradiance spectra										x							
		phytoplankton absorpt. spectra										x							
		yellow subst. absorpt. spectra										x							
A	**nutrients**	NO2-, NO3-, SO4-, PO4-						x											
I	**heavy metals**	Ag, Cu, Zn, Fe						x											
R	**major inorganic**	F-, Cl-, Si, sea salt, gypsum, alumino-silicates						x x											
S	**Physical**	granulometry	x										x						
E	**Interstitial water nutrients**	nitrate	x (v)																
D		nitrite	x (v)																
I		ammonia	x (v)																
M		phosphate	x (v)																
E	**Interstitial wa-**	Chlorophyl-a	x																

	ter		(v)																
N	pigments	Chlorophyl-c	x (v)																
T		fuccoxanthine	x (v)																
	Metals	Cd,Cu,Pb,Zn											x						
	Chlorobiphenyls	CB(28,52,101,118,138,153,180,37,77,81,126,169)													x				
B	Fish - varia	genetic structure			x														
I		parasites : spp. and incidence			x														
O		stomach analysis	x																
T	Sea urchin	embryotoxicity test												*					
A	biological effect	metamorphosis test												x					
	Starfish metals	Cd,Cu,Pb,Zn											x						
	chlorobiphenyls	CB(28,52,101,118,138,153,180,37,77,81,126,169)													x				
	biological effect	MFO[1] activity											*						
		amoebocyte RO[2] species											*						
		embryotoxicity test												x					
		amoebocyte phagocytic activity											x						
		metallothioneins											*						
	Seabirds - marine mammals																		
	metals	Cd,Cr,Cu,Fe,Ni,Pb,Zn														x			
		selenium														x			
		total Hg																x	
		methyl Hg																x	
	organic	hydrocarbons																x	
		PAHs															*		
		polar lipids																x	
		total lipids														x			
		PCBs																x	
	organochlorines	DDE																x	
		DDT																x	
		aldrin																x	
		lindane																x	
		heptachlor epoxide																x	
	ecology	diversity																	x
		density																	x
		# per sp, devel.,		x															x

	plumage stage																	
varia	pathology															x		
	parasites															x		
	metallothioneins														x			
Plankton: phyto, bacterio,	composition								x									
nanozoo & microzoo	enumeration								x									
mesozoo	composition									x								
	abundance									x								
Benthos (meio, macro, epi, hyper)																		
ecology	diversity index	x																
	# species	x																
	density	x																
	biomass	x																
	dominance index	x																
	length freq. distribution	x																
	weight freq. distribution	x																
Maps - satellite derived	suspended matter										x							
	chlorophyll-a										x							

[1] MFO mixed function oxidases

[2] RO : reactive oxygen

p : particulate; s : dissolved, v : vertical profile, c : continuous measurements

* : parameter presently not measured

BNODC

IDOD Newsletter

Issue 4 January 2001

Foreword

Here is the fourth issue of the IDOD Newsletter.

Although we are working hard on the setting up of the information system, we won't bother you this time with technical information on data banking. We concentrate on the data themselves !

Have a nice reading...

The IDOD team

MUMM
MANAGEMENT UNIT OF THE NORTH SEA
MATHEMATICAL MODELS

UCS
Katholieke Universiteit Leuven
University Center of Statistics

Laboratoire SURFACES
Université de Liège

INSIDE THIS ISSUE

1. Sample registration onboard of the Belgica

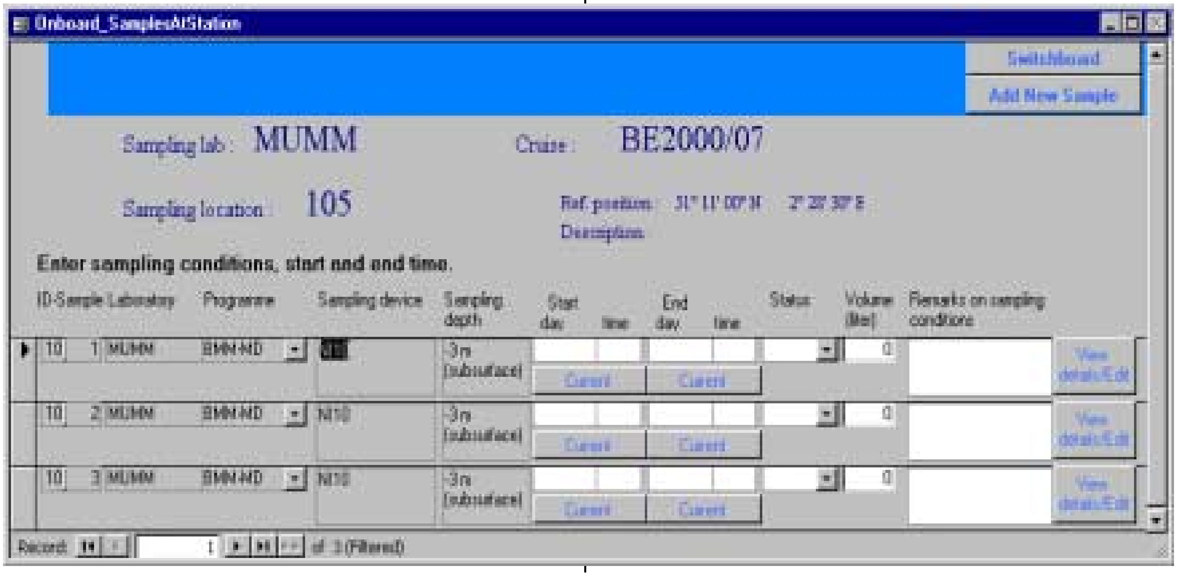
by MUMM

On the demand of data managers and data providers, IDOD has developed an application to record scientific activities onboard of the R/V Belgica. In practice, basic operations like CTD-profiles, collections of water bottles and sediments grabs, and net deployments are considered for registration.

It is obvious that datasets are only useful if the accompanying meta-information like sampling time, sampling position and sampling conditions is available. By entering this information at the moment of sampling, a more accurate and complete description of the sampling occasion is obtained, resulting in a better quality of the data in the IDOD database. The time needed for describing and documenting datasets and the risk on errors are substantially reduced.

The application also aims at facilitating the reporting obligations before and after the scientific cruise.

The application in MSAccess guides the scientists through a process of entering their planned samples. Afterwards, the sampling programme can automatically be generated. This sampling programme will be part of the cruise programme



Sampling lab : MUMM   Cruise : BE2000/07

Sampling location : 105   Ref. position: 51° 11' 00" N   2° 28' 30" E
Description:

Enter sampling conditions, start and end time.

ID-Sample		Laboratory	Programme	Sampling device	Sampling depth	Start day	time	End day	time	Status	Volume (liter)	Remarks on sampling conditions
10	1	MUMM	BMN-ND	[illegible]	-3 m (subsurface)						0	
10	2	MUMM	BMN-ND	N10	-3 m (subsurface)						0	
10	3	MUMM	BMN-ND	N10	-3 m (subsurface)						0	

directed by the chief scientist.

Onboard, the scientists from the different laboratories log on to view their own sampling programme. When the Belgica arrives at a sampling station, the scientist gets a list of his planned samples for this location by selecting the station. For every sample, the scientist only has to introduce the sampling time (start and end) and the conditions.

After the cruise, a file is generated for every laboratory containing all their samples with sampling time, sampling conditions, latitude, longitude and other ODAS parameters like wind speed and air temperature. The cruise report on the scientific activities is also automatically generated.

Another advantage of the system will be the automatic generation of Cruise Summary Reports.

Actually the application is tested onboard of the Belgica by scientists of MUMM. Some improvements and extensions still have to be developed. In a next phase (probably around mid-2001), all the laboratories having scheduled a campaign onboard of the Belgica will be asked to test the application.

2. Statistical analysis of the 1977–1996 monitoring data for seawater

(Part 3) – by UCS

Introduction

In previous IDOD newsletters the univariate analysis and the correlation analysis of the 1977-1996 monitoring data for seawater has been detailed. In this newsletter, the methods used to model the spatial continuity of the variables are described. As explained in the first IDOD newsletter, the univariate distribution, the correlation between the variables and the quantification of spatial continuity is used to set up quality checks of the data. The tools developed as part of this research will be also incorporated into the SAT program (statistical analysis tool) that will allow users of the IDOD database to examine the data in a similar manner.

Spatial continuity of the data

Geographically distributed data typically show some form of spatial continuity and hence some correlation between neighboring values. Such dependency can be used to advantage in the quality control by verifying one measurement value against another measured at a nearby location.

In order to perform such a test, it must be known how the difference between two measurements varies as a function of the distance between the two measurements. This dependence is expressed through the variogram that models how the expected value of the squared difference of two measurements varies with the distance h between the two measurements. The semi-variogram corresponds to 1/2 of this value and can be estimated on the basis of data as follows:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{|s_i - s_j| = h} (y_i - y_j)^2$$

where

$\gamma(h)$ is the estimated semi-variance at distance h,

h is the lag (that is the distance between sample locations),

y_l is the value of variable y at location s_l,

N(h) is the number of pairs of observed data points.

The previous estimation formula only works well if a sufficient number of data is available for each distance h considered. In the present data set this is not the case and therefore a continuous variation with the distance DIST is assumed as follows:

$$\gamma = a + b^*DIST^c.$$

A power model of this type allows for the fact that the data would show a (unknown) trend. For small values of c, the previous model is however indeterminate (i.e. the same model is obtained if a is increased while b is decreased by the same amount). In such a case, an alternative linear model is used that does not pose this problem. The model is in this case of the following simplified form:

$$\gamma = a + b^*DIST$$

When in this last model, b is estimated to be negative, the semi-variogram is assumed to be simply constant as follows:

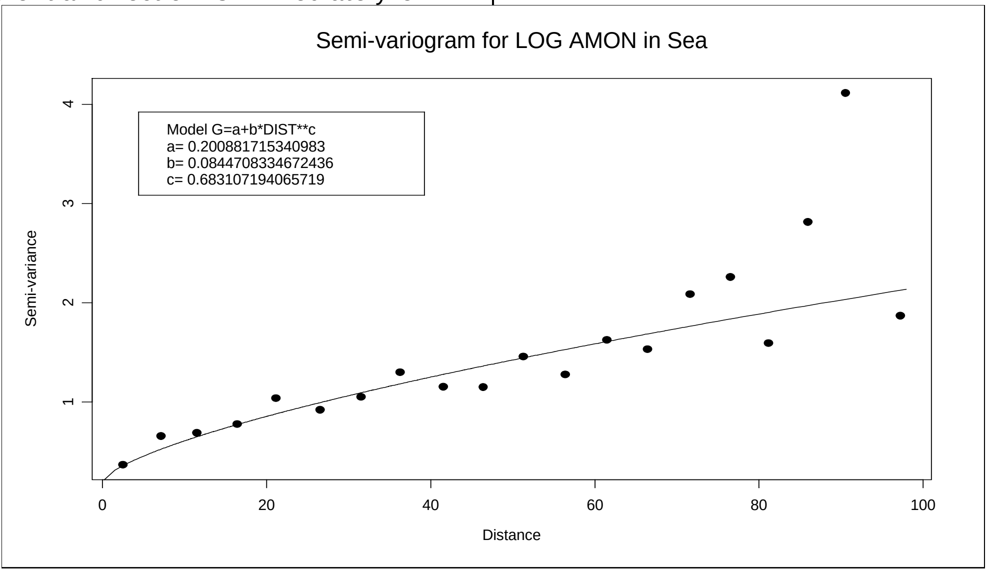
$$\gamma = a$$

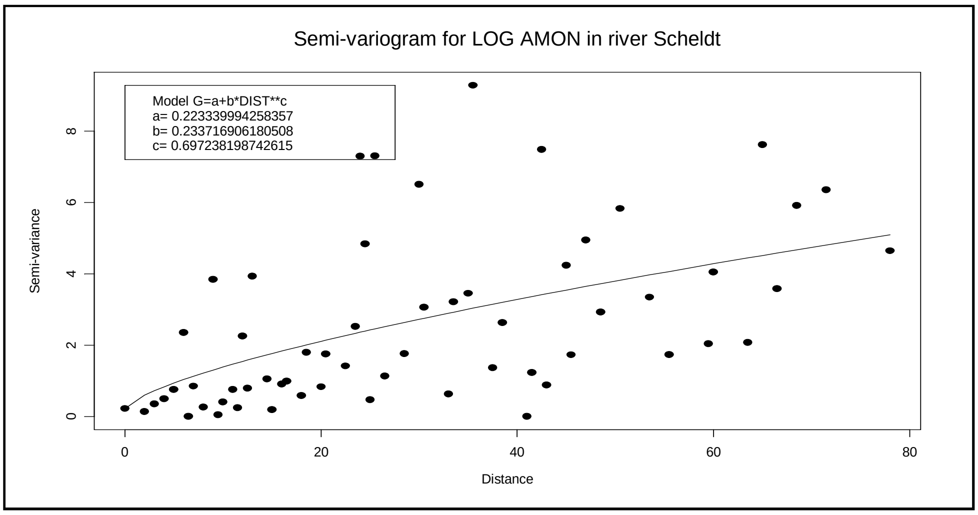
This last model corresponds to a case where there is no spatial correlation between neighboring stations. In practice this means that the measurement value cannot be checked against another measurement, even if such measurement is taken at a short distance.

Because the spatial continuity differs for observations in the river Scheldt and for observations in the North Sea, separate variograms have been estimated to model the differences in these two cases.

For observations in the river Scheldt the distances are measured along the river banks: it is natural to believe that differences between two measurements will increase as a function of this 1-D spatial coordinate. In the case of the open sea, no such preferential direction is immediately evident and, in this study the distance between two stations is defined to correspond to the Euclidean distance between the two stations. There is some evidence that perhaps better fits could be obtained by considering the distance to the coast (or the distance to the nearest river) as a preferential direction. This alternative modeling option can be further pursued, when the GIS-system that is connected to the IDOD database is available, but has not been further investigated in this study.

Fitting the semi-variogram to the observed differences between the measurements as a function of the distance measure is not a trivial problem and UCS has developed for this purpose a specific algorithm, that will be explained in one of the following newsletters. The figures below show the result for the case of the ammonium measurements in the river Scheldt and in the North Sea.





Both for the observations in open sea and for the river Scheldt a clear upward trend is observed as a function of the distance. However, the variogram is clearly higher in the river Scheldt indicating that in this case more variation between the measurements is to be expected for the same distance measure. Visually, the fit appears to be in both cases a satisfactory representation of the trend observed in the original data: scatter of the nonparametric estimates γ(h) (that are obtained by averaging the squared differences within a given distance interval) about the fitted model is to be expected and can be attributed to random variation; that this scatter would be larger for the Scheldt river is also logical because the estimates γ(h) are in this case based on fewer observations.

Application of the algorithm to the other variables shows that the variograms estimated in the river Scheldt generally show an upward trend (either the power model or the linear model is found to lead to the best fit). For the North Sea on the other hand, the upward trend is found to be less significant and, in particular for metals, it is found that a constant semi-variogram will typically lead to the best fit. As noted earlier, in this last case, spatial correlation would be of little use in a quality control check. In the other cases however, the variogram often shows a very steep increase and, if measurements exist at a small distance, it is clear that these measurements could be extremely useful to check the validity of another value. ♦

3. Spatial data analysis

by SURFACES

To compensate the lack of data measurements during an oceanographic campaign, we have used the tracking data collected in continuous during the ship displacement. These continuous data were sampled every 30 minutes. Such data are also very useful for the validation of the interpolation procedures realised on basis of the data collected at the sampling stations. Two parameters, temperature and salinity, were analysed for the data campaigns of 1996 available in the database prototype. The tracking data have been provided by MUMM-Oostende. Both parameters are independently considered in two distinct areas: open sea shelf area and Scheldt estuary.

Spatial distribution

The figure 1 shows the spatial data distribution of the tracking for campaign BE96/01. It is obvious that the distribution is quite different in the continental shelf shallow water area and in the Scheldt estuary. In the first area, a 2D spatial distribution is visible; while in the Scheldt estuary area, the sampling locations are organised along a line.

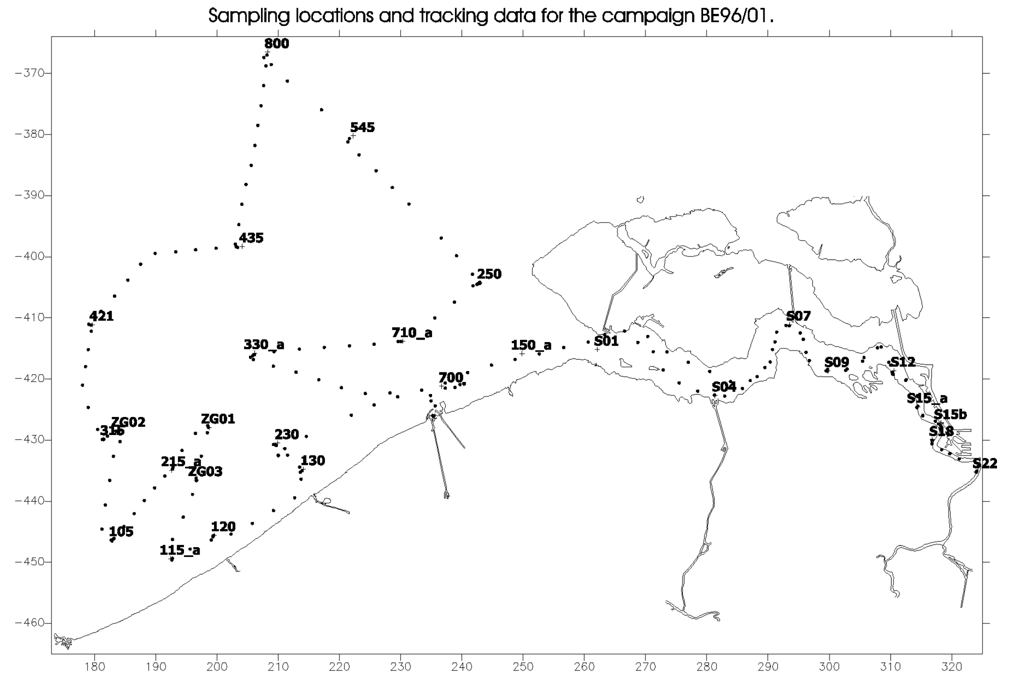


Figure 1

Moreover, in this last case, the estimation of the cross variability of the parameters is not possible because the ship track is nearly the same for all campaigns. Only the long-estuarine spatial variation can be quantified by plotting the salinity or temperature as a function of the distance along the Scheldt from an origin point. The distances are calculated with a propagation algorithm working on a grid. The chosen pixel size was 0.125 km.

Nevertheless, many incoherencies were found in the data locations: some sampling points being located on the continental area. An erosion process based on focal maximum filtering was then applied on the rasterized polygon cover of the coastlines to include these measurements in the estuarine area.

Scheldt estuary water salinity

In the Scheldt estuary, it appears clearly that the salinity is strongly correlated with the distance from the origin point located a few kilometres upstream the station S22 (proximity of Antwerp), as shown in the figure 2.

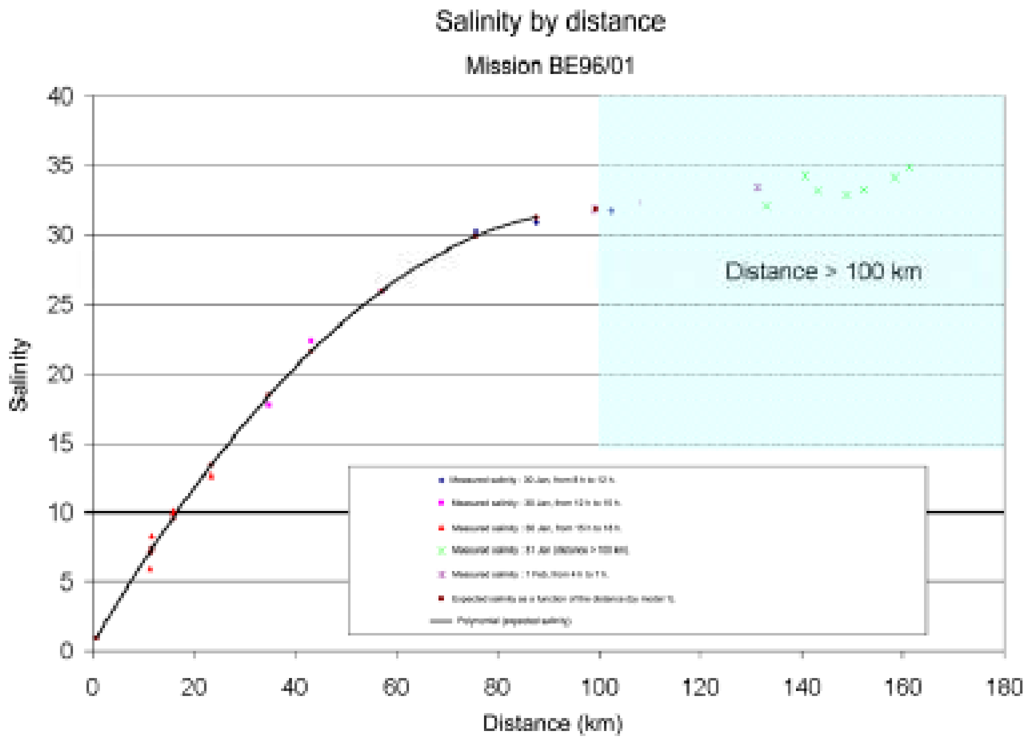


Figure 2.

The salinity begins to increase linearly and then the rate of increase diminishes as a function of the distance until reaching a sill after ~100 km from the origin point. The salinity model is a second order function and a quasi-Newtonian method was used to adjust its coefficients. The third dimension was neglected due to the lack of data across the Scheldt channel. The validation of this model was achieved by use of the 30 minutes along track data set. Unfortunately, the computation of the residual values (figure 3) cannot be considered as a true validation process because it is affected by a local-temporal effect (tidal coefficient, meteorological marine conditions, continental Scheldt hydrological conditions, tidal propagation along the estuary, time acquisition versus tidal state, etc...). Nevertheless, the residual values distribution is quite homogenous and there is a strong autocorrelation between the successive measurements. This autocorrelation does not appear on the spatiomap of figure 3 because the plotted data for similar distances are coming from different tracks realised at different times and places.

The location of the origin is arbitrary and the salinity magnitude and trend are probably influenced by the previously mentioned factors when the distance is increasing. Moreover, the spatial variation magnitude is obviously correlated with the relative importance of continental water discharge and the upstream and downstream propagation of seawater.

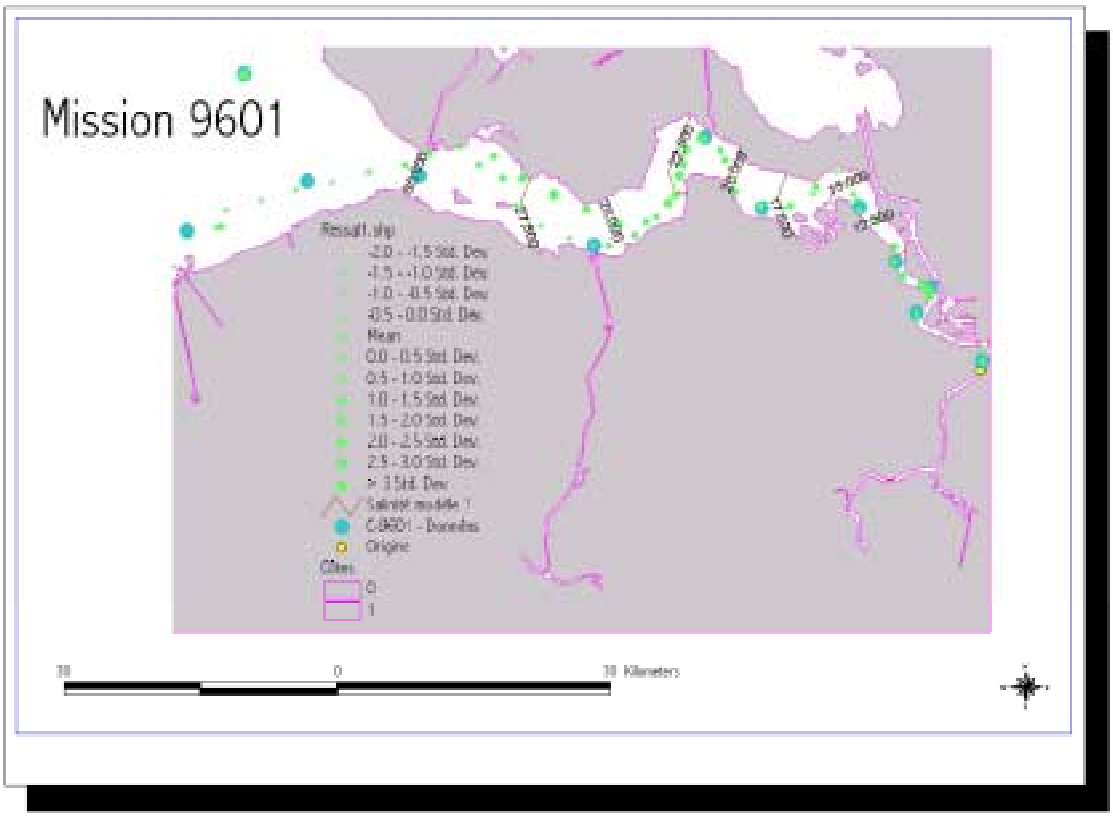


Figure 3.

Sea water temperature

On the continental shelf area characterized by shallow water, the interpolation process required more sophisticated techniques. Inverse distance weighted, linear TIN, quintic TIN, ordinary and, more recently, universal kriging interpolation methods were tested. The three first techniques are not well-adapted to the stations distribution resulting of the Belgica campaigns. In this case, the kriging method looks to be the best because more adapted to the low density of point and avoiding bull-eye effect around the sampling point location.

Some trend is present in the data collected at the sampling station and the universal kriging is thus well suited for these datasets. At the opposite, the ordinary kriging can also be considered when using the continuous track data. In fact, this last kind of interpolation technique shows shorter wavelength variability probably correlated with depth variations near the coast were water depth is not to large and when water confinement can be observed in particular bathymetrical and tidal conditions. The following figure 4 shows the result obtained by universal kriging for the temperature parameter of the campaign BE9601.

Digital Bathymetric Model (DBM) of the Vlaamse Banken area

A bathymetric model of the Vlaamse Banken area was generated from the 507785 sounding points measured between 1992 and 1999 and from the Vlaamse banken quoted points data set. The water volumes preservation was a constraint in the DBM construction. Grids were generated at four resolution levels of 5, 10, 15 and 20 m from a TIN created using the soundings of the RAAN95 campaigns.

The quality of the results obtained is not significantly different for the four resolution levels, if we consider the concept of water volumes preservation as expressed by the UGMM team. One high resolution grid has been computed for each zone. They all have been mosaïcked and then degraded at a 100 m resolution after a focal mean filtering in a square window with a 100 m size side. The resulting Vlaamse Banken grid (figure 5) has been computed in the original data set UTM reference system and then reprojected in the Postel reference system.

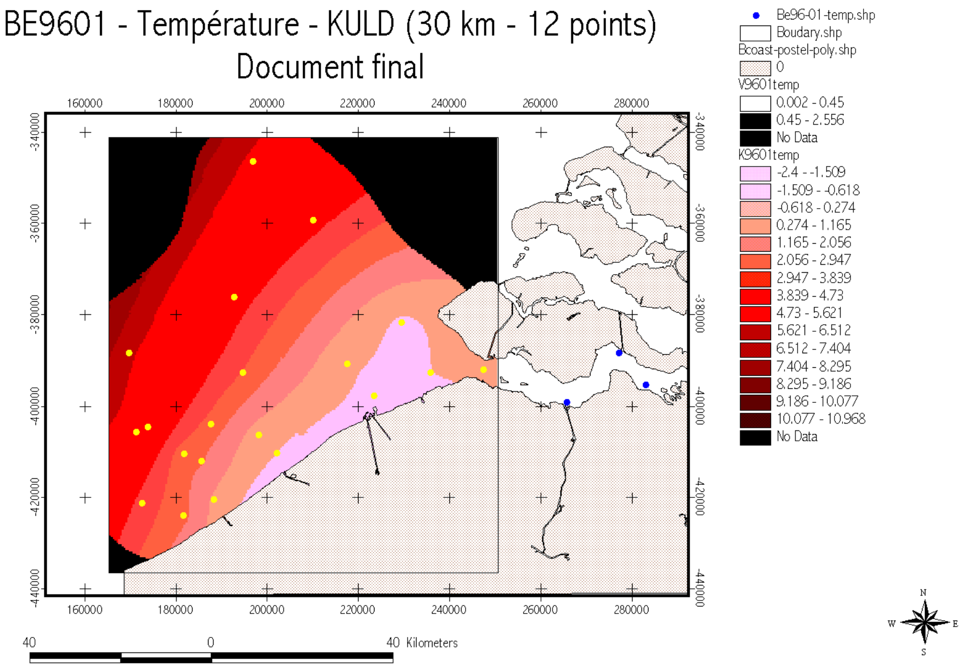


Figure 4.

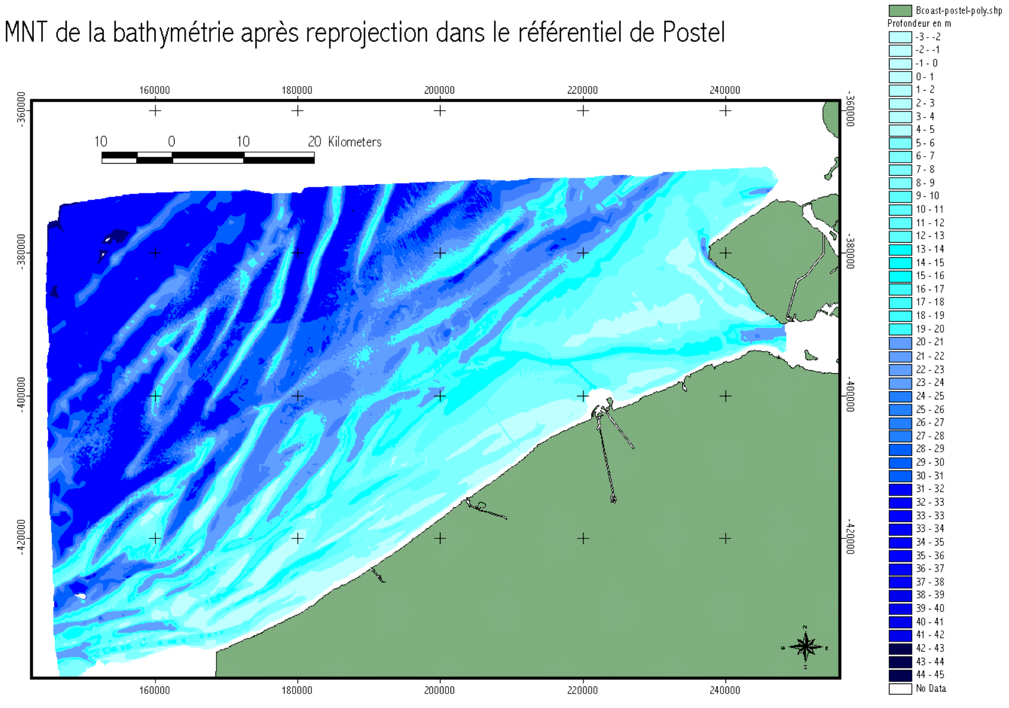


Figure 5.

4. Belgian National Oceanographic Data Centre

by MUMM

MUMM has recently been designated by the IOC (Intergovernmental Oceanographic Commission) as the Belgian National Oceanographic Data Centre (NODC).

A representative of MUMM was therefore present at the International Oceanographic Data and Information Exchange (IODE) meeting (http://ioc.unesco.org/iode). The IODE system has been established in 1961 to "enhance marine research, exploration, and development by facilitating the exchange of oceanographic data and information between participating Member States". The IODE system is composed of a network of more than 60 oceanographic data centres in as many countries around the world.

With the advance of oceanography from a science dealing mostly with local processes to one which is also studying ocean basin and global processes, researchers depend critically on the availability of an international exchange system to provide data and information from all available sources. Additionally, scientists studying local processes benefit substantially from access to data collected by other Member States in their area of interest. The economic benefit of obtaining data by exchange as opposed to collecting it oneself is huge.

Groups of Experts

IODE has already provided a considerable level of support. With regards to the problematic of the data and information exchange, the results of two Groups of Experts are especially important :

- The Group of Experts on Technical Aspects of Data Exchange (GE-TADE) focusing on the development of technical solutions for the management, exchange and easier integration of oceanographic data, the development of guidelines for formatting oceanographic information and the developments in a common data format which conforms to other major data collection programmes.
- The Group of Experts on Marine Information Management (GE-MIM) developing manuals and guides, regional Information Networks, creating a web based clearing house on training opportunities, monitoring of the development concerning electronic copyright legislation, evaluating the Integrated Library Management software, developing a toolkit for training,…

IODE activities and products

IODE contributes to following activities :

- ASFA: Aquatic Sciences and Fisheries Abstracts
- GODAR: Global Oceanographic Data Arcaeology and Rescue
- GLODIR: Global Directory of Marine (and Freshwater) professionals
- MEDI: Marine Environmental Data Information Referral System
- ODINAFRICA: Ocean Data and Information Network for Africa
- IODE Resource Kit: Comprehensive reference tool for ocean data and information managers

  It is a CD-ROM based product that contains a range of marine data-management and information-management materials, including software, quality control and analysis strategies, training manuals, and relevant IOC documents.
- OceanPortal: the one-stop shop for ocean data and information.
- Marine XML: XML standard development project

A Marine XML can provide a standardised data structure that can then support the creation of generic software, rather than the continuation of the present situation of building 'one-off' applications for each user or agency, that presently results in 'islands of information'. The Marine XML consortium aims to develop a specification for an XML framework to support the broad spectrum of marine data in a coordinated, structured and integrated manner.

The IDOD project ("Integrated and Dynamical Oceanographic Data Management") is funded by the Office for Scientific, Technical and Cultural Affairs in the frame of the programme "Sustainable management of the North Sea".

Plan d'appui scientifique à une politique de Développement Durable

Programme "Gestion Durable de la mer du Nord"

CONVENTION

entre

- Les Services fédéraux des affaires Scientifiques, Techniques et Culturelles, ci–après dénommés "les SSTC",
- l'Unité de Gestion du modèle Mathématique de la mer du Nord ("UGMM"), département de l'Institut Royal des Sciences Naturelles de Belgique, responsable du centre de données "IDOD",
- et «TITLE» «CONTACT», promoteur du projet de recherche **«REFERENCE»** ci–après dénommé(e) "le Promoteur".

Contexte

0.1 La présente convention est conclue entre les parties pour préciser leurs droits et obligations propres et réciproques en matière de données dans le cadre du programme "Gestion Durable de la mer du Nord" (ci–après dénommé "le programme").

0.2 Le programme est une des actions du "Plan d'appui scientifique à une politique de développement durable". Ce plan est une initiative du gouvernement fédéral belge et est mis en œuvre par les SSTC.

0.3 Dans ses lignes directrices, le programme précise : "*Afin de rassembler les données fragmentaires présentes chez les équipes de recherche, il est nécessaire d'évaluer la qualité de ces données et de les grouper de façon homogène dans une banque de données centrale belge, qui aura à les diffuser de façon optimale dans l'ensemble des réseaux thématiques et répondre à la demande extérieure. De leur côté, les équipes des réseaux thématiques devront fournir leurs données à cette banque, rapidement et de façon structurée.*" ("Plan d'appui scientifique à une politique de développement durable – Dossier d'informations générales")

Définitions

1.1 **Jeu de données** : tout ensemble de valeurs alphanumériques généré par l'exécution du contrat de recherche «REFERENCE» ainsi que les données descriptives ("méta–informations") correspondantes.

1.2 **Données descriptives** : toutes les données documentaires permettant de traiter et d'analyser les résultats scientifiques contenus dans le jeu de données.

1.3 **Auteur** : le(s) scientifique(s) responsable(s) d'un jeu de données.

1.4 **Embargo** : Conformément à l'usage, l'Auteur et les personnes que le Promoteur désigne à cette fin disposent d'une période d'utilisation exclusive de ses jeux de données, afin de les exploiter et de produire des publications.

1.5 **Réseau thématique** : le cas échéant, un ensemble des promoteurs associés, au sens défini dans le contexte du programme.

1.6 **Base de donnée IDOD** : l'ensemble du système d'information conçu et géré par le centre de données IDOD dans le cadre du contrat de recherche MN/DD/60 qui lie l'UGMM aux SSTC. Le contenu de la base de données IDOD n'est pas limité aux résultats obtenus dans le cadre du programme.

1.7 **Centre de données IDOD** : les moyens et le personnel affectés par l'UGMM à la bonne exécution du contrat de recherche MN/DD/60, ainsi que, pour les besoins de leurs recherches et développements spécifiques dans le cadre du Programme, le personnel affecté par la Katholieke Universiteit Leuven (Universitair Centrum voor Statistiek, Promoteur : J. Billiet, contrat de recherche MN/DD/61) et par l'Université de Liège (Laboratoire SURFACES, Promoteur J.-P. Donnay, contrat de recherche MN/DD/62).

Les SSTC

2.1 veillent à ce que le Promoteur transmette au centre de données IDOD les jeux de données acquis dans le cadre du Programme, conformément aux dispositions contractuelles.

Le Promoteur

3.1 s'engage à transmettre au centre de données IDOD les jeux de données correspondant aux travaux d'une année civile pour le 31 mars de l'année suivante au plus tard, conformément au contrat de recherche «REFERENCE» ;

3.2 fournit, pour chaque jeu de donnée couvert par la présente convention, les informations reprises en Annexe 1 ;

3.3 se conforme, autant que possible, aux recommandations du centre de données IDOD en matière de qualité, de documentation et de méthode de transfert des données ;

3.4 désigne l'Auteur des jeux de données qu'il transmet ;

3.5 désigne les personnes à considérer comme entrant dans la catégorie "A" des utilisateurs de la base de donnée (cfr 4.5) ;

3.6 informe les SSTC et le centre de données IDOD des dispositions pratiques mises en oeuvre pour assurer la bonne exécution de ses obligations, entre autres dans le contexte du réseau thématique auquel il participe (par exemple, désignation d'un *Project Data Manager*) ;

3.7 informe le personnel qu'il affecte à la bonne réalisation du contrat «REFERENCE» de son obligation de se conformer aux règles définies par la présente convention ;

3.8 veille à résoudre les litiges qui naîtraient, en matière de propriété intellectuelle, en cas de discordance entre ses obligations vis-à-vis des SSTC, les conventions qui le lient, ou ses collaborateurs, à son Institution et/ou la législation en vigueur.

L'UGMM

4.1 développe et gère la base de donnée IDOD, en conformité avec les obligations définies dans le contrat de recherche MN/DD/60 qui la lie aux SSTC ;

4.2 informe le personnel visé à l'Art. 1.7 de son obligation de se conformer aux règles définies par la présente convention ;

4.3 est responsable de la base de données IDOD pour sa conception, sa gestion et son évolution, y compris pour les outils de diffusion de l'information ;

4.4 est responsable du contenu de la base de données IDOD pour les informations qu'elle a elle–même fournies et pour les produits dérivés qu'elle aurait elle–même élaborés, en conformité avec l'Art. 4.9 ;

4.5 applique, sur les jeux de données qui lui sont transmis dans le cadre du programme, les règles d'accès suivantes :

- Par jeu de données, cinq catégories d'utilisateurs sont définies :
  A. le Promoteur, l'Auteur du jeu de données ainsi que les collaborateurs du Promoteur nommément désignés par le Promoteur,
  B. tout autre personne désignée dans une convention similaire conclue entre les SSTC, l'UGMM et un promoteur dans le cadre du Programme,
  C. les membres du personnel de l'état et, en particulier, de celui des SSTC et de l'UGMM,
  D. les autres utilisateurs, pour un usage scientifique ou éducatif,
  E. les autres utilisateurs, pour tout autre usage.

- Chacun de ces utilisateurs reçoit un droit d'accès personnel à ce jeu de données, selon le schéma suivant :

Utilisateur	Accès
A	Libre
B	Comme source d'information pour sa propre recherche, avec le consentement explicite et préalable de l'Auteur et selon le principe de réciprocité : libre Sinon : après embargo
C	Pour utilisation dans le cadre de ses activités de support à une politique de développement durable : libre Sinon : après embargo
D	Après embargo
E	Après embargo, sur la base d'une convention *ad hoc* cosignée par l'utilisateur, l'Auteur et l'UGMM (pour le centre de données IDOD)

- Dans le cadre du Programme, l'Embargo est fixé à 24 mois après la date officielle de transmission du jeu de données au centre de données IDOD (Cf. Art. 3.1).

- L'existence même d'un jeu de données n'est pas soumise à embargo.

4.6 veille à ce que tout utilisateur, lors d'une diffusion directe ou indirecte des données, se conforme aux règles en usage en matière de référence à l'Auteur et aux SSTC ;

4.7 s'engage à informer les SSTC des difficultés qu'elle rencontrerait dans l'obtention des jeux de données dans le cadre du programme ;

4.8 s'engage à mettre tous ses moyens en œuvre pour garantir la sauvegarde à long terme des jeux de données qui lui sont confiés ;

4.9 s'engage à assurer la promotion de la base de données IDOD et de son contenu en y associant le Promoteur et l'Auteur,

Concertation

À la demande des SSTC, l'UGMM organise l'animation d'un groupe d'utilisateurs de la base de données IDOD. Ce groupe est le forum de dialogue entre les gestionnaires de la base de données IDOD, ses utilisateurs, et les producteurs d’information. Il est donc l'instance première pour régler de manière harmonieuse les difficultés qui pourraient naître de l'application de la présente convention.

Pour les SSTC :	Le Promoteur :	Pour l'UGMM :
Nom :	«CONTACT»	Nom : ....
Date :	Date :	Date : ....

Annexe 1 : Données nécessaires à l'identification du jeu de données

Chaque fourniture d'un jeu de données au centre de données IDOD sera accompagnée des informations suivantes :

A. Données nécessaires à l'identification du jeu de données :

1. Nom du jeu de données.
   Il est suggéré d'utiliser le code suivant :
   Acronyme ou code SSTC du projet + année de référence + numéro de séquence

2. Code SSTC du projet dans le cadre duquel ces données ont été récoltées.

3. Identification du promoteur du projet.
   Nom, institution, service, adresse.

4. Identification de l'Auteur des données.
   Nom(s), institution, service, adresse.

5. Identification des personnes relevant de l'Art. 3.5 de la Convention pour ce jeu de données.
   Nom(s), institution, service, adresse.

6. Période couverte par le jeu de données.

7. Description de la structure et du contenu de chaque fichier.

B. Données descriptives ("méta–informations")

(Les informations nécessaires sont décrites dans le document IDOD–003.)

Plan voor wetenschappelijke ondersteuning van een beleid gericht op Duurzame Ontwikkeling

Programma "Duurzaam Beheer van de Noordzee"

OVEREENKOMST

tussen

- De federale Diensten voor Wetenschappelijke, Technische en Culturele Aangelegenheden, hierna de “DWTC",
- de Beheerseenheid van het Mathematisch Model van de Noordzee ("BMM"), departement van het Koninklijk Belgisch Instituut voor Natuurwetenschappen, dat verantwoordelijk is voor het datacentrum "IDOD",
- en «TITLE» «CONTACT», promotor van het onderzoeksproject **«REFERENCE»** hierna “de Promotor” genoemd.

Context

0.1 Onderhavige overeenkomst wordt tussen partijen gesloten om hun eigen en wederzijdse rechten en plichten inzake data nader te bepalen in het kader van het programma “Duurzaam Beheer van de Noordzee” (hierna “het programma” genoemd).

0.2 Het programma is één van de acties van het “Plan voor wetenschappelijke ondersteuning van een beleid gericht op duurzame ontwikkeling”. Dit plan is een initiatief van de Belgische federale regering en wordt ten uitvoer gebracht door de DWTC.

0.3 In zijn richtlijnen bepaalt het programma: "*Om de bij de onderzoeksgroepen fragmentair aanwezige meetgegevens voor de vooropgestelde doelstellingen bruikbaar te maken, dienen al deze gegevens geëvalueerd te worden op hun kwaliteit en dienen ze op homogene wijze gegroepeerd te worden in een centrale, Belgische gegevensbank, die ze op optimale wijze zal verspreiden binnen het geheel der thematische netwerken en op externe vraag. Van hun kant dienen de ploegen van de thematische netwerken hun gegevens op snelle en gestructureerde wijze aan deze databank te leveren.*" ("Plan voor wetenschappelijke ondersteuning van een beleid gericht op duurzame ontwikkeling – Algemeen Informatiedossier")

Definities

1.1 **Dataset**: elke verzameling alfanumerieke waarden die tot stand komt bij de uitvoering van het onderzoekscontract «REFERENCE» alsook de overeenkomstige beschrijvende gegevens (“meta-informatie”).

1.2 **Beschrijvende gegevens**: alle documentaire gegevens aan de hand waarvan de wetenschappelijke resultaten bevat in de dataset, verwerkt en geanalyseerd kunnen worden.

1.3 **Auteur**: de wetenschapper(s) verantwoordelijk voor een dataset.

1.4 **Embargo**: Zoals gebruikelijk beschikken de Auteur en de personen die de Promotor in dit verband aanduidt over een periode waarin zij het exclusief gebruik hebben van hun datasets, waardoor zij de kans krijgen de data te verwerken en te publiceren.

1.5 **Thematisch netwerk** : eventueel, een groep van geassocieerde promotoren, zoals gedefinieerd in het programma.

1.6 **IDOD Gegevensbank**: het hele informatiesysteem zoals dat ontworpen en beheerd wordt door het IDOD datacentrum in het kader van het onderzoekscontract MN/DD/60 tussen de BMM en de DWTC. De inhoud van de IDOD gegevensbank blijft niet beperkt tot de resultaten die in het kader van het programma ontstaan.

1.7 **IDOD Datacentrum**: de middelen en het personeel ingezet door de BMM voor de goede uitvoering van het onderzoekscontract MN/DD/60, evenals, voor hun specifieke onderzoek en ontwikkelingen in het kader van het Programma, het personeel verbonden aan de Katholieke Universiteit Leuven (Universitair Centrum voor Statistiek, Promotor : J. Billiet, onderzoekscontract MN/DD/61) en aan de Université de Liège (Laboratoire SURFACES, Promotor J.-P. Donnay, onderzoekscontract MN/DD/62).

De DWTC

2.1 zien erop toe dat de Promotor de datasets verworven in het kader van het programma aan het IDOD datacentrum overmaakt, overeenkomstig de contractuele bepalingen.

De Promotor

3.1 verbindt zich ertoe om de datasets die voortvloeien uit de werkzaamheden van één kalenderjaar uiterlijk tegen 31 maart van het daarop volgende jaar aan het IDOD datacentrum over te maken, overeenkomstig met het onderzoekscontract «REFERENCE»;

3.2 verstrekt voor elke dataset die onder onderhavige overeenkomst valt, alle informatie zoals vermeld in bijlage 1;

3.3 houdt zich, in de mate van het mogelijke, aan de aanbevelingen van het IDOD datacentrum op het vlak van gegevenskwaliteit, -documentatie en wijze van overdracht;

3.4 geeft de Auteur op van de datasets die zij/hij doorstuurt;

3.5 duidt de personen aan die thuishoren in de categorie “A” van de gebruikers van de gegevensbank (cf. 4.5);

3.6 brengt de DWTC en het IDOD datacentrum op de hoogte van de praktische maatregelen getroffen om de goede uitvoering van zijn verplichtingen te verzekeren, onder andere in de context van het thematisch netwerk waarin hij deelneemt (bvb. aanduiding van een *Project Data Manager*);

3.7 brengt het personeel dat hij inzet voor de goede uitvoering van het contract «REFERENCE» op de hoogte van hun verplichting om zich te houden aan de regels vastgelegd in onderhavige overeenkomst;

3.8 ziet toe op het oplossen van eventuele geschillen inzake intellectuele eigendom die kunnen ontstaan uit tegenstrijdigheden tussen zijn verplichtingen ten overstaan van de DWTC, de overeenkomsten die hem en zijn medewerkers aan hun Instelling binden en/of de geldende wetgeving.

De BMM

4.1 ontwikkelt en beheert de IDOD gegevensbank overeenkomstig de verplichtingen zoals die vastgelegd werden in het onderzoekscontract MN/DD/60 met de DWTC;

4.2 brengt het personeel vermeld in art. 1.7, op de hoogte van de verplichting zich te houden aan de regels vastgelegd in onderhavige overeenkomst;

4.3 is verantwoordelijk voor de IDOD gegevensbank, voor het ontwerp, het beheer en de evolutie ervan, inclusief de tools tot verspreiding van de informatie;

4.4 is verantwoordelijk voor de inhoud van de IDOD gegevensbank wat betreft de gegevens die zij zelf geleverd heeft en voor de afgeleide producten die zij zelf uitgewerkt zou hebben, overeenkomstig met Art. 4.9;

4.5 past op de datasets die haar in het kader van het programma overgemaakt worden, onderstaande toegangsregels toe:

- Per dataset worden vijf gebruikerscategorieën gedefinieerd:
  - A. De Promotor, de Auteur van de dataset alsook de medewerkers van de Promotor, zoals met naam aangewezen door de Promotor;
  - B. elke andere persoon die middels een gelijkaardige overeenkomst tussen de DWTC, de BMM en een Promotor in het kader van het Programma, aangewezen werd;
  - C. de personeelsleden van de Staat, en in het bijzonder die van de DWTC en de BMM;
  - D. de andere gebruikers, voor wetenschappelijk of educatief gebruik;
  - E. de andere gebruikers, voor elk ander gebruik.

- Ieder van deze gebruikers krijgt een persoonlijk toegangsrecht tot deze dataset, volgens onderstaand schema:

Gebruiker	Toegang
A	Vrij
B	Als informatiebron voor eigenonderzoek, mits de uitdrukkelijke en voorafgaande toestemming van de Auteur en volgens het reciprociteitsbeginsel : vrij Zoniet: na embargo
C	Voor gebruik in het kader van zijn activiteiten tot ondersteuning van een beleid van duurzame ontwikkeling: vrij Zoniet: na embargo
D	Na embargo

E	Na embargo, op basis van een *ad hoc* overeenkomst meegetekend door de gebruiker, de auteur van de gegevens en de BMM (voor het IDOD datacentrum)

- In het kader van het Programma is het Embargo vastgelegd op 24 maanden na de officiële datum van overdracht van de dataset naar het IDOD datacentrum (Cf. Art. 3.1).
- Het bestaan zelf van een dataset is niet aan een embargo onderworpen.

4.6 zorgt ervoor dat elke gebruiker, ten tijde van een rechstreekse of onrechstreekse verspreiding van de gegevens, op de correcte wijze refereert naar de Auteur en de DWTC;

4.7 verbindt zich ertoe de DWTC de moeilijkheden te melden die zij zou ondervinden bij het verkrijgen van de datasets in het kader van het programma;

4.8 verbindt zich ertoe alle middelen in te zetten om de bewaring op lange termijn van de haar overgemaakte datasets, te verzekeren;

4.9 verbindt zich ertoe in te staan voor de promotie van de IDOD gegevensbank en de inhoud waaraan de Promotor en de Auteur geassocieerd worden.

Overleg

Op vraag van de DWTC organiseert de BMM de werking van een gebruikersgroep van de IDOD gegevensbank. Deze groep vormt de basis voor een dialoog tussen de beheerders van de IDOD gegevensbank, de gebruikers en de informatieleveranciers. Deze groep is dan ook in eerste instantie aangewezen om de eventuele moeilijkheden die uit de toepassing van onderhavige overeenkomst zouden kunnen ontstaan, op harmonieuze wijze op te lossen.

Voor de DWTC :	De promotor :	Voor de BMM :
Naam :	«CONTACT»	Naam : ....
Datum :	Datum :	Datum : ....

Bijlage 1 : Gegevens noodzakelijk voor de identificatie van de dataset.

Elke gegevenstoelevering aan het IDOD datacentrum zal vergezeld zijn van de volgende informatie :

A. Gegevens noodzakelijk voor de identificatie van de dataset :

1. Naam van de dataset.
   Er wordt voorgesteld om de volgende code te gebruiken :
   Acroniem of DWTC code van het project + referentiejaar + volgnummer

2. De DWTC code van het project waarvoor deze gegevens verzameld werden.

3. De identificatie van de Promotor van het project.
   Naam, instituut, dienst, adres..

4. De identificatie van de auteur van de gegevens.
   Naam (namen), instituut, dienst, adres.

5. De identificatie van de personen volgend uit het Artikel 3.5 van de Overeenkomst voor deze dataset.
   Naam (namen), instituut, dienst, adres.

6. De periode waarop de gegevens betrekking hebben.

7. De beschrijving van de structuur en de inhoud van elk bestand.

B. Beschrijvende gegevens ("meta-data")

(De vereiste gegevens zijn beschreven in het document IDOD–003.)

UCS

Universitair Centrum voor Statistiek

Integrated and Dynamical Oceanographic Data Management - IDOD

Scientific Report

June 2002

Contribution of KUL-UCS

KATHOLIEKE UNIVERSITEIT LEUVEN

University Center of Statistics
Katholieke Universiteit Leuven
de Croylaan 52b
B-3000 Leuven
Belgium

1. Introduction

During the fifth and last year of the IDOD project UCS has concluded the work on the development and integration of the following two software tools into the project:

1. SAT: a statistical analysis tool that allows to perform in a simple and user-friendly manner basic statistical analyses on data retrieved from the database. This program can be operated over the web;
2. SQC: a stand-alone program that allows to screen data in the database and perform a statistical quality control on those data. Use of this program is reserved for the datamanager who is responsible for maintaining the quality of the data of the database.

With these two programs two basic objectives that were set out at the start of the project are accomplished:

1. to make the data widely available and allow for intelligent use and interpretation of the data;
2. to qualify the data to the extent possible such that the data can be confidently used and misinterpretations based on anomalous data are avoided.

In this final scientific report the concepts of both programs and the work involved in their development is briefly summarized. User manuals describing in detail the functionality of both programs are shown in Appendices 1 and 2.

2. Role of SQC and SAT in the Overall Dataflow

Figure 1 shows schematically the place of SQC and SAT in the overall dataflow. Raw data from the data providers are entered into the IDOD database. At this stage, the first step of a 3-level quality assurance is applied: it is verified that the data conform to a number of data acceptance tests (i.e. metadata must be present, etcetera). When entering the data into the database, a second-level of quality assurance concerning data integrity is applied (i.e. verification of correct formatting, verification that references to measurement locations are known in the database, etcetera). After these two stages, the data are available to the users but are "unqualified" at the third level of the quality assurance (the statistical quality control).

The work performed by UCS concerns then this third level of the quality assurance and the examination of the data by users with the help of a statistical analysis tool.

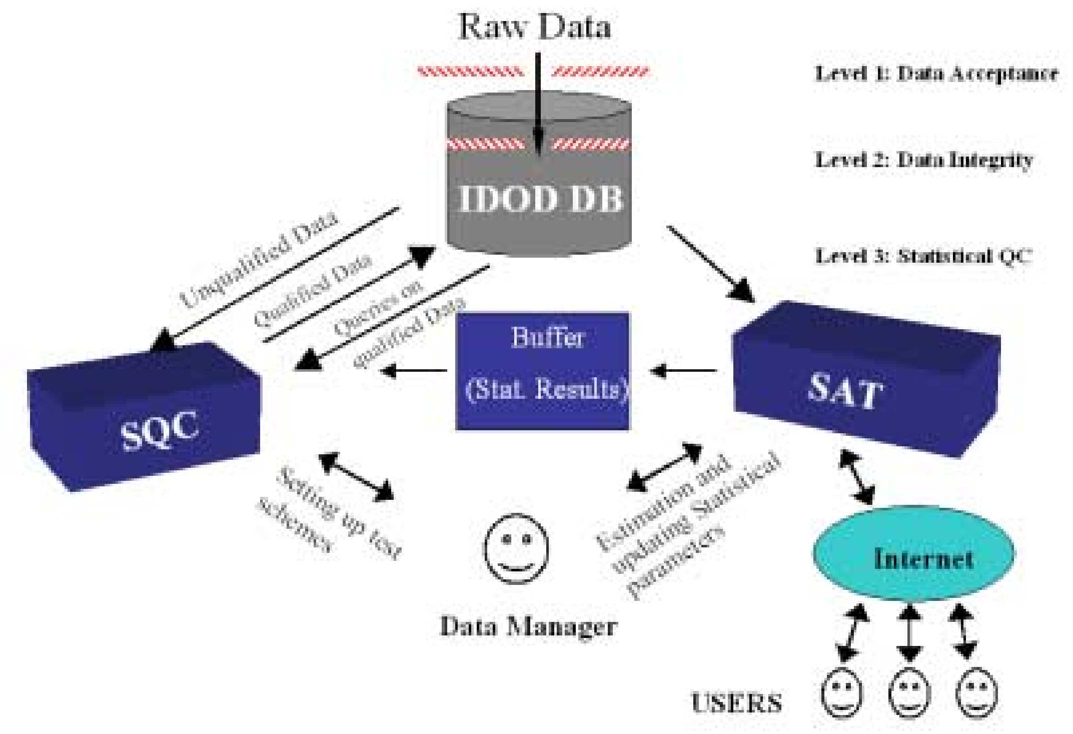


Figure 1: Role of SQC and SAT in the overall dataflow

As shown at the right side of the diagram the SAT program can be run through the Internet by multiple users. Data can be retrieved by the users and examined through a variety of analysis tools. This allows to view the data not only in their raw numeric form, but allows to also understand interrelationships, trends, etcetera which are typically most important to scientific users.

The same SAT program is also useful to the datamanager for the purpose of characterizing normal statistical patterns of the data: as an example, the datamanager can examine how one variable statistically relates to other variables in the database.

The statistical information that is derived from this analysis can then be send by the datamanager to a special buffer (an ACCESS database). For this purpose, SAT includes some special functions that are reserved for use by the datamanager only. Having examined the data and derived the necessary information, the datamanager can then declare in the SQC program a series of tests based on this statistical information. Once such a "test scheme" has been declared, the data sets can be retrieved from the database into the SQC program for qualification at the third-level of the quality assurance. The result of this operation is the assignment of quality labels to each individual datapoint that indicate which tests were passed and which tests failed for each datapoint. This multivariate outcome is further summarized into a single numeric value between 0 and 1 that summarizes to which extent the datapoint conforms to "normal" behavior: i.e. values close to 1 indicate that nothing unusual is noted for the datapoint; values close to 0 indicate that the measurement is highly unusual.

Once a test scheme has been set up and is found to be satisfactory (i.e. data with low numeric quality values are indeed found to be "anomalous", while data with high numeric quality values are indeed found to be "normal"), the test scheme can be routinely applied to new incoming datasets.

The dataflow scheme of Figure 1 has been purposely designed to avoid that the third-level quality assurance would be a bottleneck in making data available to the users: when datasets arrive that contain new types of measurements or a different combination of such measurements, it will typically be necessary to develop a new test scheme. Doing so requires extensive use of SAT to understand the marginal distributions of the data, the interrrelationships between the variables and the spatial correlation of the data and can be a time-consuming proces. For this reason, the SQC labelling has not been retained as mandatory but only as recommendable.

It is also clear that the third-level quality assurance should be viewed as a dynamically evolving feature that will improve in time as the normal statistical pattern of the data is better understood and more data become available in the database to establish these patterns. Because the database has become operational only at the end of the project, the SQC program has been developed as a general tool to organize the statistical quality control. The number of actual test schemes that has been implemented so far is limited. Nevertheless, first applications show that the strategy is effective and indeed in some cases measurements with a-typical behavior are detected. It should be stressed that an "a-typical" measurement need not necessarily be an "erroneous" measurement. On the contrary: "a-typical" measurements may be data that are of most interest and signal that some relationships are badly understood or prone to unexpected large deviations. However, one should recognize that in practice even with the best precautions and measurement practices, erroneous measurements do occur (i.e. because of instrument error, contaminations, mislabelling, human errors, …) and of course it is highly important that such measurements are detected. Ultimately it is the responsibility of the data manager and the users themselves to make this distinction between "valid" and "erroneous" measurements. This is why in the database a user quality label is reserved for each datapoint that can override the automated label that is assigned by SQC. The SQC labelling is thus intended only as a tool that draws attention to "a-typical" data. If very low numeric quality values are obtained, the associated measurements should be further researched to identify the source of the error or their a-typical behavior (i.e. by contacting the data provider and searching for additional meta-information or explanations of the a-typical behavior).

3. The Statistical Analysis Tool

The objective of the Statistical Analysis Tool (SAT) program is to make a number of statistical analyses tools readily available to the users for examining the data. The development effort involved and the final functionalities that have been implemented are reviewed next. A detailed user manual is attached in appendix.

3.1 Development Effort

At an early stage of the project it has been decided to use a general statistical software package to support the actual statistical calculations. After a comparison of the relative advantages and disadvantages of the different packages, the S+ statistical software has been chosen because it is highly performant, has many functionalities and allows for preprogramming statistical operations. Operating the S+ software requires however expert knowledge of the software and of statistical analysis and therefore a user-interface had to established that efffectively '"hides" the underlying package and reduces the input choices to a few easy-to-understand input values. At the same time, the overall SAT tool had to operate through the Internet and allow for its use by multiple users.

The work performed by UCS during the second and third year of the project concentrated primarily on the development of appropriate S+ pre-programmed functions that in combination with a Visual Basic shell allow to execute standardized statistical analyses in a user-friendly manner. This work comprised: a) designing the functionality and reducing it to simple choices, b) implementing and testing the S+ software, c) writing the necessary visual Basic modules for entry of the parameters in a graphics-user-interface, passing the parameters to the S+ functions and returning the final results in a visual or numeric display.

The last two years of the work performed by UCS have focused entirely on making this same functionality available in a web-based environment. This involved several software problems all of which were resolved in the course of the project: i.e. building web-pages, passing the arguments of the web-pages to the underlying Visual Basic process, avoiding conflicts when multiple users request analyses at the same time, displaying results in web pages, etcetera.

While most of the functions that have been implemented correspond to standard statistical analyses some analyses were specifcally designed and programmed in the S+ language for the purpose of the IDOD project and in particular to support the statistical quality control. Specifically:

1. in verifying the statistical relationship of one measurement with other simultaneously measured data, it may often occur that some of the other

measurements are missing and as a consequence for different measurement instances, different sets of covariate variables may be available for the verification. For this reason a "subset-regression" functionality has been build into the SAT program that allows to rank different potential sets of covariate variables according to the accuracy with which they can predict the response variable (i.e. the measurement under examination). The results of the alternative regressions as well as the ranking are then stored in the statistical buffer (see Figure 1) and automatically used by SQC to perform the best test considering the available measurement values for a given record (i.e. measurement instance);

2. an important means for validating data consists in checking the spatial difference between one measurement versus the closest measurement in space that is of the same type and falls within a given time window. To execute such a test, statistical information about how the difference between such measurements increases with distance must be derived. The relationship of the variance of this difference with the distance measure is referred to as a "variogram". Because the data from which the variogram can be estimated are sparse and are not on a regular grid, and because the squared differences that may be used to estimate the variance do not follow a normal distribution, a specific algorithm has been developed for this purpose and has been implemented in the SAT program. This algorithm is fully detailled in the Master of Statistics thesis by Plevoets (1999).

3.2 Implemented Functionalities

The functionalities of the SAT program are grouped into 3 broad categories of which only the last one includes actual statistical analyses:

- Import/Query

  This functionality allows to enter the data that are previously queried from the IDOD database (using a query form developed by MUMM/Ulg) into the SAT tool. Data that have been previously extracted by the user can be reloaded.

- Data Handling

  This functionality supports transformations, filtering and viewing of the different variables contained in the dataset.

- Statistical Analysis

  The possibilities to perform a statistical analysis are further subdivided into following groups:

- Summary Statistics.

  This group of options allows to derive either numerical or graphical summary statistics (i.e. boxplots, histograms, Q-Q normality plot, density plot).

- Normality Test

  This option allows to check more in detail the normality of a variable in the dataset through the use of chi-square goodness-of-fit testing and testing of the Kolmogrov-Smirnov statistics.

- Trend fitting

  The trend analysis is used to look for the relationship between a variable of interest (the response variable, e.g. salinity) and another variable (the regressor variable, e.g. distance from coast). With SAT a linear, quadratic and nonparametric function can be fitted to the data and visually displayed.

- Correlation Analysis

  This option is split into Correlation/Covariance Matrix and Scatterplotmatrix. The correlation/covariance matrix function calculates the correlation between the different variables in a dataset. The scatterplotmatrix visualizes the correlation between the different variables of a data set.

- Regression Analysis

  The regression analysis functions allow to examine the relationship between one variable and one or more of the other variables. The multiple regression option allows to declare general models (i.e. including interaction terms and or quadratic terms or higher) and derive the regression results for the model. The subset regression option determines for a list of potential regressor variables a ranked sublist (according to the R^2 statistic) of the sets of most explanatory variables. This option is particularly useful for the statistical quality control of the data as explained earlier in Section 3.1.

- Variogram Analysis

  The variogram analysis allows to investigate how the variance of the difference between two measurements of the same type varies with the distance between those variables. Option "variogram calculation" determines nonparametric estimates for different distance intervals. Option "variogram fit" fits a power law variogram to the data following the algorithm developed by Plevoets (1999).

The SAT program also includes a "SQC modelling" option that is intended for use by the data manager only to set up test schemes. The suboptions "distribution test", "regression test" and "spatial test" are variations on the standard SAT options "trend fitting", "regression subset" and "variagram fit" that allow to automatically store the results of the analysis into a buffer that can be accessed by the SQC program for use in execution of the tests.

4. The Statistical Quality Control

The objective of the Statistical Quality Control (SQC) program is to organize in a systematic fashion the valdiation of the data by verifying that the measurements conform with normal statistical patterns. As in the previous chapter, the development effort and implemented features are discussed separately.

4.1 Development Effort

Background

The conceptual design of the SQC program has been decided upon again at an early stage of the project (i.e. year 1) after examination of a typical datset (see for instance "Exploratory Statistical Analysis Of The 1996 Belgian Monitoring Data For Seawater" in appendix to the scientific rapport of 1998). Specifically, on the basis of this analysis it was decided that apart from deterministic tests (i.e. checking for limits that are applicable for a given dataset) the three characteristics that are amenable to statistical verification concern:

1. verification that a measurement falls within reasonable bounds of the marginal distribution that is expected after removal of eventual spatial and/or seasonal trends,
2. verification that a measurement falls within reasonable bounds of the distribution that is expected given the measurement values of other variables taken at the same location and time,
3. verification that the difference of a measurement with a measurement of the same variable taken at another location falls within reasonable bounds.

Because at that time the database was not operational and therefore no specific test schemes could be developed, it was further decided that a general supporting software had to be developed that would allow in a relatively easy manner to declare testing schemes and that would be able to combine the results of the individual tests into a single numeric quality label. Furthermore it was noted that in many cases some of the variables may be missing for some of the measurement instances and the test scheme should accommodate for such instances. The general conceptual framework that followed from these considerations is described next.

Conceptual Framework

A-priori one would expect any measurement to be a good-quality observation of the physical process under examination. The degree-of-belief in this statement may however vary, for instance depending on the data provider, the type of instrument used or other covariate variables that may influence the quality of the measurement.

This type of information can be mathematically captured by assigning to each measurement an a-priori probability p that the datapoint is of "normal" quality. Datapoints of "normal" quality display the normal type of random variation to be expected. In principle, the value of p can be objectively defined: i.e. after examination of the data and expert identification of "normal" and "abnormal" datapoints, the value of p can be taken equal to the proportion of "normal" datapoints. Variation of p with covariate conditions would be also possible by classifying the data into different categories according to these covariate conditions. In the present version of SQC, the p values is differentiated according to the measurement type but does not take covariate information into account.

Without any further information (i.e. other data measured in the databank), the final numerical quality label q corresponds directly to the a-priori value p. When however other data are available, it is possible to verify the statistical consistency of the datapoint with those other data (i.e. consistency in terms of marginal distribution, correlation with other variables, consistency in space, ….). Consider then such a single test T which leads to a discretized test result with 3 possible outcomes: i.e. "0"=test passes, "1"="test fails", "N"= the test cannot be applied because of missing data. Because of the intrinsic random nature of the data, one cannot exclude that a "normal" datapoint does lead to a test outcome "1". However, it is unlikely that this would happen and much more likely to be the result of "abnormality". The probability of a "normal" measurement falling into any of the test results is typically defined by the nature of the test and can be objectively defined: for instance, "normal" datapoints fall with a probability of 5% outside 95% prediction intervals. The probability of an "abnormal" datapoint falling into the different test intervals can only be objectively defined if specific information is available about the deviating pattern. This is seldom the case and deviations may have many origins: i.e. instrumental errors, transmission errors, coding errors, … For this reason, subjective probabilities need to be assigned to the different test outcomes in the case of "abnormal" data. These subjectively assigned probabilities can be however used to "calibrate" the classification scheme (i.e. if for a learning dataset normal and abnormal data have been identified).

The information about the power of the test to differentiate between normal and abnormal data can then summarized by an "information" matrix as shown in Table 1: the value α correspond to the probability that a "normal" datapoint leads to failure of the test; similarly the value β corresponds to the probability that an "abnormal" datapoint leads to passing the test.

test result	0=pass	1=fail
"normal"	$\alpha_0 = 1-\alpha$	$\alpha_1 = \alpha$
"abnormal"	$\beta_0 = \beta$	$\beta_1 = 1-\beta$

Table 1 - "Information" matrix: probability of a measurement leading to a test result for "normal" and "abnormal" data

Knowing the outcome i of the test, the a-posteriori probability q of a datapoint being normal is then found using Bayes theorem as:

$$q \mid i = \frac{p\alpha_i}{p\alpha_i + (1-p)\beta_i} \qquad (1)$$

If $\alpha_i = \beta_i$, then the test is non-informative and the a-priori and a-posteriori probability are identical. If $\alpha_i > \beta_i$, then the a-posteriori probability increases reflecting the fact that circumstantial evidence is found that the measurement does conform to the regular statistical patterns. When $\alpha_i < \beta_i$, then the a-posterioir probability will decrease and a lower quality label is assigned, since the datapoint is found to lead to a test result that is unusual for "normal" data.

When different tests are applied, Equation (1) can be sequentially applied. In doing so, an assumption of independence is made which is difficult to verify, but from a practical point of view this assumption should be acceptable in sofar the different tests truly consider different statistical patterns. One may also note that the order of application of the tests does not influence the result.

As an alternative to this scheme of assigning a numeric quality indicator one could consider the use of a P-value that is assigned to each measurement: in statistical terminology, the P-value refers to the probability that the noted deviation would be as large or larger if the measurement conforms to the normal statistical pattern. Values that show a deviation that is unusually large would be assigned a small P-value, whereas values that show small deviations would be assigned a large P-value.

The Bayesian scheme that is here proposed has however following advantages over the use of P-values to quantify data validity:

1. it can incorporate in a natural way a-priori information regarding the quality of the data that is derived from information other than purely statistical tests (i.e. data provider, instrument type, covariate information, ….);

2. there is no reason to associate different quality labels to measurements that fall within a normal range of variation as would be the case if a P-value is used;
3. while the use of a P-value may seem statistically more rigorous, combining the P-values obtained from different tests would involve (similar to the Bayesian scheme) making some assumptions of independence and a pragmatic rule to avoid that P-values are systematically biased towards lower values if many tests are applied. The Bayesian scheme has the advantage that the a-priori value can both increase (if the datapoint is found to be conform with regular statistical behavior) and decrease (if the datapoint is found to be not conform with regular statistical behavior, or more precisely is more likely to conform to irregular behavior);
4. related to the previous point, the Bayesian scheme deals in a natural way with missing data. In actual data, it may often occur that some of the data are missing and hence some of the tests that are normally applied can not be executed. If for instance, no tests can be applied then the numerical quality label equals the a-priori assignment. If a test can be applied and the outcome is "0" (i.e. the test is passed) then the numerical quality label increases which is in correspondance with the intuitive feeling that there is more confidence on this datapoint since it is "confirmed" by other data.

The fact that in the Bayesian scheme the outlying character that is found in 1 test can be compensated by the normal character that is noted in another test is better appreciated through the following example. Suppose that two tests are applied. We further assume that the information matrices of the two tests are identical.

The a-posteriori probability that follows if one test fails while the other does not fail can be calculated by sequential application of Equation (1). For this specific example, it is however more easily calculated by considering the information matrix for the multivariate outcome of the test as is shown in Table 2.

Test result	0/0	1/0	0/1	1/1
"normal"	α_0^2	$\alpha_0(1-\alpha_0)$	$\alpha_0(1-\alpha_0)$	$(1-\alpha_0)^2$
"abnormal"	β_0^2	$\beta_0(1-\beta_0)$	$\beta_0(1-\beta_0)$	$(1-\beta_0)^2$

Table 2 - "Information" matrix for the multivariate outcome of two tests with binary outcomes (0 or 1) and identical information matrices

A measurement that fails one test (outcome 1) while passing another test (outcome 0) would then be assigned the a-posteriori quality indicator:

$$q=\frac{p\alpha_0\left(1-\alpha_0\right)}{p\alpha_0\left(1-\alpha_0\right)+\left(1-p\right)\beta_0\left(1-\beta_0\right)} \tag{2}$$

The a-posteriori value q equals the a-priori probability when $\alpha_0(1-\alpha_0)=\beta_0(1-\beta_0)$ or $\alpha_0=1-\beta_0$. For instance if outcome 0 is expected with 99% chance for "normal" data and only with 1% chance for "abnormal" data, then the "charging" evidence of one test is neutralized by the "favoring" evidence of the other test.

This possibility to take both information in favor of normal behavior and information pointing towards abnormal behavior into account is crucial to a good detection scheme. For instance, in a dataset where multiple variables are measured simultaneously it will often occur that only one of the variables is anamolous, while the others are normal. When applying the tests, the normal variables will however also appear abnormal when compared to the anomalous measurement. However, because in the Bayesian scheme both evidence in charge and evidence in favor is accounted for, only the variable that shows anomalous behavior relative to different variables will be assigned a lower quality label than has been previously assessed.

The previous example is also interesting because it shows that when the values β are set equal to the complement of α, then the outcome "'pass" in one test is counteracted by the outcome "fail" in another test and the final a-posteriori probability is equal to the a-priori probability. This may however not be necessarily desirable and to establish a good test scheme some calibration of the values of α and β is undoubtly necessary to obtain reasonable results.

The final numerical quality indicator should therefore not be seen as a precise indication of the fraction of points that are indeed "a-typical", but as a ranking indicator of the degree of validation of the measurement. Values of q larger than the initial value indicate that the statistical investigation weighs in favor of a normal behavior, while values less than the initial value p would indicate the detection of one or more anomalies that in combination are judged to weigh in favor of a-typical behavior. The individual test results are of course also preserved and thus allow detailled investigation of the reason for the assignment of an eventually low value. Furthermore it is possible to extract those data that show a specific test pattern and through further investigation of such data to gain further understanding of the actual reason for the anomaly.

Practical Implementation of the Data Flagging Scheme

To numerically implement the data flagging scheme into an operational software some constraints must be made on the manner in which the tests are applied. These constraints and the practical application of the test scheme are explained next.

For a given datatype (that is a collection of different measurement types) the measurements are considered as an ordered datastream that is sequentially passed through the tests. Ordering would be normally according to the time of measurement but also other ordering schemes can be conceived (i.e. for spatial data a certain spatial ordering scheme may be used). The "oldest" data are by convention treated first.

The tests that are applied use either information that is stored in the database (i.e. data that have been previously validated) or data in the datastream that have already passed the tests. To arrive to an algorithm that can be implemented without requiring excessive storage requirements, the different tests are grouped according to the timewindow that is used: i.e. tests that consider only the data in a single record (and thus simultaneously measured) have a timewindow of length 1 and require only storage of the measurements in a single record; tests that consider data in multiple records (i.e. because a temporal difference or another statistic is used that depends on multiple records) require storage of multiple records.

The overall organization of the data flagging scheme then corresponds to:

1. enter a new record and update the values in the time-window buffer storing previous results;
2. assign a-priori probabilities in accordance with the measurement type;
3. apply the different tests, where a test is characterized by following attributes:
   - the variable to be tested;
   - the algorithm to be applied leading to the a discrete outcome;
   - the information matrix.
4. knowing the test results for the different measurement variables the a-priori p value is updated to a value q

As noted earlier, the final result is invariant with respect to the ordering of the tests in such a "test scheme". However, in some cases it may be desirable to apply first a subset of tests, update the value of q and then consider measurements with very low value of q as unreliable for the further execution of tests. In such a case, the end result is of course dependent on the order in which the tests are applied. This enhancement is incorporated by grouping tests into "stages". Within each stage all tests are applied, after which the a-priori probability is updated. Next a filter is applied that changes values with a value of q below a treshold to "missing" values such that no further testing is applied to these data and, more importantly, they are not used to further test other measurements. When applying the following stages, the a-posteriori probability that is obtained from application of the previous stages is used as an a-priori probability. The individual adjustments from the different stages are however also maintained, as well as the individual results of the tests.

4.2 Implemented Functionalities

The previous conceptual framework has been implemented into the SQC program which is described in detail in the user manual shown in Appendix. Following functionalities have been implemented:

1. the declaration of measurement types (i.e. significant waveheight, …) eventually further broken down according to the type of instrument or data provider;
2. the declaration of data types (i.e. a group of measurement types that are delivered in a single dataset to be tested);
3. the declaration of individual test schemes: i.e. length of time window, algorithm and information matrix;
4. the declaration of test schemes: i.e. declaration of stages, the tests that need to be applied in each stage and the filtering of data between stages;
5. the visualization and reporting of test schemes;
6. the export of the quality validation results such that it can be incorporated into the IDOD database.

For the individual test schemes the three generic types of tests indicated earlier in Section 4.1 are considered (1. distribution test, 2. regression test, 3. spatial test). The specific values for application to a given dataset can be derived through application of the corresponding SAT program and are stored in a buffer that is directly accessed by the SQC software.

Included in the SQC manual is an example of application that illustrates how a specifc test scheme can be developed, declared and applied to the data to flag "a-typical" measurements.

5. Summary

The contribution of UCS to the IDOD project has resulted in two software tools: SAT that allows users to perform statistical analysis through the web on the IDOD database allowing for an intelligent interpretation of the data; SQC that allows the datamanager to construct test schemes that can be used to validate the data and assign quality labels that indicate the degree to which a measurement conforms with the normal statistical pattern.

The two tools contribute significantly to the "value" of the databank and it is expected that they will promote the actual use of the databank by practitioners.

References

Plevoets, B., June 1999, "Semi-variogram estimation from sparse point data (using generalized linear models techniques)", M.Sc. Thesis, ISPS program, KULeuven

REPORT FROM BELGIUM ON THE FIRST APPLICATION OF THE COMPREHENSIVE PROCEDURE.

This report constitutes the Belgian national report on the initial application of the Comprehensive Procedure, making use of the Common Assessment Criteria for the Eutrophication Status of the OSPAR Maritime Area as agreed by OSPAR in 2002.

The present document has been processed and put together by the Management Unit of the North Sea Mathematical Models on the basis of the monitoring data and with the support of the IZEUT (Identification of maritime Zones affected by EUTrophication) project, commissioned by the Federal Office for Scientific, Technical and Cultural Affairs[1].

The structure of the report follows the agreed reporting format :

1. Area
2. Description
3. Assessment
4. Classification
5. Discussion
6. Other information

1. AREA

The area covered by this report is the Belgian Continental Shelf (BCS).

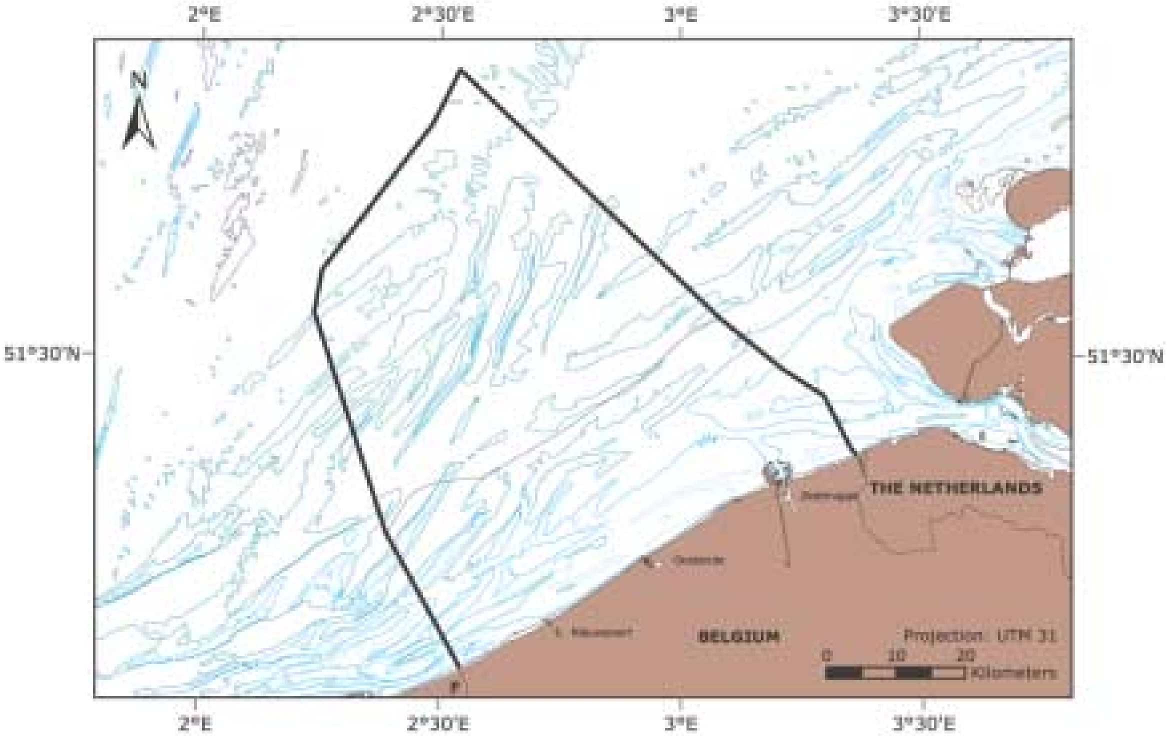


Figure 1 : Belgian Continental Shelf (BCS), contour in bold. The fine line indicates the limit of the territorial waters.

[1] IZEUT (Identification of maritime Zones affected by EUTrophication) project; contract number MN/11/93.

2. DESCRIPTION

The Belgian continental shelf (BCS) is located in the southeastern part of the Southern Bight of the North Sea (Figure 1). It is bordered in the south by the Straits of Dover and in the north by the Dutch coastal zone. The surface of the Belgian continental shelf (BCS), which corresponds to the Exclusive Economic Zone (EEZ), is 3500 km^2. The overall length of the Belgian coastline is 67 km. It is a relatively shallow area with the deepest water of 40 m found in the northwest. The 10-meter isobath runs along the coast from southwest to northeast. The area is characterized by extensive sandbanks ten of kilometers long, a few kilometers wide and with height of 20-25 m.

The tidal motion, in particular the semi-diurnal M2 component, is the dominant energetic feature of the hydrodynamics of the BCS. It induces strong along-shore currents (e.g. 0.6-1.1 m s^{-1}). Shallow depths and strong tidal currents ensure an almost permanent complete mixing of these coastal waters. The physical oceanography of these shallow waters is also considerably influenced by winds (mainly southwesterly) which induced significant residual currents (Verlaan & Groeendijk, 1993).

The main residual circulation pattern of water masses results in a northeastward flow of Atlantic waters through the Dover Strait. The BCS is a highly dynamic system characterised by waters resulting from the mixing between the inflowing Atlantic waters and freshwater inputs from the rivers Yser, Scheldt and Rhine. The geographical extent of river plumes varies mainly as a function of wind, with tidal effects giving smaller scale modulations. The wind is in turn related to the North Atlantic Oscillation (NAO) index.

3. ASSESSMENT

Although monitoring data are available up to the year 2001, the present assessment has been carried out on the basis of data for the years 1999, 1998, 1997 and 1995. This appears to be the best combination between recent data and a rich data set. Only Quality Assured data were used.

This section of the national report is illustrated for the most recent of these four years: 1999.

3.1 Degree of Nutrient Enrichment (Category I Assessment Parameter)

3.1.1 Trend analysis

Figure 2A shows the evolution of nutrient enrichment (ammonium, nitrate+nitrite, inorganic nitrogen, phosphate and silicic acid) of Belgian coastal waters over the 1974-2001 period calculated for an average salinity of 33.5 (Rousseau *et al*, 2002). The standard deviation has been calculated according to Scherrer (1984).

No significant trend in the evolution of dissolved inorganic nitrogen (on average 29 µM) is observed over the 1974-2001 period. Interestingly, after 1985, a significant opposite change between the forms of inorganic nitrogen (ammonium and nitrate+nitrite) is observed, which coincided with similar transformations of nitrogen forms in the upper Scheldt estuary (Van Damme *et al.*, submitted). Statistical analysis shows that silicic acid decreases slightly but significantly (at $p < 0.05$) over the 1974-2001 period.

Most spectacular is the decreasing trend in phosphate winter concentration ($p > 0.005$) from ~2µM in the 1974-1984 period to 0.8 µM nowadays. This trend corresponds to the clear decrease of phosphate concentration recorded in the upper estuary since the mid 1980's (Van Damme *et al.*, submitted) which itself was due to the reduction in P emission (attributable in particular to the lower content of P in washing powder).

The DIP vs salinity regression was statistically compared with similar calculations conducted on available "French waters" data sets when available. It is assumed that a significant difference between slopes would indicate a difference in nutrient loads to the two coastal areas. Reversibly, similar slopes would reflect a similar nutrient enrichment along the aquatic continuum eastern Atlantic-North Sea waters. It is quite interesting to note that nutrient-salinity slopes were not statistically different for French and Belgian areas except for phosphate enrichment prior to 1980. During this period the phosphate enrichment of French coastal waters was significantly higher possibly reflecting a higher use of phosphate in washing powder.

The contrasted changes in phosphate and nitrogen winter concentrations altered the N:P:Si balance of nutrients available to phytoplankton growth. The extent of this change can be appreciated on figure 2B which compares the trends of N:P, Si:P and N:P molar ratios in winter over the 1974-2001 period with nutrient requirements of coastal phytoplankton (Redfield *et al.*, 1963) and diatoms (Brzezinski, 1985). One major observation is the important shift of N:P ratios from values close to the Redfield ratio during the period 1974-1985 to nitrogen excess values after the mid-1980's. The largest changes were observed during the 1990's when N:P increased from values around 20 to more than 30 from 1992 onwards. This suggests that phosphate is nowadays the nutrient limiting phytoplankton growth. Note that light could also be a limiting factor.

On the other hand the inter-annual evolution of N:Si and Si:P molar ratios gives some information on the evolution of diatom nutrient limitation. Clearly nitrogen availability largely exceeded the silicon availability for diatoms during the whole period. More interesting is the evolution of the Si:P ratio which clearly indicated silicon limitation during the 1974-2001 period. However, since mid-1995, the system looks more balanced with respect to Si and P, explained by the decrease in P emission.

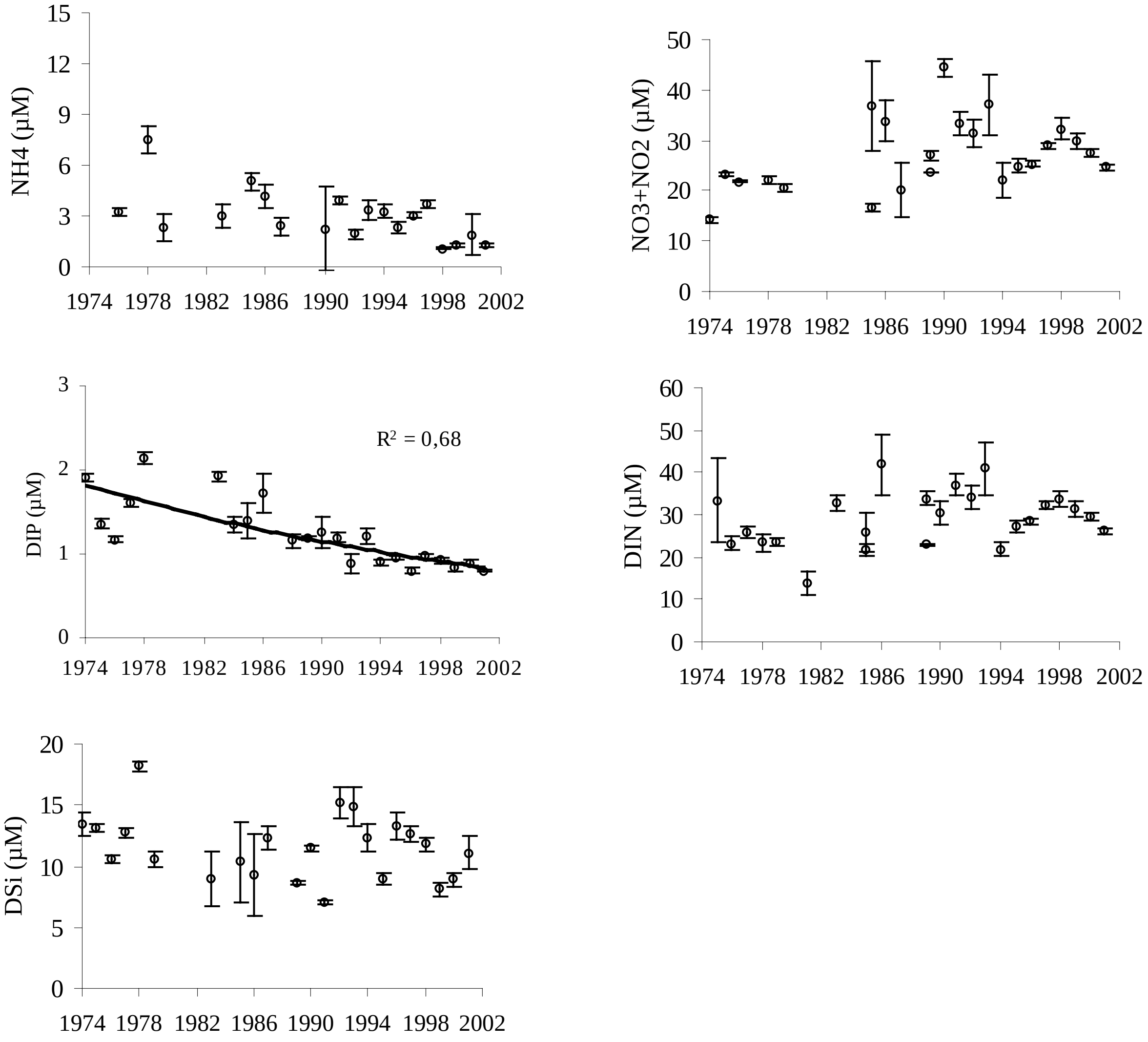


Figure 2A: 1974-2001 evolution of winter nutrient concentrations at salinity 33.5 (Rousseau *et al*, 2002).

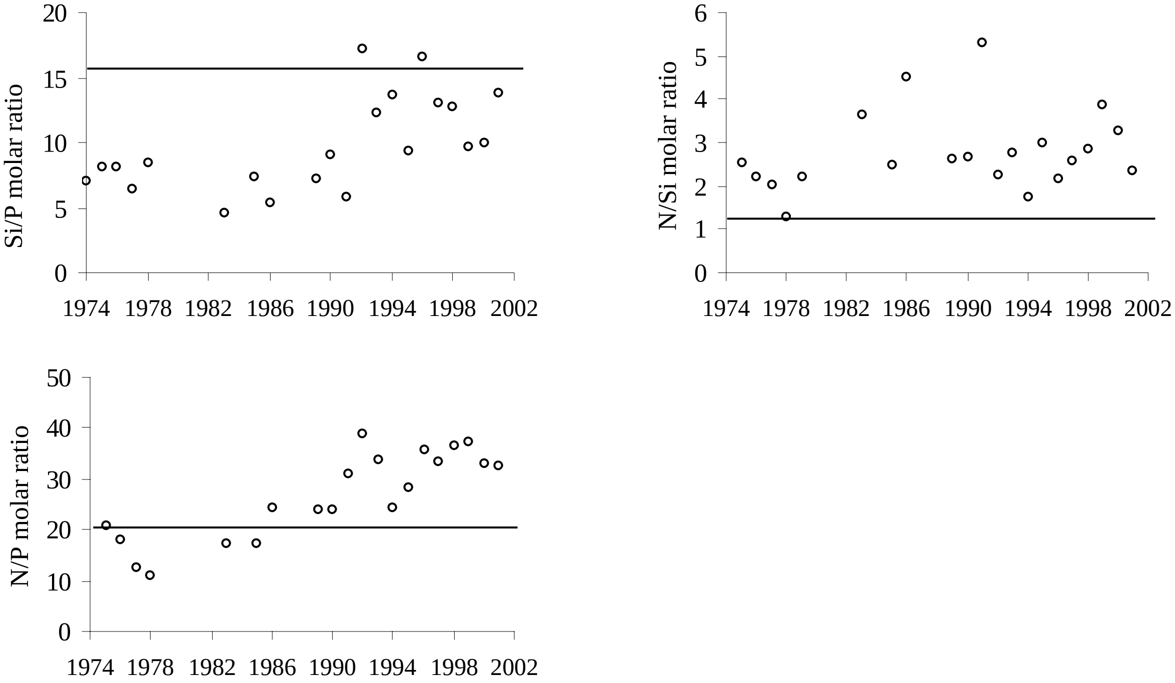


Figure 2B: Nutrient molar ratios calculated for the Belgian coastal waters over the period 1974-2001. The horizontal lines indicate N:P phytoplankton (Redfield *et al.*, 1963) and N:Si & Si:P diatom (Brzezinski, 1985) stoichiometry (Rousseau *et al*, 2002).

3.1.2 Transboundary nutrient fluxes

The BCS is directly influenced by in-flowing Atlantic waters themselves enriched by freshwaters of the rivers Seine (mostly) and Somme. The residual circulation of water masses in the Channel follows a SW-NE axis along the Southern Bight of the North Sea under the combined effects of tides, dominant winds and average sea level (Salomon and Breton, 1991).

The intensity and nutrient concentration of water inflow from the Channel to BCS varies considerably according to meteorological conditions, mainly the wind stress itself driven by the North Atlantic Oscillation (NAO), the dominant mode of climate variability in northern Europe (Breton *et al.*, submitted).

It is therefore recommended to consider the SW transboundary flux of nutrients, in addition to the “local” inputs to the BCS, in any interpretation of the present assessment.

3.1.3 Atmospheric deposition

Atmospheric deposition of nutrients to the BCS has been estimated, based on a literature review. Table 1 gives estimations of atmospheric deposition of N in the BCS.

The values for N vary from 1.29 to 2.64 kT yr^{-1}. Nitrogen oxides (NO and NO2-) are the main N species deposited over the Southern Bight of the North Sea. Fluxes calculated or modelled per surface unit are lower for the North Sea than those of the Southern Bight. This is linked to the gradient of atmospheric gas and particules concentrations from the coast to offshore. From this review, an average flux of 2 kT y^{-1} of N can be considered for the atmospheric depositions over the BCS. This corresponds to about 4 % of the direct riverine inputs of N (as reported under the OSPAR RID programme).

A value, for P, of 0.095 kT y-1 is considered as specific to the continental coastal waters of the North Sea and is thus applicable to the BCS (Nelissen and Stefels, 1988). This corresponds to less than 2 % of the direct riverine inputs of P (as reported under the OSPAR RID programme).

Table 1: Atmospheric deposition of nitrogen to the North Sea and Southern North Sea related to the BCS area. The Southern North Sea represents the zone between the Straits of Dover and 56°N.

Zone	**Nutrient**	**Flux/area** **TN $km^{-2}y^{-1}$**	**Deposition to BCS** **10^3 TN y^{-1}**	**Method**	**Reference**
North Sea	$NO + NO_2 + NH_3 + NH_4^+$	1.311	2.644	Measured	Brockmann et al., 1988
North Sea	$NH_3 + NH_4^+$ $NO + NO_2$ Total	0.270 0.380 0.650	0.545 0.767 1.312	Modelled	vanJaarsveld, 1992
North Sea	$NO + NO_2 + NH_3 + NH_4^+$	0.337	1.295	Measured (1987-1995)	OSPAR QSR, 1987 OSPAR-CAMP, 1998
Southern North Sea	$NH_3 + NH_4^+$ $HNO_3 + NO_3$ DON Total	0.369 0.587 0.035 0.992	0.745 1.185 0.071 2.001	Measured	Rendell et al., 1993
Southern North Sea	$NO + NO_2 +$ $NH_3 + NH_4^+$	0.957	1.929	Measured	Nelissen & Stefels, 1988 (from OSPAR QSR 1987)
Continental coastal waters	$NO + NO_2 + NH_3 + NH_4^+$	1.039	2.096	Measured	Nelissen & Stefels, 1988 (from OSPAR QSR 1987)

3.1.4 Winter nutrient concentrations

Winter nutrient concentrations were measured at monitoring stations of the national grid by MUMM. The geographical distribution of winter DIN and DIP concentrations in 1999 (figures 3 and 4) shows a decreasing gradient from coastal to offshore waters although very few offshore stations were visited due to bad weather conditions. A decreasing gradient from the Scheldt mouth to the French-Belgian border is also visible.

Maximum DIN concentration recorded in 1999 is 126.6 µmol while that of DIP is 1.5 µmol l^{-1}. Offshore concentrations were 12 µmol DIN and 0.49 µmol l^{-1} DIP. For DIN, all stations but one (offshore) show high winter concentration (higher than 15 µmol l^{-1}). For DIP, most of the monitoring stations show high winter concentration (higher than 0.8 µmol l^{-1})

Silicate (figure 5) shows high values close to the Scheldt mouth and a decreasing gradient northwestwards. Concentrations below 4µmol l^{-1} are observed in the west part of the BCS. This can be explained by the low silicate in-flowing Atlantic waters.

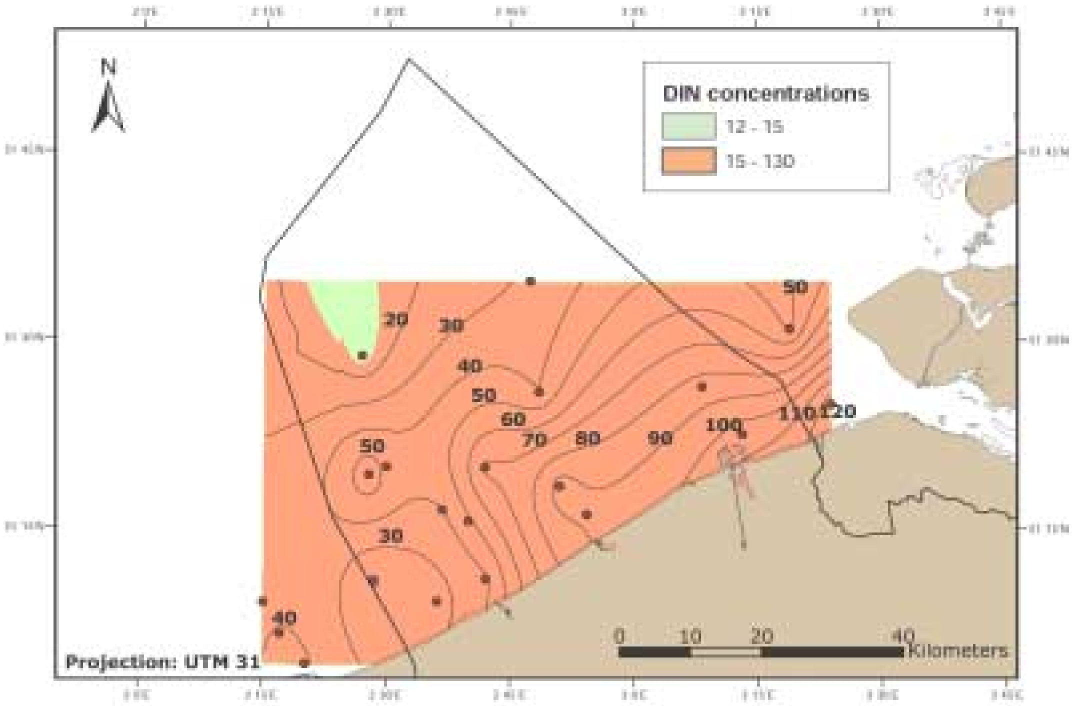


Figure 3: Distribution of winter DIN concentrations (in µmol l^{-1}) over the BCS in 1999. The monitoring stations are indicated.

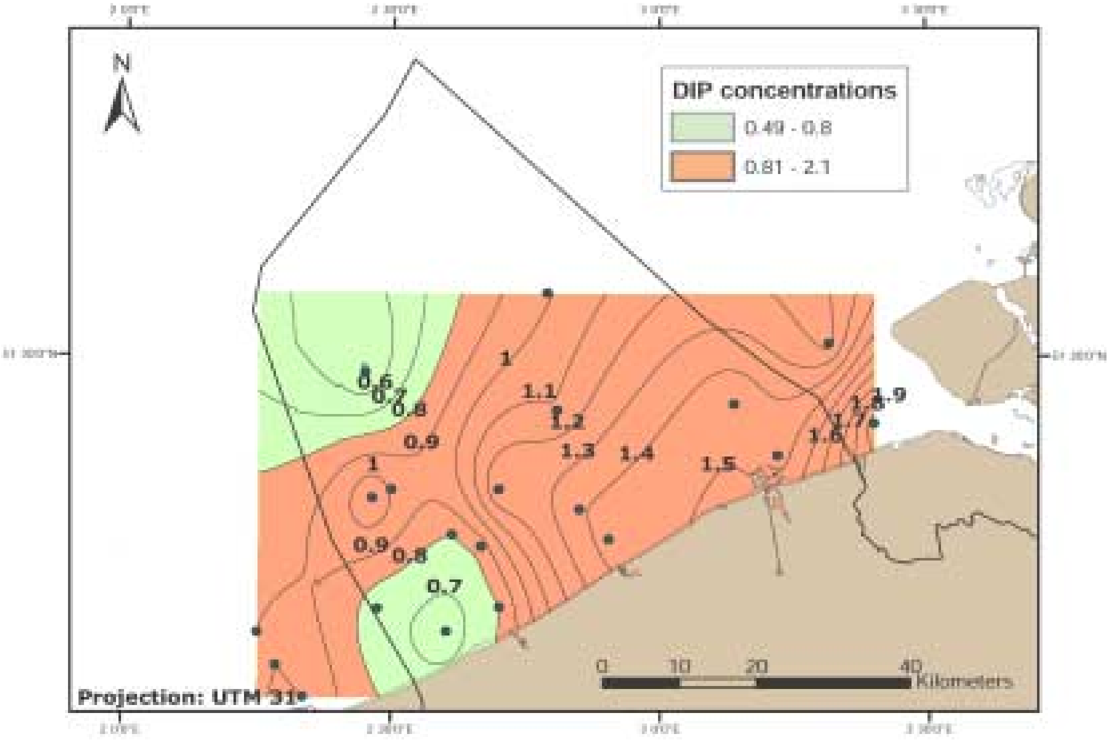


Figure 4: Distribution of winter DIP concentrations (in µmol l^{-1}) over the BCS in 1999. The monitoring stations are indicated.

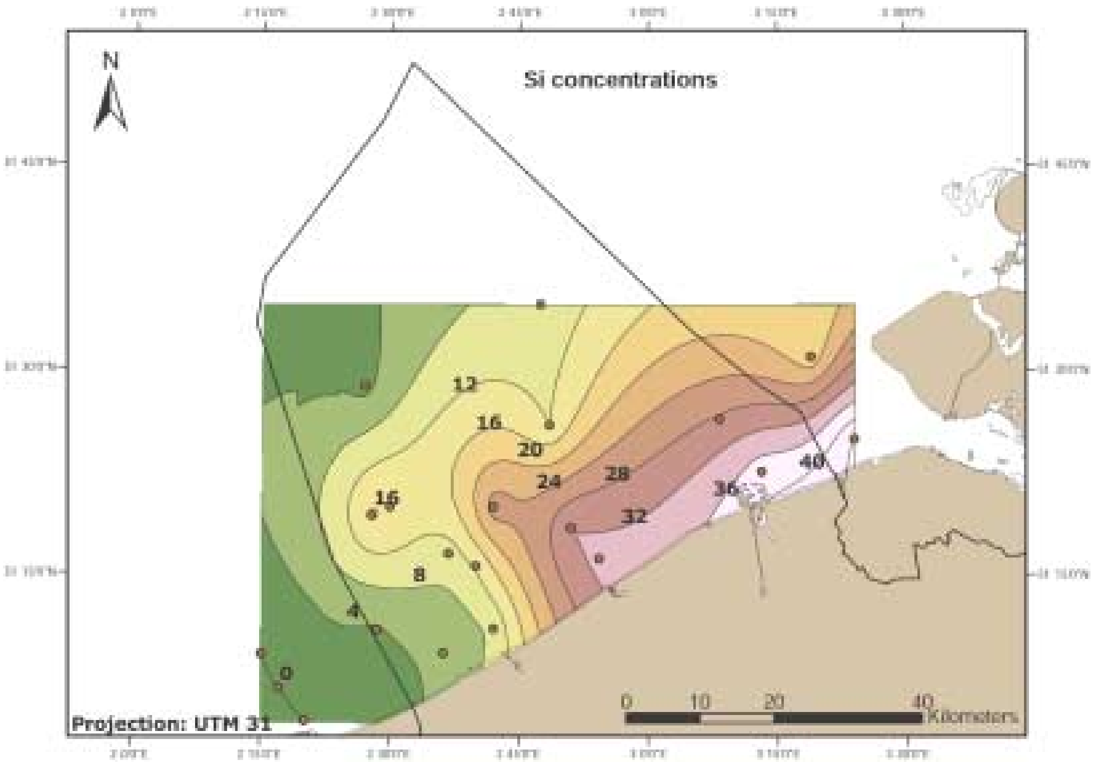


Figure 5: Distribution of winter Si concentrations (in µmol l^{-1}) over the BCS in 1999. The monitoring stations are indicated.

3.1.5. Nutrient ratios

Winter nutrient ratios indicate, for the BCS, a large excess of DIN over DIP and Silicate compared to phytoplankton stoichiometry over the whole BCS. Figure 6 shows that all the N/P ratios are higher than the reference value of 16 (Redfield *et al.*, 1963). Maximum values are located offshore of Oostende and Zeebrugge. The N/P ratio progressively decreases towards offshore waters.

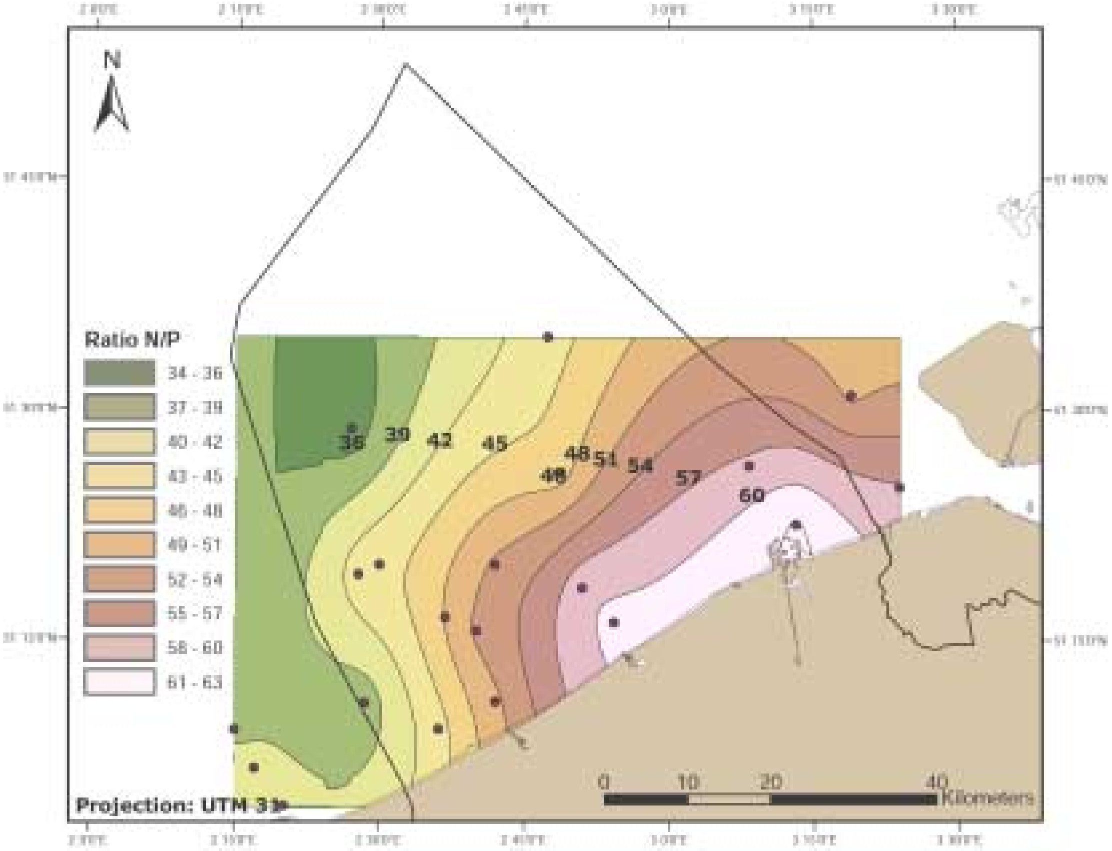


Figure 6 : mapping of the N/P ratio in winter 1999 on the BCS

Further analysis of available data indicates an excess of DIN over Silicate with respect to coastal diatom requirements (i.e. N/Si= 1; Brezinski, 1985). The larger excess occurs in the shallow coastal waters at the French-Belgian border and in offshore waters. The lowest values are recorded in areas of high DIN concentrations. This suggests higher silicate availability close to river and harbour channel discharges.

With respect to the P/Si ratio required for diatoms (P/Si = 0.063; Redfield *et al.*, 1963; Brezinski, 1985), a large excess of DIP over Si appears at the French-Belgian border as well as in more offshore waters. Well-balanced P/Si ratios are recorded in the center of the BCS but DIP deficit occurs in the area close to the Oostende harbour channel.

Altogether, this analysis indicates that the coastal area near the French-Belgian border and the offshore waters are characterised by silicic acid deficiency with respect to N and P.

3.2 Direct Effects (Category II Assessment Parameter)

3.2.1 Maximum and mean Chlorophyl a concentrations

Because of the transient feature of Phaeocystis blooms, only, Chl-a concentrations measured during the Phaeocystis spring bloom in April-May were considered. The period for maximum Chl-a levels was determined by the Chl-a concentrations recorded during the weekly sampling at a station located in the centre of the BCS (figure 7). Because of the limited size of the BCS, it is assumed that this weekly monitoring allows us to identify the start of the spring phytoplancton bloom.

Figure 7: seasonal variation of Chl a concentration at station 330 in the middle of the BCS – weekly monitoring.

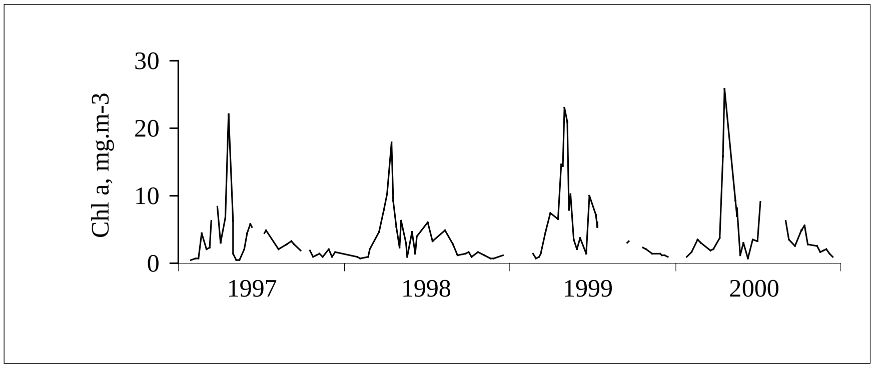


Figure 8 indicates much higher Chl-a concentrations (up to 42.6 µg Chl-a l^{-1}) at coastal stations with maximum values close to the Scheldt mouth as well as in front of the Oostende and Nieuwpoort channels.

Generally there is a progressive decrease of Chl-a concentrations along a coastal-offshore gradient with the lowest concentrations recorded close to the Belgian-Dutch border in the less-enriched marine waters (8.06 µg Chl-a l^{-1}).
The 15µg/l treshold (OSPAR criteria, see also item 4 Clasification) has been represented.

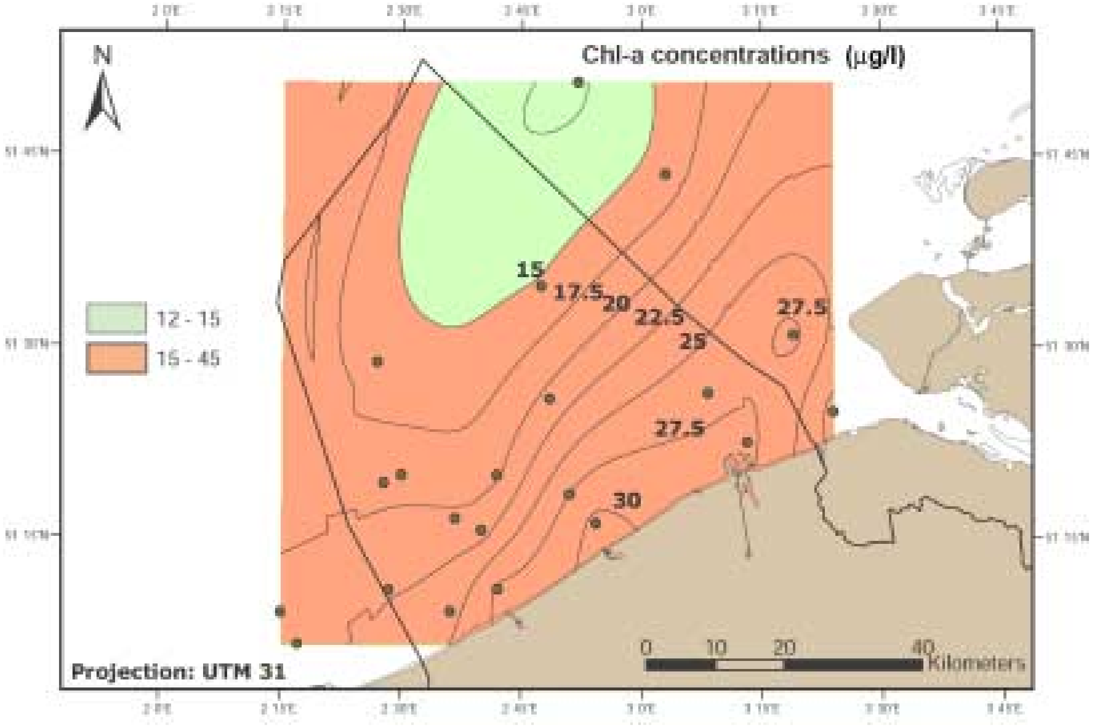


Figure 8: distribution of Chl-a concentration (in µg/l) in spring 1999. The monitoring stations are indicated.

3.2.2 Region specific phytoplankton indicator species

Current knowledge reports that Phaeocystis colony blooms constitute the eutrophication-related problems in Belgian coastal waters and more generally in the eastern Southern Bight of the North Sea. The success of this non-siliceous phytoplankter results in both its ability to use anthropogenic nitrate in P-regenerated conditions (Rousseau, 2000) and its "resistance" to grazing (Gasparini et al., 2000).

Most of the adverse impacts of Phaeocystis colony blooms in the Southern Bight of the North Sea concern biological resources, the fishing industry and tourism which are of obvious socio-economical interest in the area. These have been reported as deposits of foam on the beaches or as clogging of fishing nets as a consequence of huge accumulation of ungrazed gelatinous Phaeocystis colonies. Reports of these damages are however mostly anecdotal and the loss for tourism and fishing industry of the region is not established. These blooms are not directly detrimental to human health as are other toxic blooming algae.

Foam accumulation on the beaches is the only visible negative impact of Phaeocystis blooms in the Southern Bight of the North Sea.

Phaeocystis-related phenomena have been suggested as affecting negatively the benthic community, fish production and the quality of the beaches although such effects have never been demonstrated**.**

3.2.3 Macrophytes

In contrast with neighbouring countries no macroalgal mats have been recorded in Belgium.

3.3 Indirect Effects (Category III Assessment Parameter)

Most of the organic matter derived from ungrazed colonies is remineralized in the water column by intense bacterial activity (Rousseau *et al*., 2000). Although this could create locally some transient oxygen depletion, no serious oxygen problems have been reported in Belgian coastal waters in particular and more generally in the eastern part of the Southern Bight of the North Sea. The prevailing hydrodynamical conditions (strong tide and wind mixing) ensure continuous oxygen replenishment.

3.4 Other Possible Effects (Category IV Assessment Parameter)

Besides Phaeocystis-related problems, neither oxygen depletion nor human diseases due to algal toxins have been recorded to date in the Southern Bight of the North Sea. However, recent reports on P. pouchetii suggests that this impact can not be totally ignored.
For the time being aquaculture/mariculture is a minor activity throughout the BCS and no research activity on toxic algal blooms has been carried out.

4 CLASSIFICATION

From the previous section of this report, it can be assumed that the most relevant assessment parameters are winter DIN & DIP with regard to the degree of nutrient enrichment and Chl-a with regard to the direct effects of nutrient enrichment.

The following thresholds are applied to these parameters:

- For DIN, elevated concentrations are defined as those greater that **15 µmol/l**.
- For DIP, elevated concentrations are defined as those greater than **0.8 µmol/l**.
- For Chl-a, a background Chl-a concentration was determined for the application of phytoplankton assessment criteria for the BCS. Eutrophication in the BCS occurs in the form of massive, ephemeral Phaeocystis colony blooms during spring (April-May). Phaeocystis blooms are responsible for foam accumulation on the beaches, the only visible effect of ecosystem disturbance. Phaeocystis colonies generally occur simultaneously with diatom (mainly Guinardia spp.) blooms that considerably affect the Chl-a concentrations. Analysis of historical data (ASMO 98/3/Info.1-F) suggests that foam accumulation would occur from a

Phaeocystis cell concentration of 10^7 l^{-1}. This cell concentration can be converted into a Chl-a concentration of 0.5 µgChl-a l^{-1} (C:Chla=29) using the experimentally determined factor of 0.5 pgChl-a /Phaeocystis cell (Rousseau *et al.*, 1990).

The contribution of diatoms to the spring bloom has been established from the analysis of a 13 year time series at a monitoring station (330) located in the centre of the BCS. Seasonal evolution of Chl-a concentrations measured on a weekly basis indicates that the average Chl-a concentration of 9.2 µg l^{-1} corresponds to the Phaeocystis pre-bloom situation, itself determined as a Phaeocystis concentration higher than $1x10^6$ cells l^{-1}(Cadée & Hegeman, 1991).

This allows us to consider a threshold of **15 µgChl-a l^{-1}** from which nutrient over-enrichment would result in ecosystem disturbance (same value as indicated for Netherlands North Sea coastal waters in the Common Assessment Criteria).

The next figures present the results of the first application of the Common Agreed Criteria to the BCS in, respectively, 1999, 1998, 1997 and 1995.

Eutrophication status in 1999

Almost all the covered area (figure 9) qualifies as a problem area with regard to eutrophication. The status of the white section could not be determined because of lack of winter DIN/DIP information.

However, from the high offshore DIN values presented in figure 3 and the distribution of Chl-a presented in figure 8, it could be assumed that the eastern half of this white section would qualify as a potential problem area and be added to the presented potential problem area.

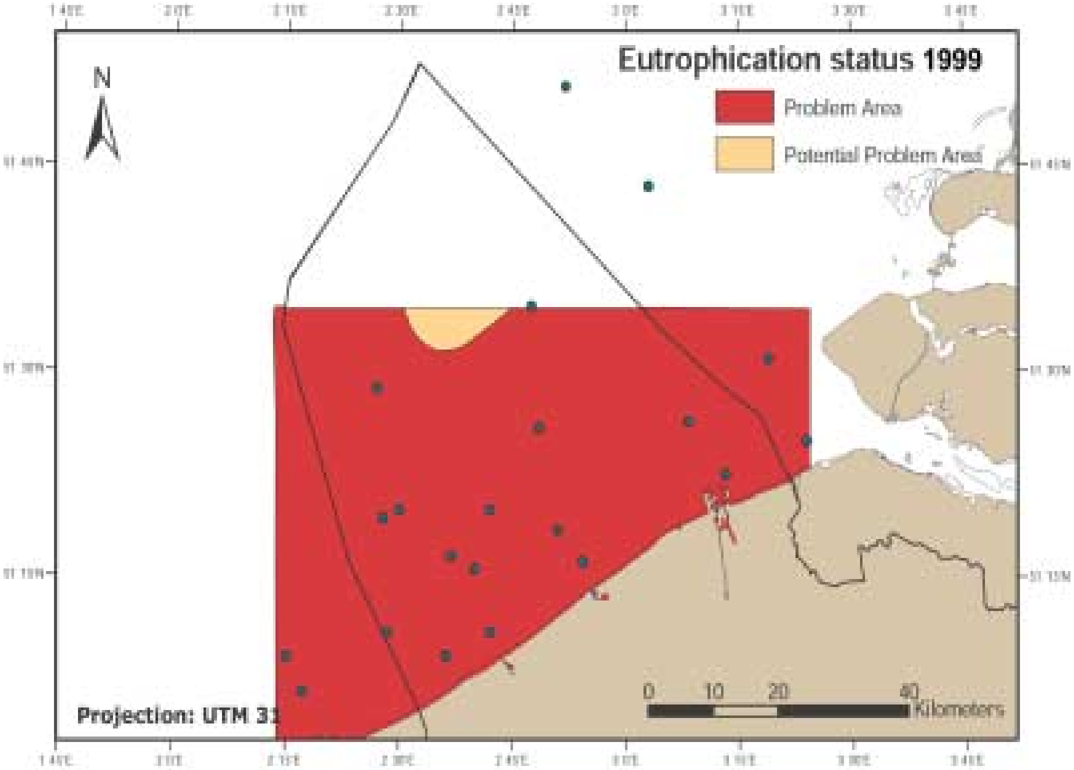


Figure 9: Assessment of the Eutrophication status of the BCS in 1999. Monitoring stations are indicated.

Eutrophication status in 1998

A problem area with regard to eutrophication is visible on figure 10. This problem area is situated offshore of Oostende, Zeebrugge and the Scheldt mouth. It develops further (North) in adjacent marine waters.

The western part of the coastal waters qualifies as a potential problem area.

Two offshore zones appear as non problem areas. The biggest one extends further North in adjacent marine waters.

With regard to the extreme offshore zone presented as potential problem area (beyond the non problem areas), it should be mentioned that there is no monitoring information available that confirms this status. The presented status of this zone is probably due to the extrapolation method applied that gave a relative greater weight to data from the many coastal stations, compared to the few available offshore stations. This will be further discussed in section 5 of this report.

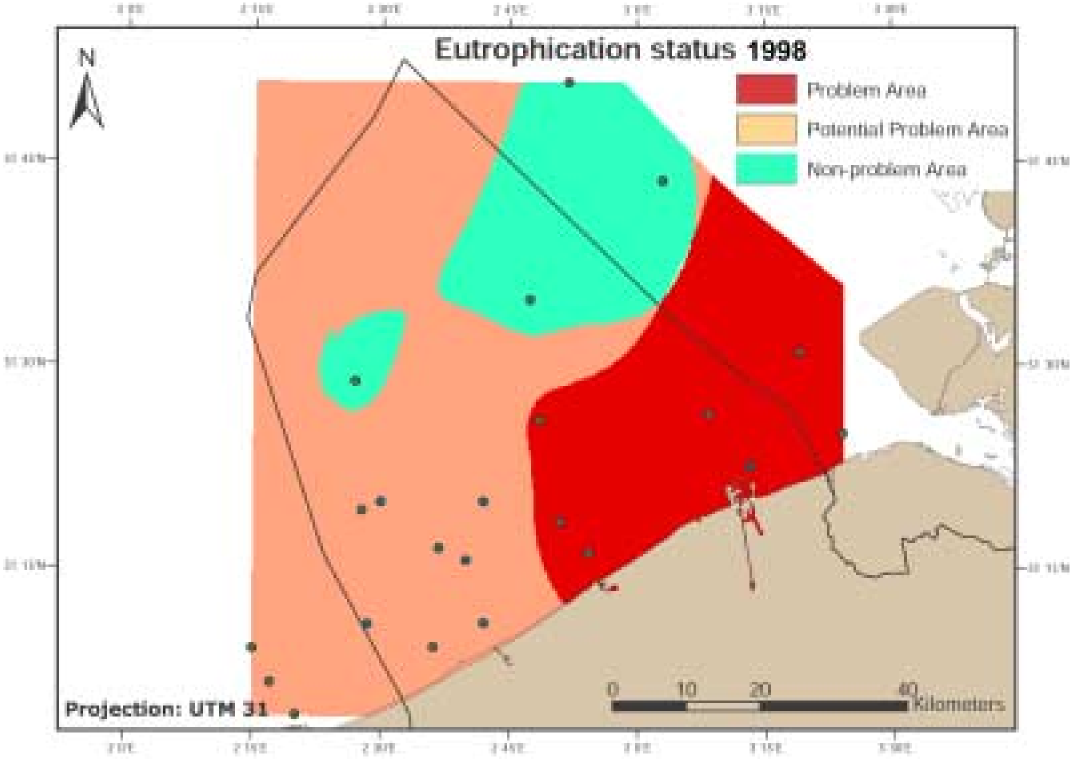


Figure 10: Assessment of the Eutrophication status of the BCS in 1998. Monitoring stations are indicated.

Eutrophication status in 1997

The problem area (figure 11) is situated offshore from Oostende, Zeebrugge and the Scheldt mouth.

Another clear problem area is visible in the South-West.

The rest of the BCS and adjacent marine waters qualifies as a potential problem area.

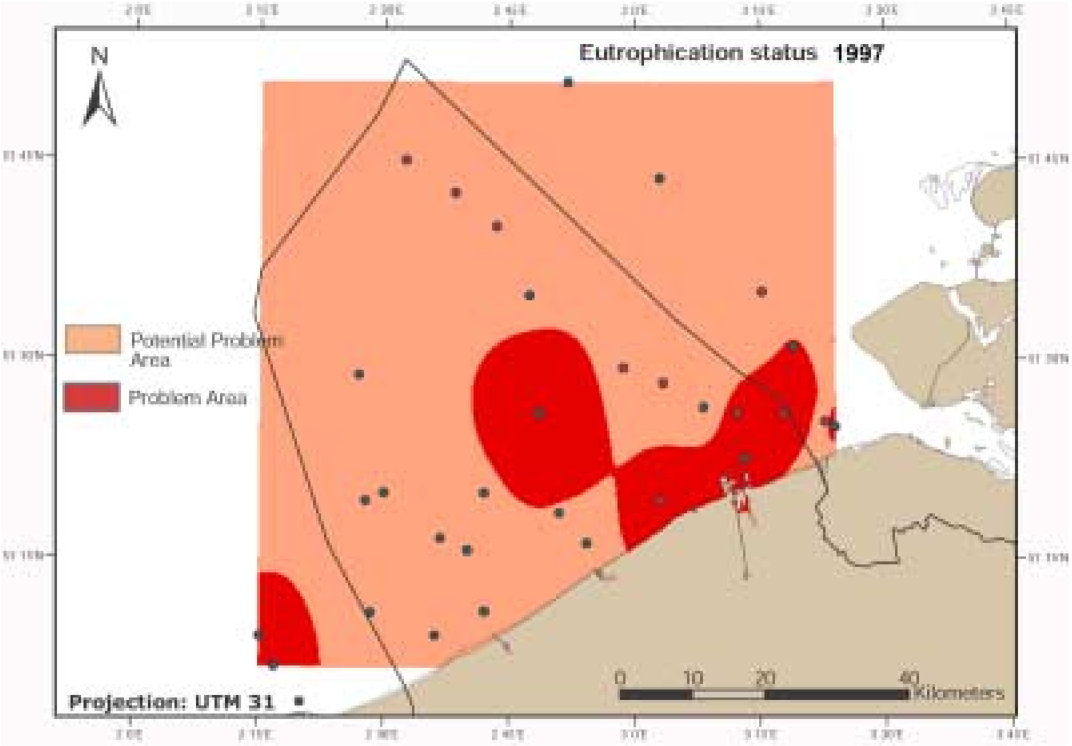


Figure 11: Assessment of the Eutrophication status of the BCS in 1997. Monitoring stations are indicated.

Eutrophication status in 1995

The problem area (figure 12) extends along almost the whole coastline and is mainly developed near Oostende and the Scheldt mouth.

A second problem area is identified in the South-West.

A non problem area is visible in the northern half of the BCS and extends apparently northwards.

The other parts of the BCS and adjacent marine waters qualify as potential problem areas.

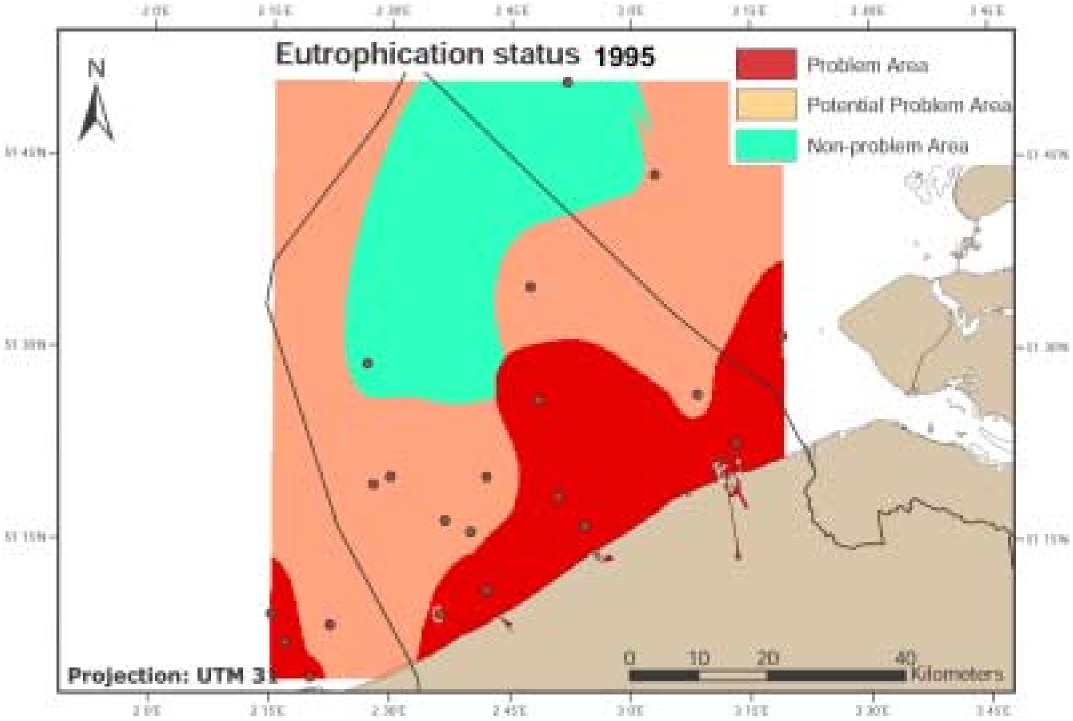


Figure 12: Assessment of the Eutrophication status of the BCS in 1995

Overall assessment and classification

Out of these four recent assessed situations, characteristic features have been demonstrated:

- A problem area is identified for coastal waters, extending from the Scheldt mouth to at least Oostende.
- Another problem area appears South-West in adjacent marine waters.

Depending on the year, the western half of the coastal waters qualifies either as a problem area or as a potential problem area.

Offshore waters can be classified as potential problem areas or non problem waters.

5. DISCUSSION

In almost all situations, winter DIN and/or DIP concentrations (Category I assessment parameter) were above the thresholds. Therefore, the Chl-a concentrations (and the subsequent distribution pattern) have been determined for the identification of the eutrophication status.

Although a strict selection of Chl-a data has been carried out for the purpose of this assessment (see 3.2.1), it should be noted that variations of Chl-a concentrations can appear due to the timing of the sampling event, light availability, phytoplankton community species composition, hydrodynamical conditions and physiological state.

These elements are most probably at the reason for the strong inter-annual variations in observed maximum Chl-a concentrations and their geographical distribution. Maximum Chl-a levels recorded in the BCS in spring 1995 1997, 1998 and 2000 were respectively 38.8, 52.5, 42.6 and 43.5µg/l.
For the same period, Chl-a concentrations recorded in offshore waters fluctuated from 2 to 5 µg/l.

With regard to Phaeocystis, the fact that it is able to use organic sources of phosphate, suggests that organic nutrient data should be acquired to supplement the existing inorganic nutrient monitoring.

The current national monitoring grid is characterised by a high density of stations in the coastal waters and a lower density in offshore waters. This distribution of sampling stations combined with a large range of Chl-a values has been somehow a limiting factor in the assessment of offshore waters and was documented in the 1998 assessment.

A further identification of zones affected by eutrophication would justify additional offshore sampling in the national programme and certainly has to be carried out by putting together data sets from adjacent countries (France, Netherlands, UK).

This improvement and widening of the monitoring network(s) will allow the integration of transboundary fluxes of nutrients thereby contributing to a better understanding of the zones affected by eutrophication.

6. OTHER INFORMATION

Annex 1: references
Annex 2: summary reporting format

ANNEX 1

References

Breton E., Rousseau V., Parent J.Y., Ozer J., Lefebvre A. & Lancelot C. Combined effect of climate and man on diatom/*Phaeocystis* blooms in the eutrophicated Belgian coastal waters (Southern Bight of the North Sea). Submitted to Limn. Ocean.

Brzezinski, M.A. 1985. The Si:C:N ratio of marine diatoms: interspecific variability and the effect of some environmental variables. J. Phycol. 21: 347-357

Cadée, G.C. and J. Hegeman. 1991. Historical phytoplankton data of the Marsdiep. Hydrobiol. Bull. 24: 111-188.

Gasparini, S., M.-H. Daro, E. Antajan, M. Tackx, V. Rousseau, J.-Y. Parent & C. Lancelot. 2000. Mesozooplankton grazing during the *Phaeocystis globosa* bloom in the Southern Bight of the North Sea. *J. Sea Res*. 43: 345-356.

Marshall, J., Kushnir, Y., Battisti, D., Chang, P., Hurrel, J., McCartney, M. and M. Visbeck. 1997. A white paper on Atlantic climate variability. http://geoid.mit.edu/accp/avehtml.html

Redfield, A.C., B.A. Ketchum and F.A. Richards. 1963. The influence of organisms on the composition of sea-water, p 26-77. In: The sea. M.N. Hill (Ed). Wiley.

Rousseau, V., Mathot, S. and C. Lancelot. 1990. Calculating carbon biomass of *Phaeocystis* sp. from microscopic observations. Mar. Biol. 107: 305-314.

Rousseau, V. 2000. Dynamics of *Phaeocystis* and diatom blooms in the eutrophicated coastal waters of the Southern Bight of the North Sea. PhD Thesis, Université Libre de Bruxelles.

Rousseau, V., S. Becquevort, J.Y. Parent, S. Gasparini, M.-H. Daro, M. Tackx and C. Lancelot. 2000. Trophic efficiency of the planktonic food web in a coastal ecosystem dominated by *Phaeocystis* colonies. *J. Sea Res*. 43: (357-372).

Rousseau, V., Breton, E , De Wachter, B., Beji, A., Deconinck, M., Huijgh, J., Bolsens, T., Leroy, D. & C. Lancelot. 2002. IZEUT : Identification of Belgian maritime zones affected by eutrophication. Implementation of the OSPAR Common Procedure to combat eutrophication. Final report.

Scherrer, B. (1984). Biostatistique, 2 éme édition, Gaëtan Morin (Ed), Paris.

Nelissen, P.H.M. & J. Stefels. 1988. Eutrophication of the North Sea. Nederlands Instituut voor Onderzoek der Zee. Report 1988-4.

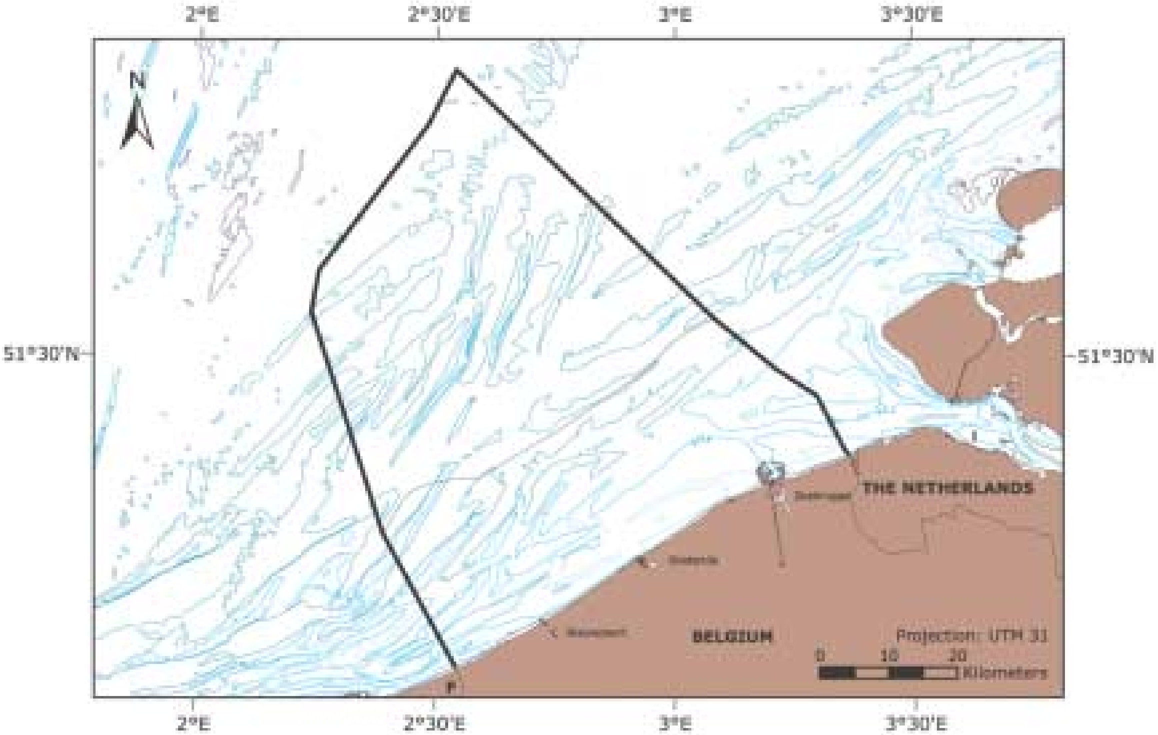
Van Damme, S., T. Ysebaert, P. Herman, M. Starink, L Hellings, M Tackx, O. Van Cleemput, P. Meire. Nutrients and trophical levels in highly polluted Scheldt estuary at a starting point of recovery (Belgium and The Netherlands): a synthesis (submitted; Estuaries).

Verlaan, P.A.J. & F.C. Groenendijk. 1993. Long term pressure gradients along the Belgian and Dutch coast. MAST G8M report DGW-93.045, Rijkswaterstaat Tidal Waters Division.

Reporting format on the results of the OSPAR Comprehensive Procedure

1. Area

Belgian Continental Shelf (BCS)



2. Description of the area

The Belgian continental shelf (BCS) is located in the southeastern part of the Southern Bight of the North Sea (Figure 1). It is bordered in the south by the Straits of Dover and in the north by the Dutch coastal zone. The surface of the Belgian continental shelf (BCS), which corresponds to the Exclusive Economic Zone (EEZ), is 3500 km2. The overall length of the Belgian coastline is 67 km. It is a relatively shallow area with deepest water of 40 m found in the northwest. The 10-meter isobath runs along the coast from southwest to northeast. The area is characterized by extensive sandbanks ten of kilometers long, a few kilometers wide and height of 20-25 m.

The tidal motion, in particular the semi-diurnal M2 component, is the dominant energetic feature of the dynamics of the BCS. It induces strong along-shore currents (e.g. 0.6-1.1 m s-1). Shallow depths and strong tidal currents ensure a permanent complete mixing of these coastal waters. The physical oceanography of these shallow waters is also considerably influenced by winds (mainly southwesterly) which induced significant residual currents (Verlaan & Groeendijk, 1993).

The main residual circulation pattern of water masses results in a northeastwardly flow of Atlantic waters through the Dover Strait. The BCS is a highly hydrodynamic system characterised by waters resulting of the mixing between the inflowing Atlantic waters and freshwater inputs from rivers Yser, Scheldt and Rhine. The geographical extent of river plumes varies mainly as a function of wind, with tidal effects giving smaller scale modulations. The wind is in turn related to the NAO index (Rogers, 1997).

3. Assessment

Category	Assessment Parameters	Description of Results	Score
Degree of Nutrient Enrichment (I)	Riverine total N and total P inputs and direct discharges (RID)	not used for the assessment	
	Winter DIN- and/or DIP concentrations	most of the time above thresholds	+
	Increased winter N/P ratio (Redfield N/P = 16)	not used for the assessment	
Direct Effects (II)	Maximum and mean Chlorophyll a concentration	variable and determinant for the assesment	+/-
	Region/area specific phytoplankton indicator species	not used for the assessment	
	Macrophytes including macroalgae (region specific)	not relevant for the assessment	
Indirect Effects (III)	Degree of oxygen deficiency	not relevant for the assessment	
	Changes/kills in Zoobenthos and fish mortality	not used for the assessment	
	Organic Carbon/Organic Matter	not used for the assessment	
Other Possible Effects (IV)	Algal toxins (DSP/PSP mussel infection events)	not relevant for the assessment	

4. Classification

Category I Causative factors	Category II Direct effects	Category III and IV Indirect effects/ Other possible effects	Classification
+	+		Problem area (see discussion below)
-	+		Problem area (see discussion below)
+	-		Potential problem area (see discussion below)
-	-		Non problem area

5. Discussion

Out of these four recent assessed situations, characteristic features have been demonstrated:

- A problem area appears in coastal waters, extending from the Scheldt mouth to, at least, Oostende.
- Another problem area appears South-West in adjacent marine waters.

Depending on the year, the western half of the coastal waters qualifies either as problem areas or as potential problem areas.

Offshore waters can be classified as potential problem areas with the presence of non problem waters themselves extending Northwards.

In almost all situations, winter DIN and/or DIP concentrations (Category I assessment parameter) were above the thresholds. Therefore, the Chl-a concentrations (and the subsequent distribution pattern) have been determinant in the identification of the eutrophication status.

Although a strict selection of Chl-a data has been carried out for the purpose of this assessment, it should be reminded that variations of Chl-a concentrations can appear due to the timing of the sampling event, light availability, phytoplankton community species composition, hydrodynamical conditions and physiological state.

These elements are most probably at the origin of the strong inter-annual variations in maximum Chl-a concentrations and their geographical distribution. Maximum Chl-a levels recorded in the BCS in spring 1995 1997, 1998 and 2000 were respectively 38.8, 52.5, 42.6 and 43.5µg/l.
For the same period, Chl-a concentrations recorded in offshore waters fluctuated from 2 to 5 µg/l.

With regard to Phaeocystis, the fact that it is able to use organic sources of phosphate would plea for the acquisition of organic nutrient data which would add value to the existing inorganic nutrient monitoring.

The actual national monitoring grid is characterised by a high density of stations in the coastal waters and a lower density in offshore waters. This distribution of sampling stations combined with a large range of Chl-a values has been somehow a limiting factor in the assessment of offshore waters and was documented in the 1998 assessment.

A further identification of zones affected by eutrophication would justify additional offshore sampling in the national programme and certainly has to be carried out by putting together data sets from adjacent countries (France, Netherlands, UK).

This improvement and widening of the monitoring network(s) will allow the integration of transboundary fluxes of nutrients thereby contributing to a better understanding of the zones affected by eutrophication.

6. Other information

-

IDOD - Integrated and Dynamical Oceanographic Data Management

K. De Cauwer, M. Devolder, S. Jans, L. Schwind, S. Scory

IDOD ?

IDOD est un système d'information sur la qualité du milieu marin.

Il est constitué d'une base de données relationnelle et d'outils pour les analyses statistiques et spatiales.

Quelles données ?

La base de données contient des données physiques, chimiques et biologiques mesurées dans l'eau, la biote et les sédiments de la mer du Nord.

Ces données sont mesurées par l'UGMM et par d'autres équipes scientifiques.

Il s'agit, par exemple, des valeurs de paramètres tels que la température de l'eau, la salinité, les concentrations des nutriments, des métaux lourds ou des PCB ou encore d'information sur les populations animales et leur état physiologique.

Toute donnée n'a de vraie valeur que si elle est bien documentée. C'est pourquoi, la base de données conserve également le maximum de « méta-informations » : projet de recherche, temps et position de l'échantillonnage, type d'écosystème, méthode d'analyse, conditions météo, ...

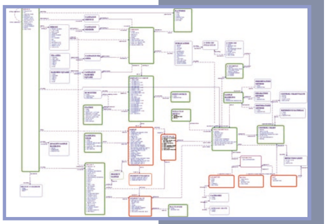
Disponibilité ?

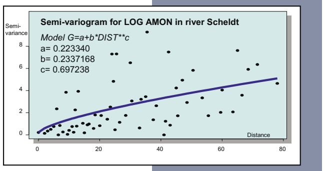
Les données publiques peuvent être consultées *via* notre site web, sur base de critères concernant la position géographique, la période, les paramètres et le format de sortie désirés.

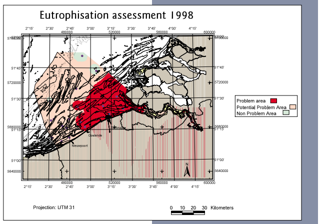
Les données soumises à diffusion restreinte doivent faire l'objet d'une demande motivée adressée à *idod@mumm.ac.be*

Quel usage ?

Ce système d'information s'adresse à un vaste public : scientifiques, autorités publiques, professionnels de la mer, ... C'est pourquoi un ensemble varié d'outils d'extraction, de visualisation et d'analyse est également disponible sur le site web. Les spécialistes du centre de données peuvent également réaliser sur demande des produits ou études spécifiques.







IDOD ?

IDOD is een informatiesysteem over de kwaliteit van het mariene milieu.

Het bestaat uit een relationele gegevensbank en instrumenten voor statistische en ruimtelijke analyses.

Welke gegevens ?

De gegevensbank bevat fysische, chemische en biologische gegevens, gemeten in water, biota en sedimenten van de Noordzee.

Deze gegevens worden gemeten door de BMM en andere wetenschappelijke teams.

Het betreft bijvoorbeeld, de waarden van parameters zoals de temperatuur van het water, de saliniteit, de concentraties aan nutriënten, zware metalen of PCB's of nog informatie over de dierenpopulaties en hun fysiologische toestand.

Elk gegeven heeft slechts echt waarde wanneer het goed gedocumenteerd is. Daarom bevat de gegevensbank ook het maximum aan "meta-informatie": onderzoeksproject, staalnametijd en -positie, type ecosysteem, analysemethode, meteorologische omstandigheden, …

Beschikbaarheid ?

Openbare gegevens kunnen *via* onze website geconsulteerd worden op basis van criteria betreffende de geografische positie, de periode, de parameters en het gewenste outputformaat.

De gegevens voor beperkte verspreiding moeten het voorwerp uitmaken van een gemotiveerde aanvraag gericht aan *idod@mumm.ac.be*

Welk gebruik ?

Dit informatiesysteem richt zich op een groot publiek : wetenschappers, openbare diensten, beroeps van de zee, … Daarom is dan ook een gevarieerd geheel van instrumenten voor de extractie, visualisatie en analyse beschikbaar op de website. Op aanvraag kunnen de specialisten van het gegevenscentrum ook specifieke producten en studies maken.

Programme « Gestion durable de la mer du Nord »
S.S.T.C.

Programma "Duurzaam beheer van de Noordzee"
D.W.T.C.

BMM
UGMM

MUMM
Gulledelle, 100
B-1200 Brussels
Tel.: +32 (0)2 773 21 11
Fax : +32 (0)2 770 69 72
http://www.mumm.ac.be/
e-mail: info@mumm.ac.be

The Belgian Marine Data Centre
http://www.mumm.ac.be/datacentre/

MUMM is a department of the
Royal Belgian Institute of Natural Sciences

museum

Sustainable Management of the North Sea

IDOD

Integrated and Dynamical Oceanographic Data Management

(January 1997 - June 2002)

MUMM Management Unit of the Mathematical Models of the North Sea

IDOD - Integrated and Dynamical Oceanographic Data Management

Teams:

- MUMM - Management Unit of the Mathematical Model of the North Sea

  K. De Cauwer, M. Devolder, S. Jans, L. Schwind, S. Scory, J. Backers, M. Moens

KUL - Universiteit Centrum voor Statistics

G. Dierckx, B. Plevoets, F. Vastmans, J. Van Dyck

ULg - SURFACES

M. Binard, Y. Cornet, F. Muller, J.-C. Sainte

MUMM

IDOD - Integrated and Dynamical Oceanographic Data Management

- *The importance of data:*

  … by itself: science, management

  … with other data: cross-analysis, correlation, verification

  … ecosystem approach

and also… data acquisition costs a lot !

IDOD MUMM

Data acquisition at sea… costs a lot!

- Example for the year 2000

Belgica	= 502 €/hour
Mean salary cost of scientists on board	= 245 €/hour
Total	= 747 €/hour

Cost of a standard campaign

(from Monday 10 a.m to Friday noon) **= 73.206 €**

IDOD MUMM

IDOD - Integrated and Dynamical Oceanographic Data Management

- Conditions:

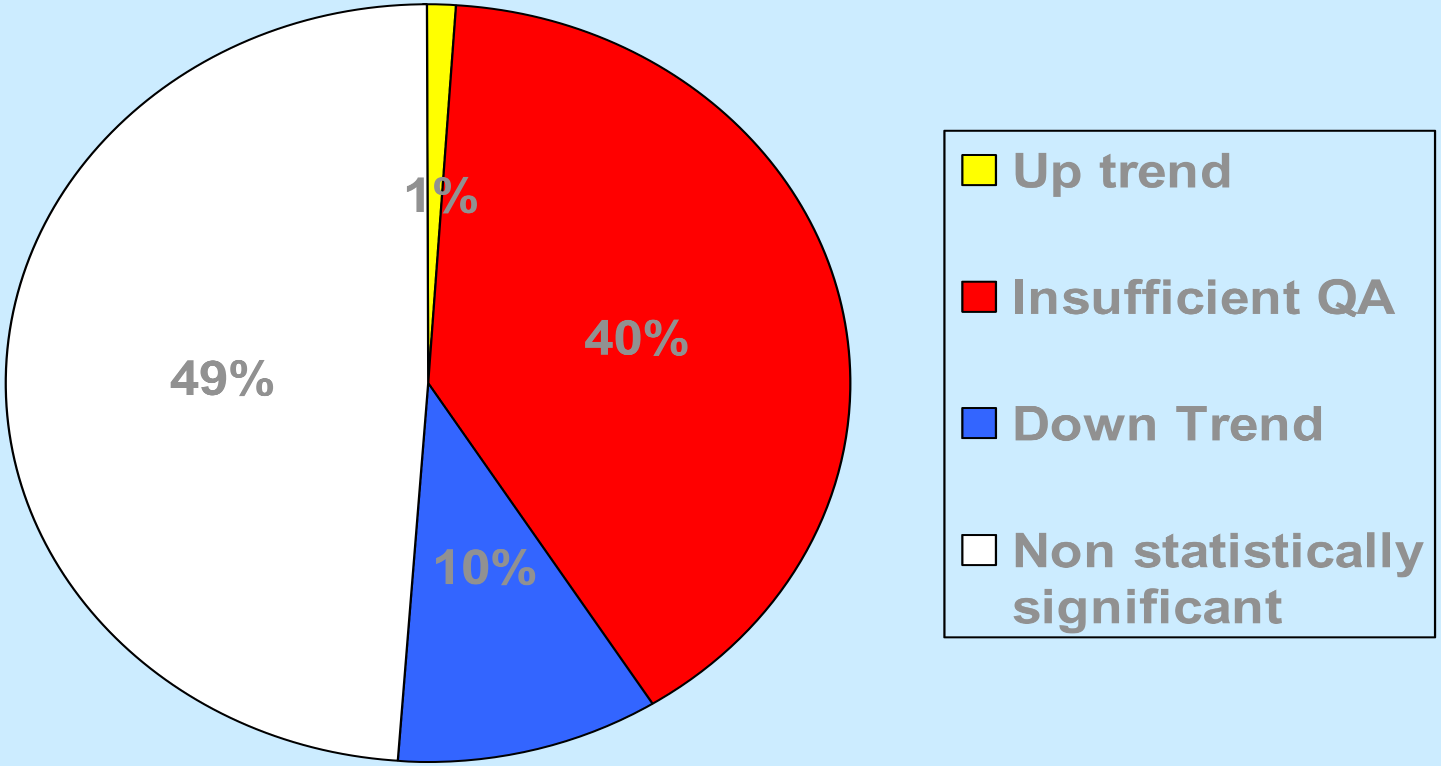
  coherent data set (quality, documentation)

  safe storage of data for the future (new analysis techniques)

  performant analysis tools

MUMM

IDOD - Integrated and Dynamical Oceanographic Data Management



OSPAR trend assessment of contamination in biota

IDOD MUMM

IDOD - Integrated and Dynamical Oceanographic Data Management

- *Objectives*

  To provide structured, homogenized and validated oceanographic data necessary for any scientific research, decision making and sustainable development…

  To establish, to manage and to promote an integrated database of marine environmental data, ensuring a smooth and scientifically sound data flow between the data producers and the end-users

  Interest of everybody…

MUMM

IDOD - Integrated and Dynamical Oceanographic Data Management

- Therefore… IDOD

  IDOD is an information system developed to store, retrieve and process oceanographic data

  IDOD = a relational database + statistical and spatial analysis tools

MUMM

How to reach these objectives?

Technical aspects

- Inventory of data sets and databases
- Set-up of
  - data base
  - quality control
  - data transfer procedures

Users tools

Products and applications

Conclusion

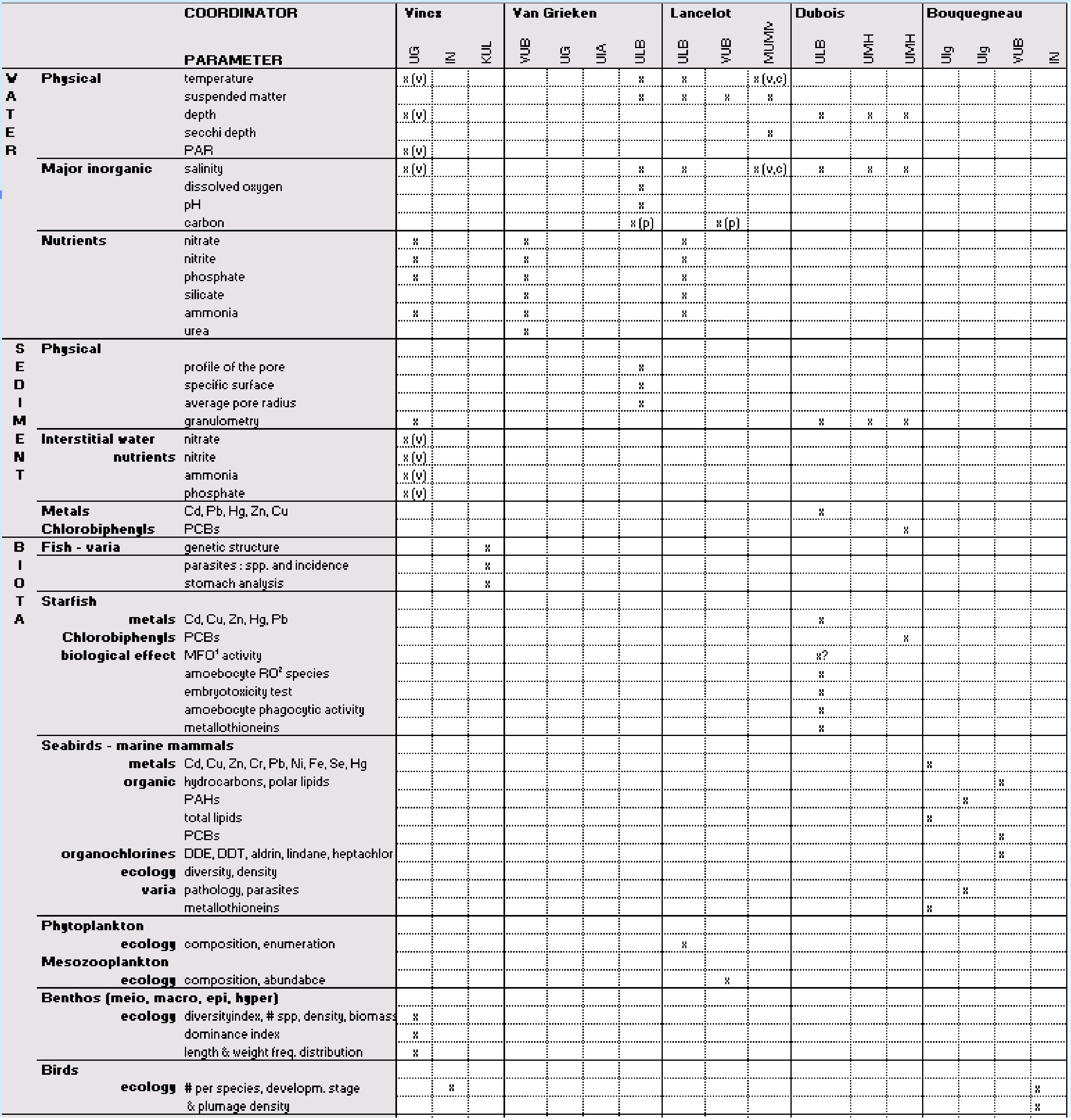
IDOD MUMM

IDOD - Integrated and Dynamical Oceanographic Data Management

Technical aspects...

Data set inventory

(data collected in the frame of the Programme)



		COORDINATOR	Vincx			Van Grieken				Lancelot			Dubois			Bouquegneau			
		PARAMETER	UG	IN	KUL	VUB	UG	UIA	ULB	ULB	VUB	MUMM	ULB	UMH	UMH	Ulg	Ulg	VUB	IN
W	**Physical**	temperature	x (v)						x	x		x (v,c)							
A		suspended matter							x	x	x	x							
T		depth	x (v)										x	x	x				
E		secchi depth										x							
R		PAR	x (v)																
	Major inorganic	salinity	x (v)						x	x		x (v,c)	x	x	x				
		dissolved oxygen							x										
		pH							x										
		carbon							x (p)		x (p)								
	Nutrients	nitrate	x			x				x									
		nitrite	x			x				x									
		phosphate	x			x				x									
		silicate				x				x									
		ammonia	x			x				x									
		urea				x													
S	**Physical**																		
E		profile of the pore							x										
D		specific surface							x										
I		average pore radius							x										
M		granulometry	x										x	x	x				
E	**Interstitial water**	nitrate	x (v)																
N	**nutrients**	nitrite	x (v)																
T		ammonia	x (v)																
		phosphate	x (v)																
	Metals	Cd, Pb, Hg, Zn, Cu											x						
	Chlorobiphenyls	PCBs													x				
B	**Fish - varia**	genetic structure			x														
I		parasites : spp. and incidence			x														
O		stomach analysis			x														
T	**Starfish**																		
A	**metals**	Cd, Cu, Zn, Hg, Pb											x						
	Chlorobiphenyls	PCBs													x				
	biological effect	MFO[1] activity											x?						
		amoebocyte RO[2] species											x						
		embryotoxicity test											x						
		amoebocyte phagocytic activity											x						
		metallothioneins											x						
	Seabirds - marine mammals																		
	metals	Cd, Cu, Zn, Cr, Pb, Ni, Fe, Se, Hg														x			
	organic	hydrocarbons, polar lipids																x	
		PAHs															x		
		total lipids														x			
		PCBs																x	
	organochlorines	DDE, DDT, aldrin, lindane, heptachlor																x	
	ecology	diversity, density																	
	varia	pathology, parasites															x		
		metallothioneins														x			
	Phytoplankton																		
	ecology	composition, enumeration								x									
	Mesozooplankton																		
	ecology	composition, abundabce									x								
	Benthos (meio, macro, epi, hyper)																		
	ecology	diversityindex, # spp, density, biomass	x																
		dominance index	x																
		length & weight freq. distribution	x																
	Birds																		
	ecology	# per species, developm. stage		x															x
		& plumage density																	x

Data types

- concentration (e.g.: nutrients) in water
- concentration (e.g.: heavy metals) in sediment
- concentration (e.g.: PCB’s) in biota
- biomass and population densities
- observations on biota : e.g. %coverage with oil, absence/presence of fish diseases
- models results

Meta Information

- General (date, time, position)
- Methods (sampling, sample handling, analysis…)
- Quality control (Control charts, intercalibration exercises such as Quasimeme, …)
- Meteorology (wind speed and direction, solar radiation, air temperature…)

Set-up of the data base

- Analysis and design
- Implementation and production

Analysis and design

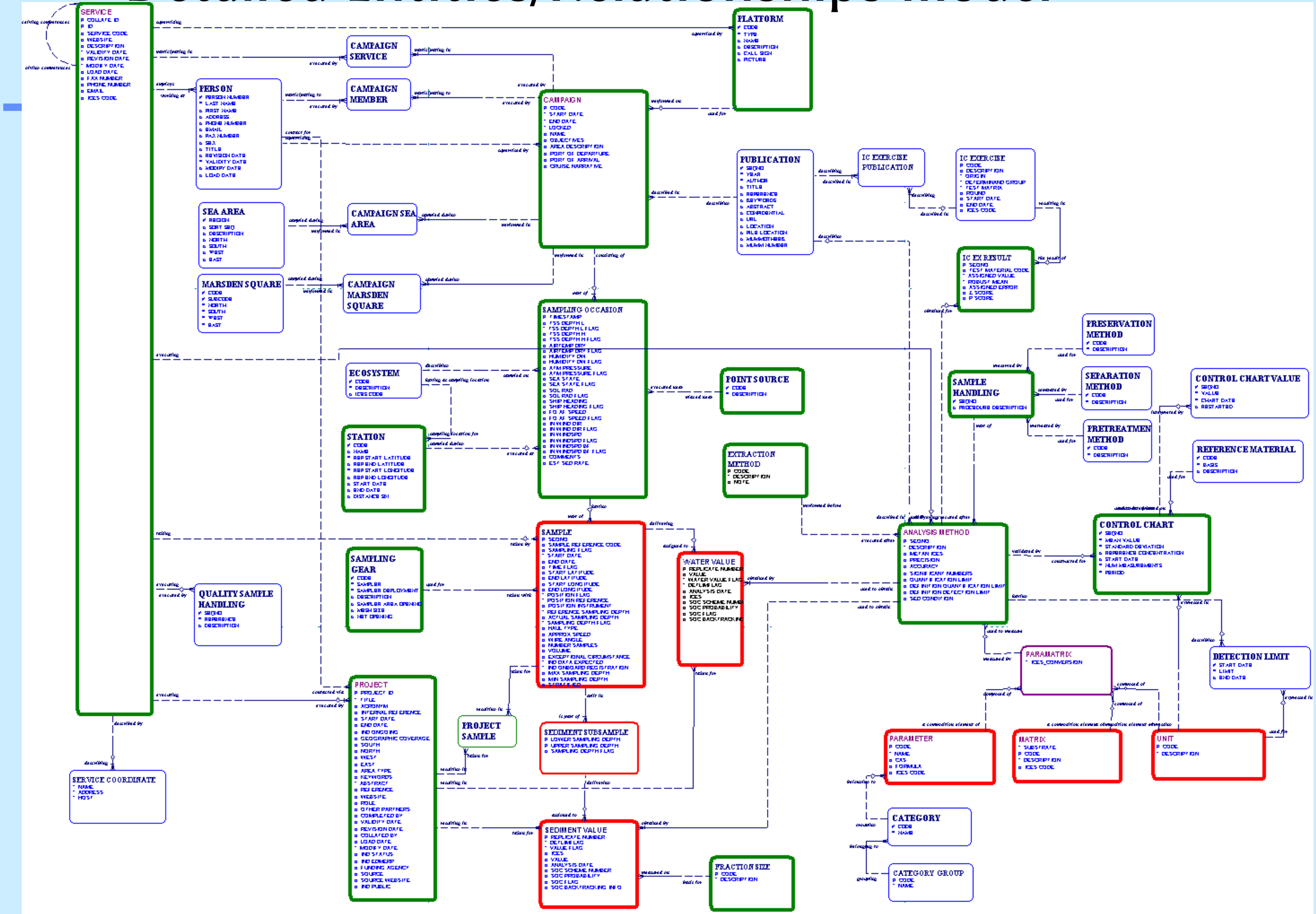
What do we want to do ?

Definition of the entities and relationships based on the preliminary analysis.

How do we want to do it ?

Description of modules for importing, retrieving and updating data into the dB.

Detailed Entities/Relationships model



Sampling with Niskin bottle

Campaigns

BE2001/03

Sampling occasions

Station 700 - 7/02/2001 09.56 u

Samples

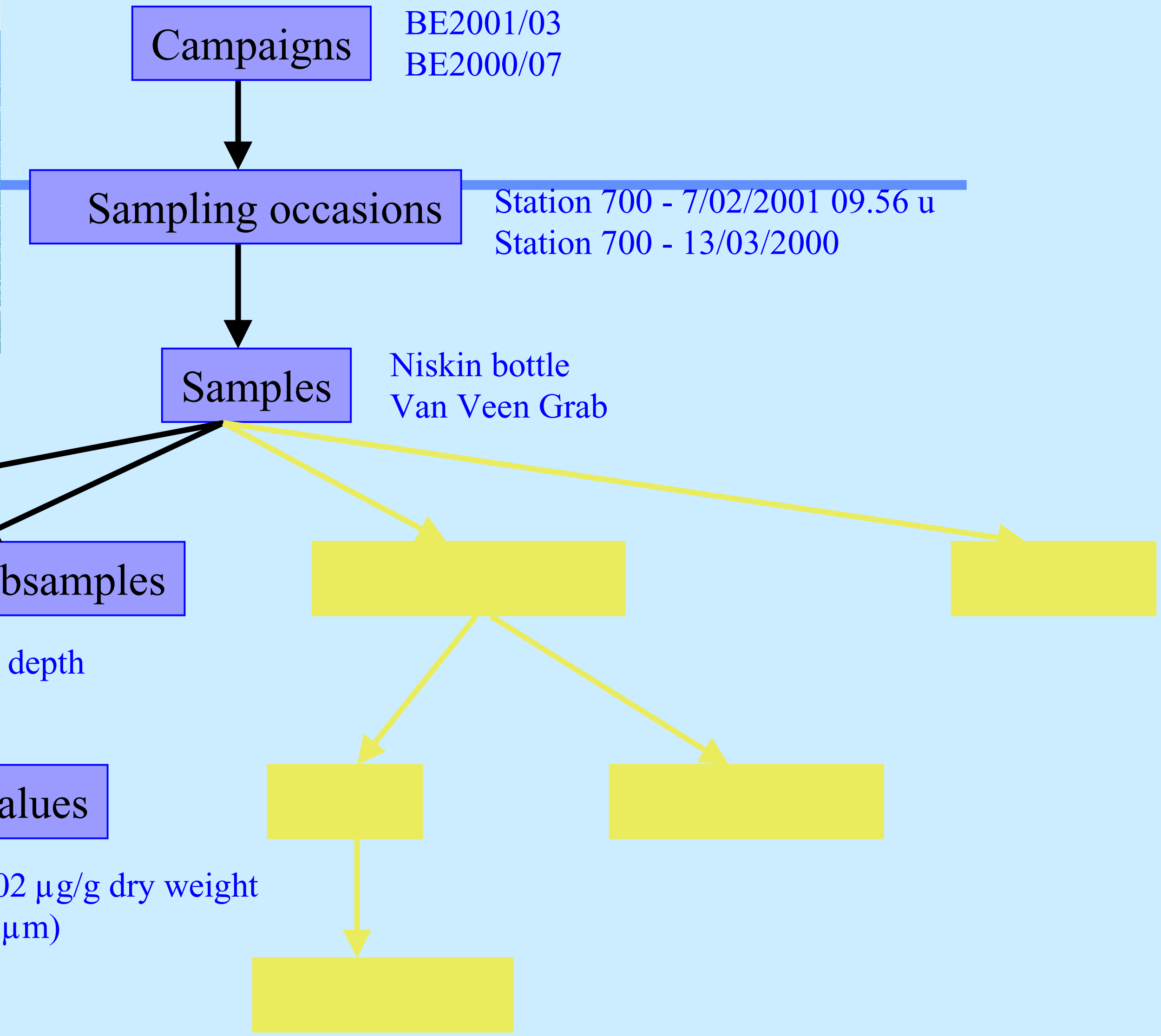
Niskin bottle

Water values

Nitrate 75 µmol/l
Ammonium 2 µmol/l
Dissolved oxygen 9.8 mg/l

IDOD MUMM

Sampling with boxcorer



Campaigns

BE2001/03
BE2000/07

Sampling occasions

Station 700 - 7/02/2001 09.56 u
Station 700 - 13/03/2000

Samples

Niskin bottle
Van Veen Grab

Water values

Nitrate : 75 µmol/l
Ammonium : 2 µmol/l
Dissolved oxygen : 9.8 mg/l

Sediment subsamples

0-10 cm depth

Sediment values

CB101 : 0.00202 µg/g dry weight
(< 63 µm)

IDOD MUMM

Fishing net

Campaigns

BE2000/03
BE2000/07
BR2000 (Broodwinner)

Sampling occasions

Station 700 - 7/02/2001 09.56 u
Station 700 - 13/03/2000
BCP - 5/10/2000

Samples

Niskin bottle
Van Veen Grab
Beam trawl

Water values

Nitrate : 75 µmol/l
Ammonium : 2 µmol/l
Dissolved oxygen : 9.8 mg/l

Sediment subsamples

0-10 cm depth

Sediment values

CB101 : 0.00202 µg/g dry weight
(< 63 µm)

Biota subsamples

5 Atlantic cod specimens
with average weight of 446 g

Tissue

Liver tissue with
58 % lipid weight

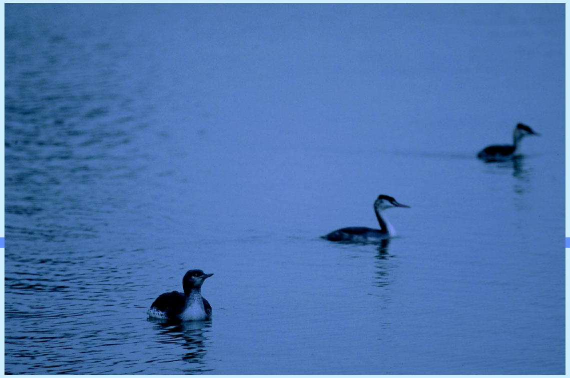
Biota values

CB101 = 0.18628 µg/g lipid weight

Observations

2 individuals are
males

MUMM



grebes

Campaigns

BE2000/03
BE2000/07
BR2000 (Broodwinner)
BE1998/13

Sampling occasions

Station 700 - 7/02/2001 09.56 u
Station 700 - 13/03/2000
BCP - 5/10/2000
Square 9312 - 2/06/1998

Samples

Niskin bottle
Van Veen Grab
Beam trawl
Eye/binocular

Water values

Nitrate : 75 µmol/l
Ammonium : 2 µmol/l
Dissolved oxygen : 9.8 mg/l

Sediment subsamples

0-10 cm depth

Sediment values

CB101 : 0.00202 µg/g dry weight
(< 63 µm)

Biota subsamples

5 Atlantic cod specimens
with average weight of 446 g

Tissue

Liver tissue with
58 % lipid weight

Biota values

CB101 = 0.18628 µg/g lipid weight

Observations

2 individues are
males

Densities

2.2 Black-headed
Gulls/km2
45.12 Common
Terns/km2

Implementation / Production

= actual programmation of the database

Software used:

- **DB**: Oracle *8.i* and related designer and development tools
- Windows-NT

Set-up of quality control

- Automatic DB Quality Control
  - 33 validation rules

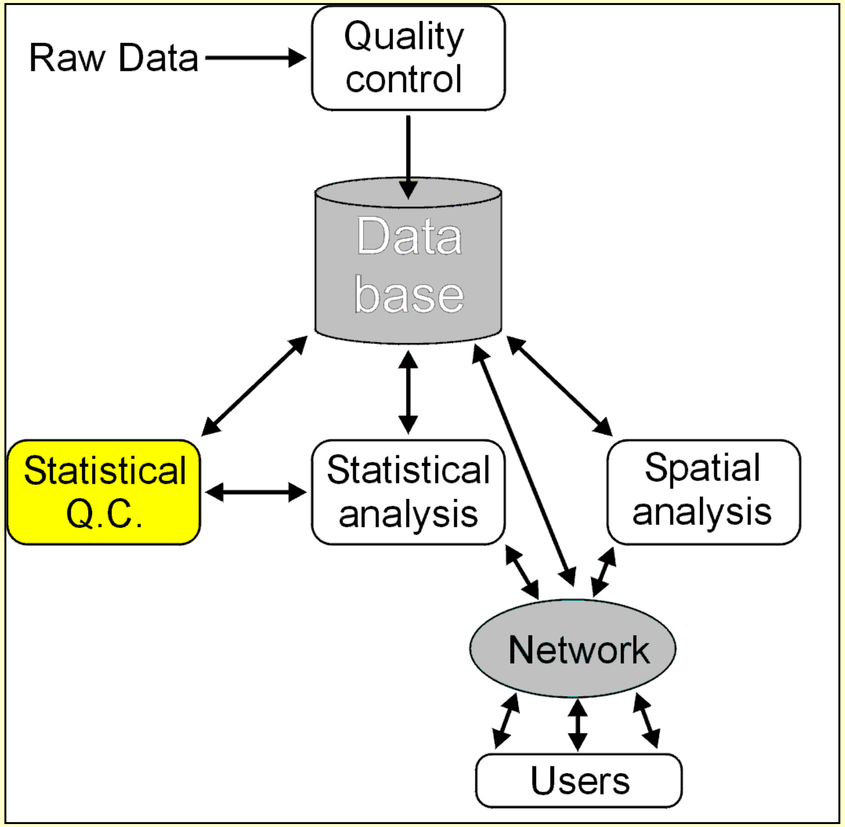
- Statistical Quality Control

IDOD MUMM

Automatic DB Quality Control

- Examples of validation rules
  - the start and end dates of a sample must be between the start and end dates of the related campaign
  - the matrix for a water value should be ‘dissolved’, ‘particulate’ or ‘total’
  - the value for dissolved phosphate must be lower than the value for total phosphorus

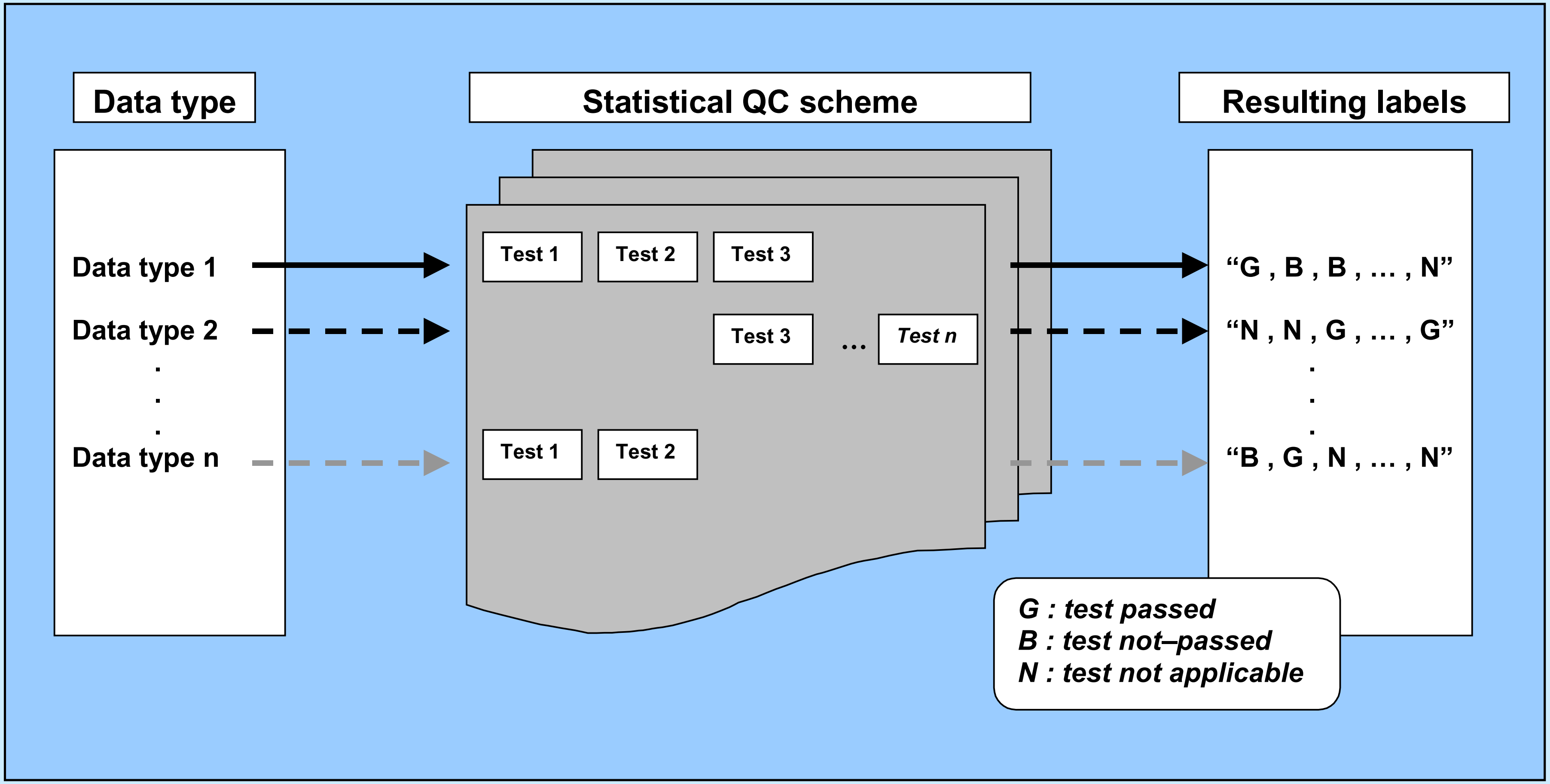
Statistical Quality Control



MUMM

SQC: Conceptual Design

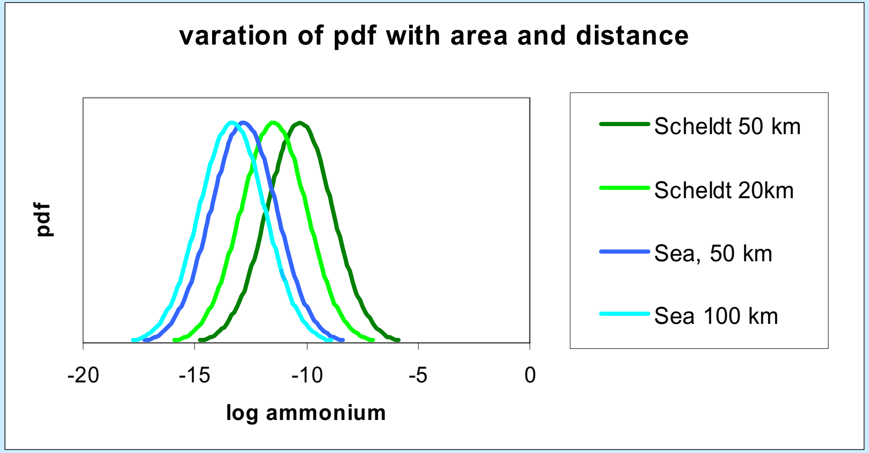
Universitair Centrum voor Statistiek



G : test passed
B : test not–passed
N : test not applicable

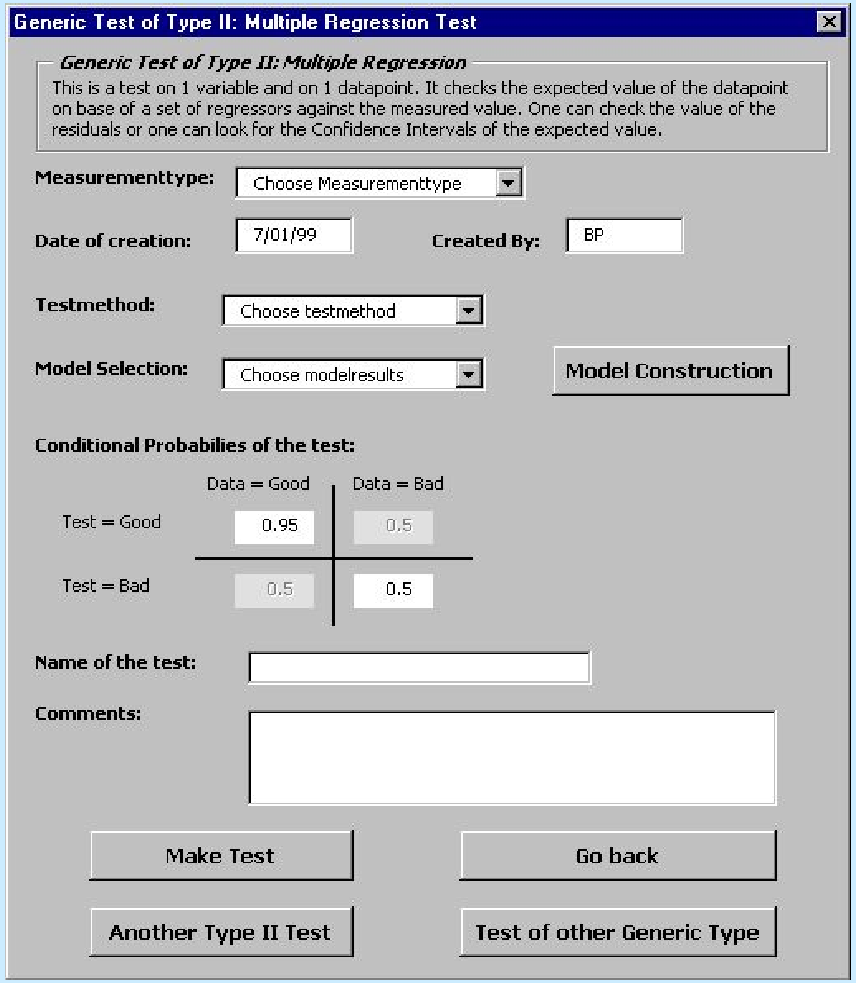
MUMM

Distribution QC



Parameters are function of

- time period
- geographical area
- location within geographical area



Generic Test of Type II: Multiple Regression Test

Generic Test of Type II: Multiple Regression

This is a test on 1 variable and on 1 datapoint. It checks the expected value of the datapoint on base of a set of regressors against the measured value. One can check the value of the residuals or one can look for the Confidence Intervals of the expected value.

Measurementtype: Choose Measurementtype

Date of creation: 7/01/99 **Created By:** BP

Testmethod: Choose testmethod

Model Selection: Choose modelresults

Conditional Probabilies of the test:

	Data = Good	Data = Bad
Test = Good	0.95	0.5
Test = Bad	0.5	0.5

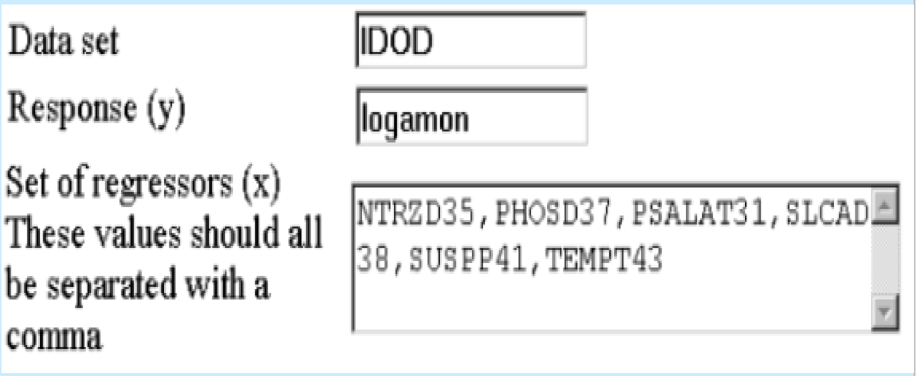
Name of the test:

Comments:

Regression QC

Universitair Centrum voor Statistiek

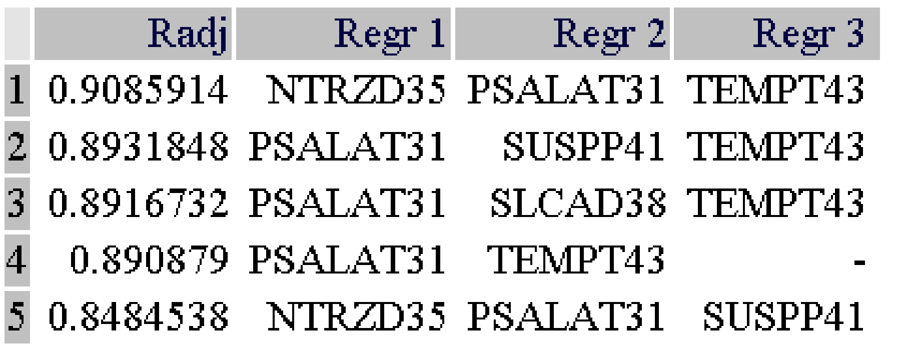
- Check against "best" applicable regression model



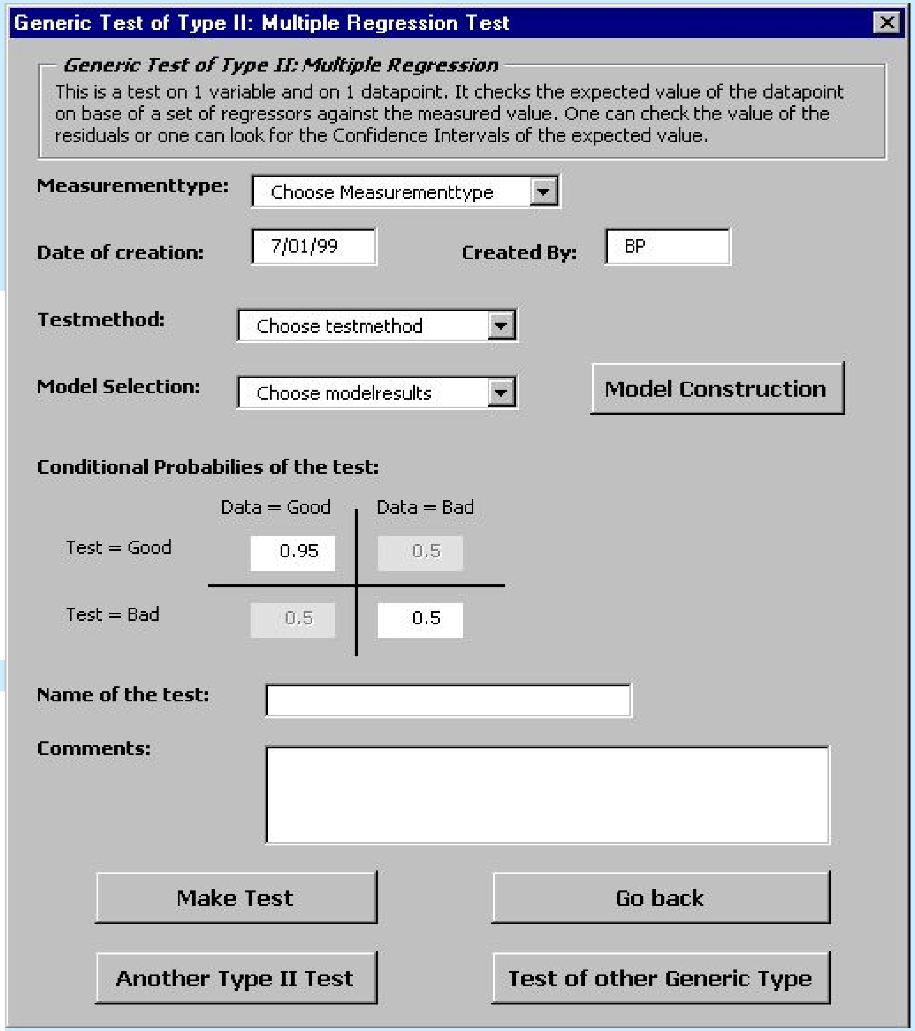
Data set: IDOD

Response (y): logamon

Set of regressors (x) These values should all be separated with a comma: NTRZD35,PHOSD37,PSALAT31,SLCAD38,SUSPP41,TEMPT43



	Radj	Regr 1	Regr 2	Regr 3
1	0.9085914	NTRZD35	PSALAT31	TEMPT43
2	0.8931848	PSALAT31	SUSPP41	TEMPT43
3	0.8916732	PSALAT31	SLCAD38	TEMPT43
4	0.890879	PSALAT31	TEMPT43	-
5	0.8484538	NTRZD35	PSALAT31	SUSPP41



Generic Test of Type II: Multiple Regression Test

Generic Test of Type II: Multiple Regression

This is a test on 1 variable and on 1 datapoint. It checks the expected value of the datapoint on base of a set of regressors against the measured value. One can check the value of the residuals or one can look for the Confidence Intervals of the expected value.

Measurementtype: Choose Measurementtype

Date of creation: 7/01/99

Created By: BP

Testmethod: Choose testmethod

Model Selection: Choose modelresults

Conditional Probabilies of the test:

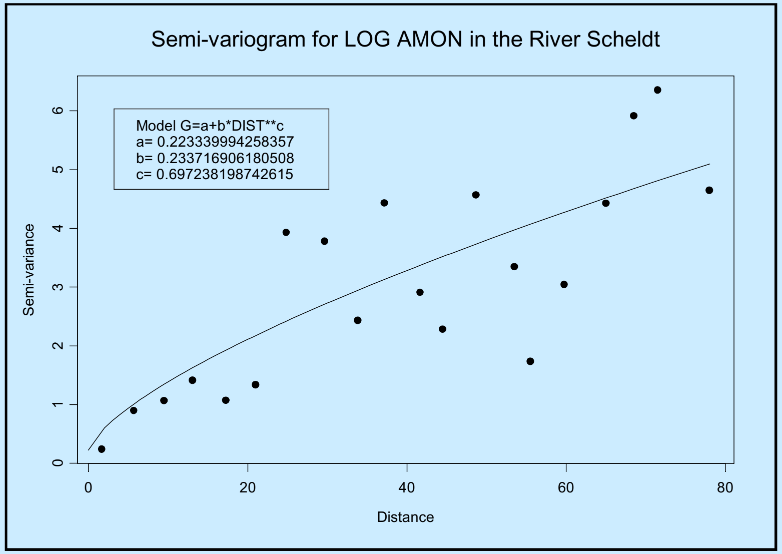
	Data = Good	Data = Bad
Test = Good	0.95	0.5
Test = Bad	0.5	0.5

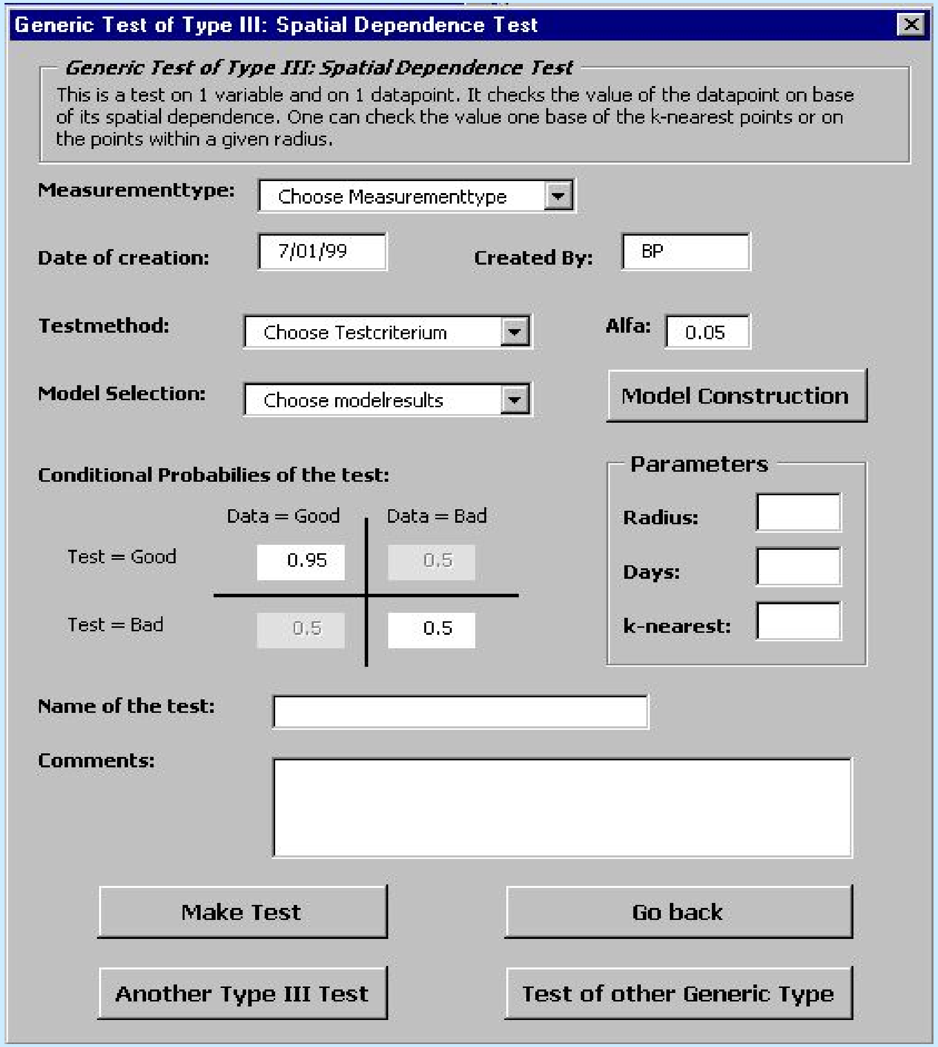
Name of the test:

Comments:

Spatial Interpolation QC

UCS
Universitair Centrum voor Statistiek





Generic Test of Type III: Spatial Dependence Test

Generic Test of Type III: Spatial Dependence Test

This is a test on 1 variable and on 1 datapoint. It checks the value of the datapoint on base of its spatial dependence. One can check the value one base of the k-nearest points or on the points within a given radius.

Measurementtype: Choose Measurementtype

Date of creation: 7/01/99 **Created By:** BP

Testmethod: Choose Testcriterium **Alfa:** 0.05

Model Selection: Choose modelresults

Conditional Probabilies of the test:

	Data = Good	Data = Bad
Test = Good	0.95	0.5
Test = Bad	0.5	0.5

Parameters

Radius:

Days:

k-nearest:

Name of the test:

Comments:

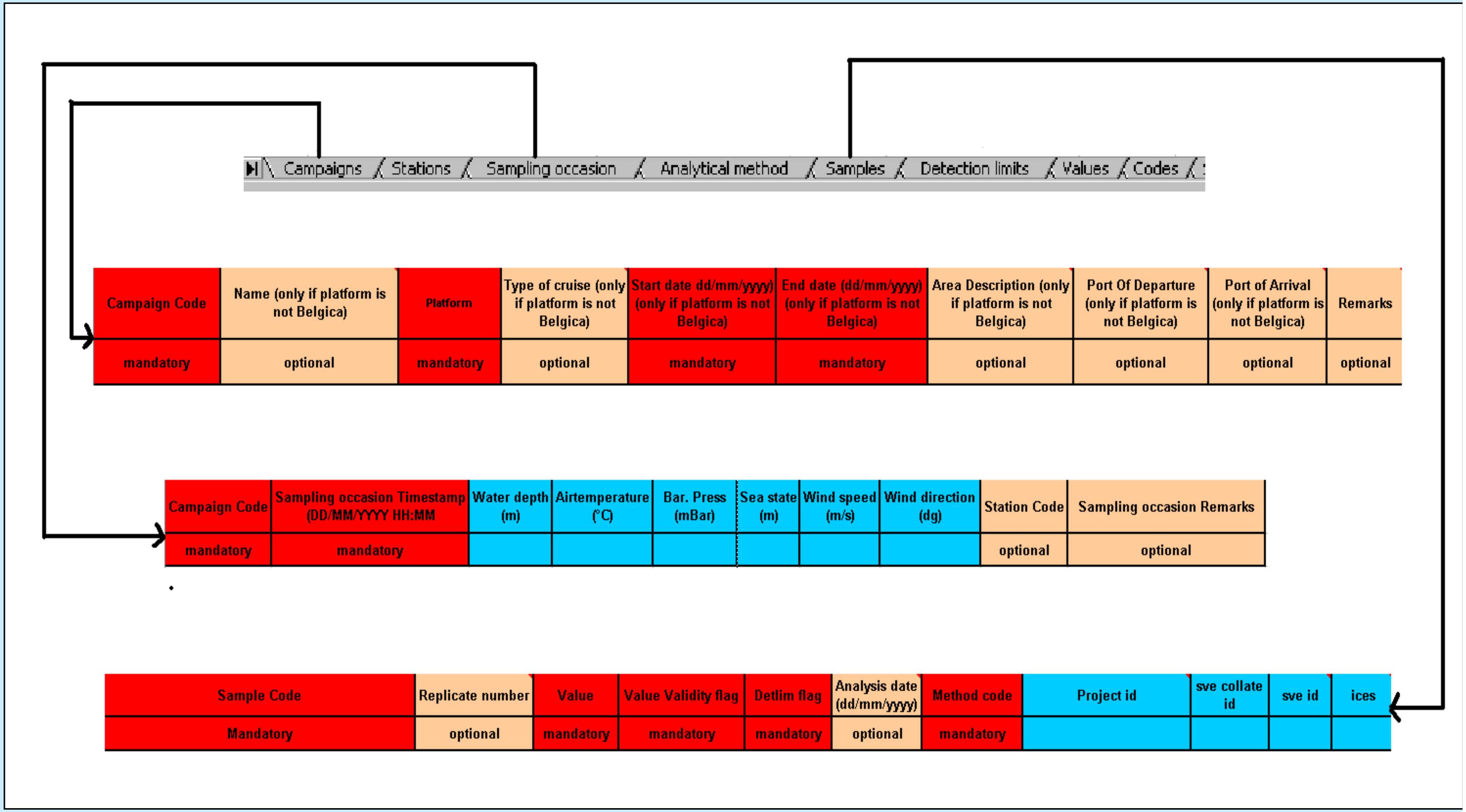
- Check of difference with neighbor(s)

IDOD MUMM

Combination of test results (Bayesian Approach)

- Test result : G = “passed” or B = “not-passed” or N = “not-applicable”
- Test characteristics:
  - What is the probability that a good datapoint would not pass the test?
  - What is the probability that a bad datapoint would pass the test
- The (a-posteriori) probability of a datapoint being good can then be calculated and stored with the value

Set-up of data transfer procedures



Campaign Code	Name (only if platform is not Belgica)	Platform	Type of cruise (only if platform is not Belgica)	Start date dd/mm/yyyy) (only if platform is not Belgica)	End date (dd/mm/yyyy) (only if platform is not Belgica)	Area Description (only if platform is not Belgica)	Port Of Departure (only if platform is not Belgica)	Port of Arrival (only if platform is not Belgica)	Remarks
mandatory	optional	mandatory	optional	mandatory	mandatory	optional	optional	optional	optional

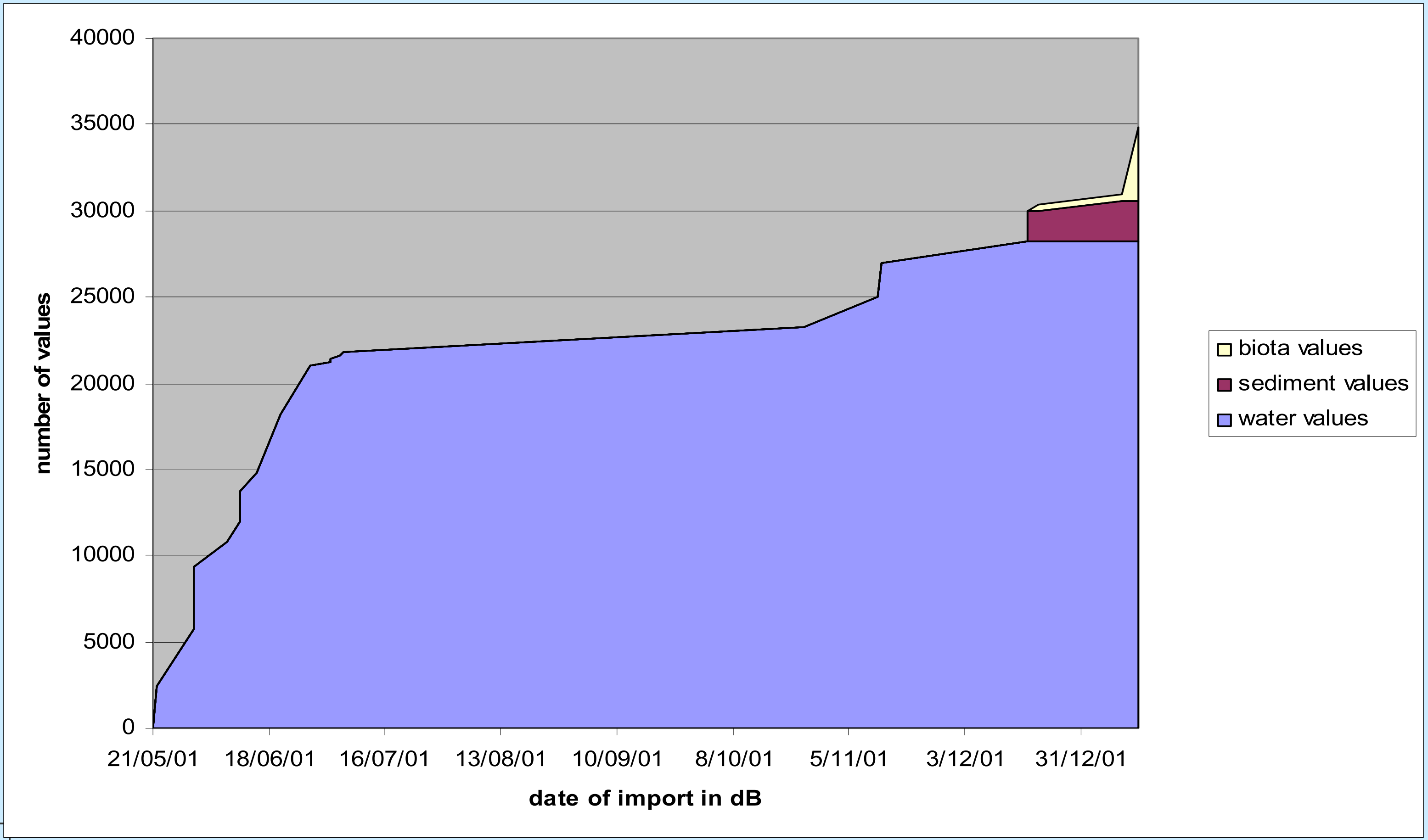
Campaign Code	Sampling occasion Timestamp (DD/MM/YYYY HH:MM	Water depth (m)	Airtemperature (°C)	Bar. Press (mBar)	Sea state (m)	Wind speed (m/s)	Wind direction (dg)	Station Code	Sampling occasion Remarks
mandatory	mandatory							optional	optional

Sample Code	Replicate number	Value	Value Validity flag	Detlim flag	Analysis date (dd/mm/yyyy)	Method code	Project id	sve collate id	sve id	ices
Mandatory	optional	mandatory	mandatory	mandatory	optional	mandatory				

Set-up of data transfer procedures
Data sets received at MUMM

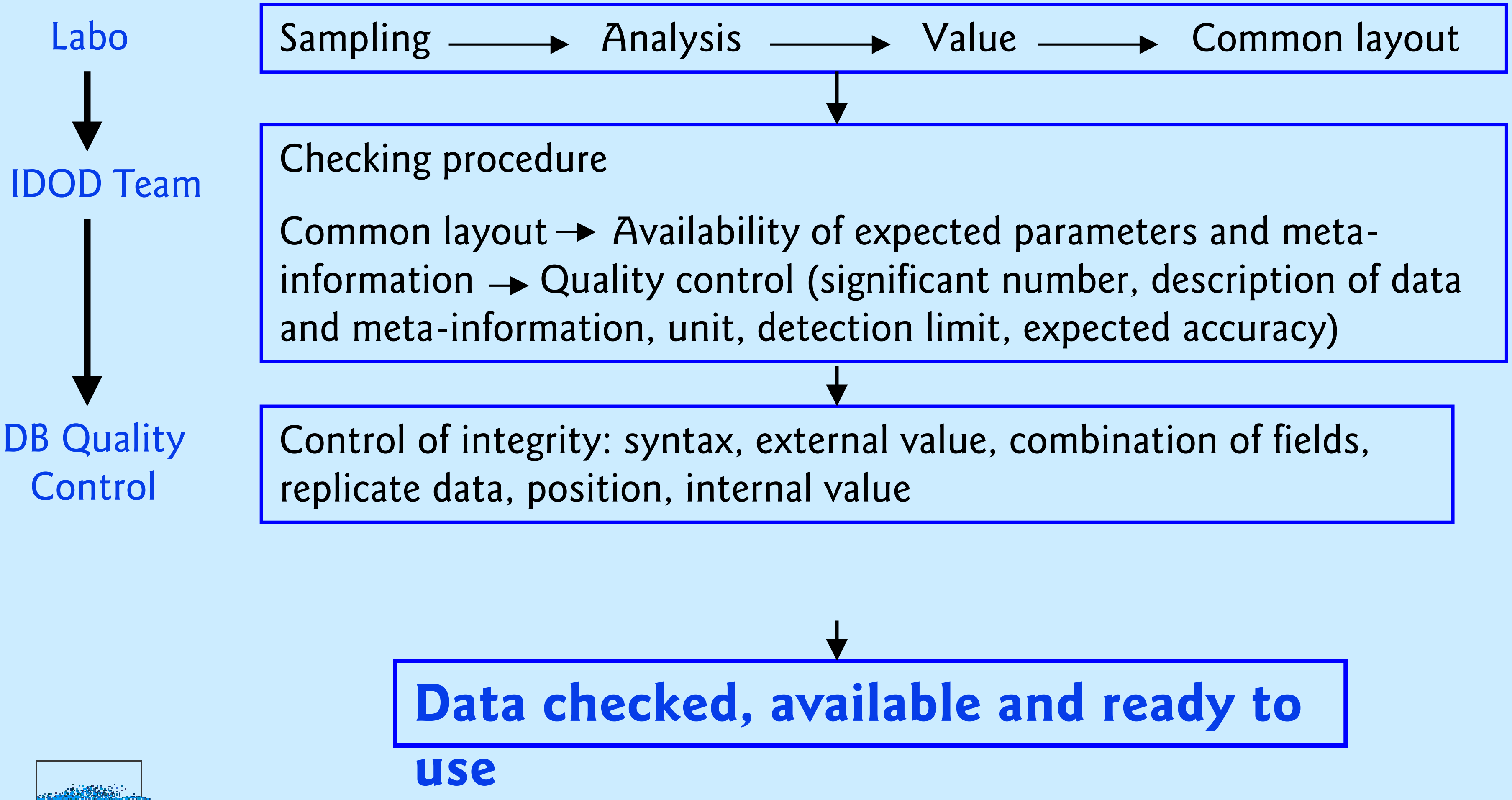
IDOD MUMM

Set-up of data transfer procedures Import of data in the DB



IDOD MUMM

Data Life



Labo

IDOD Team

DB Quality Control

Sampling → Analysis → Value → Common layout

Checking procedure

Common layout → Availability of expected parameters and meta-information → Quality control (significant number, description of data and meta-information, unit, detection limit, expected accuracy)

Control of integrity: syntax, external value, combination of fields, replicate data, position, internal value

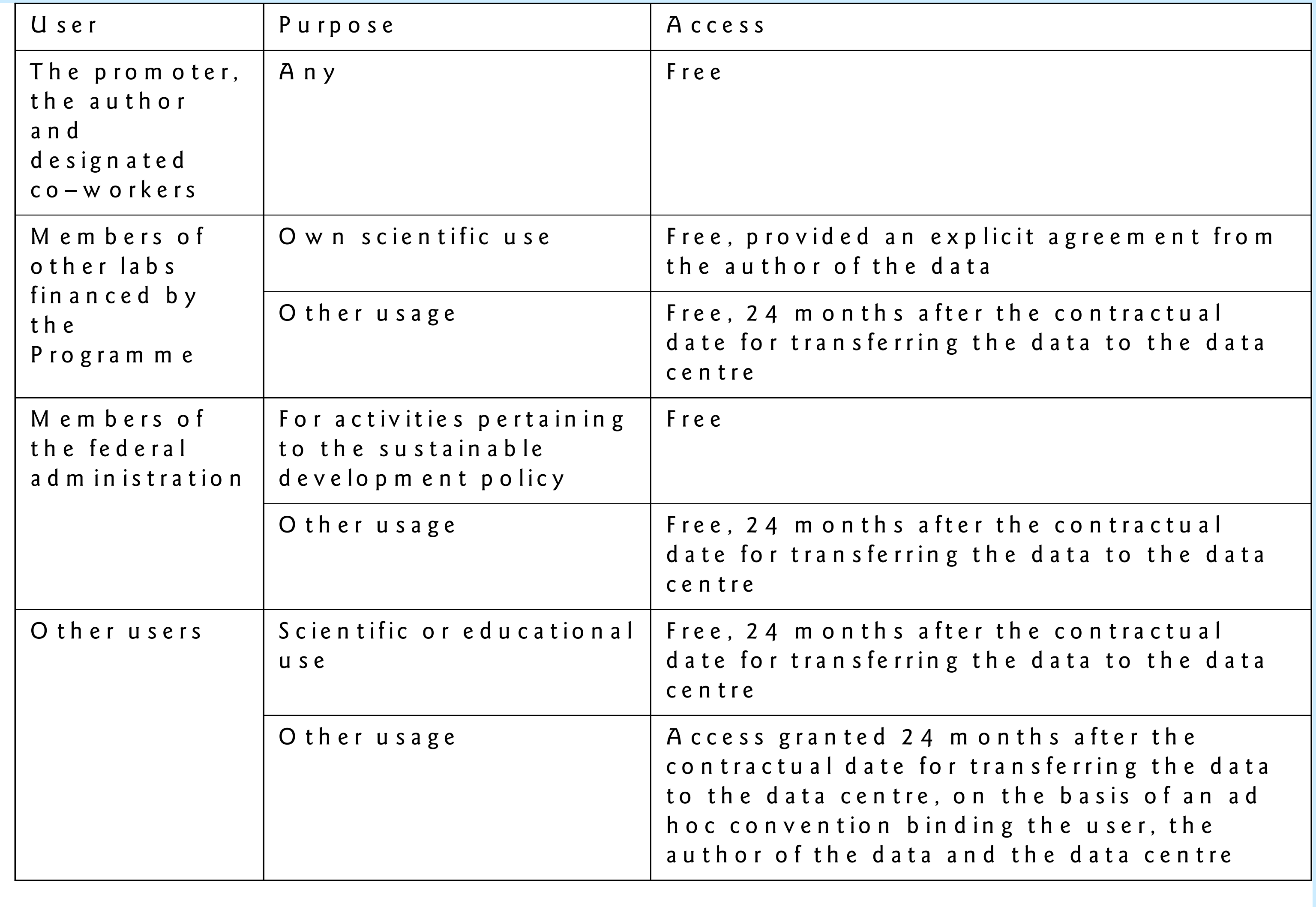
Data checked, available and ready to use

MUMM

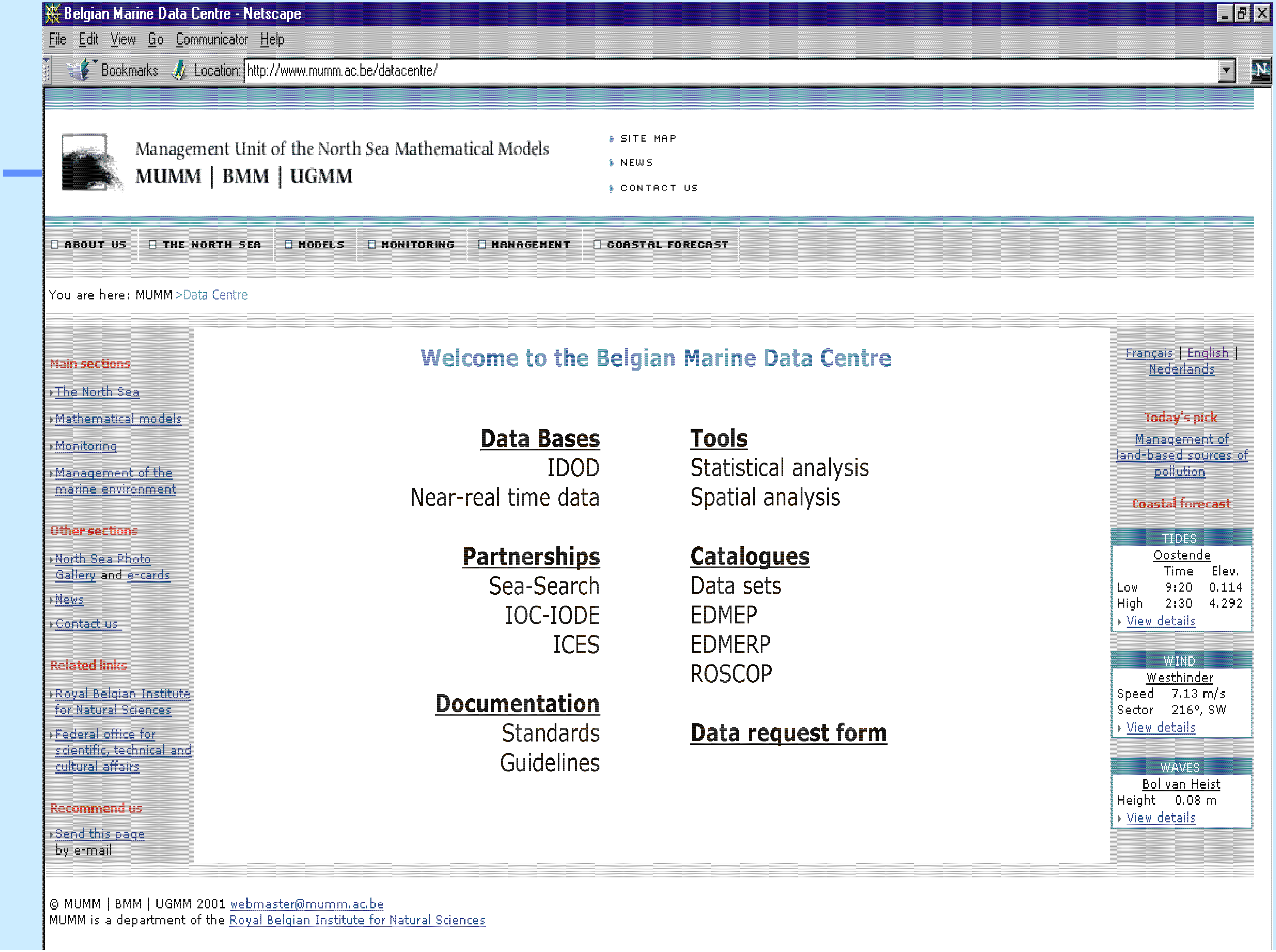
IDOD - Integrated and Dynamical Oceanographic Data Management

Users tools...

Access rules to the data gathered during the Programme



User	Purpose	Access
The promoter, the author and designated co-workers	Any	Free
Members of other labs financed by the Programme	Own scientific use	Free, provided an explicit agreement from the author of the data
	Other usage	Free, 24 months after the contractual date for transferring the data to the data centre
Members of the federal administration	For activities pertaining to the sustainable development policy	Free
	Other usage	Free, 24 months after the contractual date for transferring the data to the data centre
Other users	Scientific or educational use	Free, 24 months after the contractual date for transferring the data to the data centre
	Other usage	Access granted 24 months after the contractual date for transferring the data to the data centre, on the basis of an ad hoc convention binding the user, the author of the data and the data centre


Belgian Marine Data Centre - Netscape
File Edit View Go Communicator Help
Bookmarks Location: http://www.mumm.ac.be/datacentre/

Management Unit of the North Sea Mathematical Models
MUMM | BMM | UGMM

- SITE MAP
- NEWS
- CONTACT US

□ ABOUT US □ THE NORTH SEA □ MODELS □ MONITORING □ MANAGEMENT □ COASTAL FORECAST

You are here: MUMM >Data Centre

Main sections

- The North Sea
- Mathematical models
- Monitoring
- Management of the marine environment

Other sections

- North Sea Photo Gallery and e-cards
- News
- Contact us

Related links

- Royal Belgian Institute for Natural Sciences
- Federal office for scientific, technical and cultural affairs

Recommend us

- Send this page by e-mail

Welcome to the Belgian Marine Data Centre

Data Bases
IDOD
Near-real time data

Tools
Statistical analysis
Spatial analysis

Partnerships
Sea-Search
IOC-IODE
ICES

Catalogues
Data sets
EDMEP
EDMERP
ROSCOP

Documentation
Standards
Guidelines

Data request form

Français | English | Nederlands

Today's pick
Management of land-based sources of pollution

Coastal forecast

TIDES
Oostende

	Time	Elev.
Low	9:20	0.114
High	2:30	4.292

- View details

WIND
Westhinder

Speed	7.13 m/s
Sector	216°, SW

- View details

WAVES
Bol van Heist

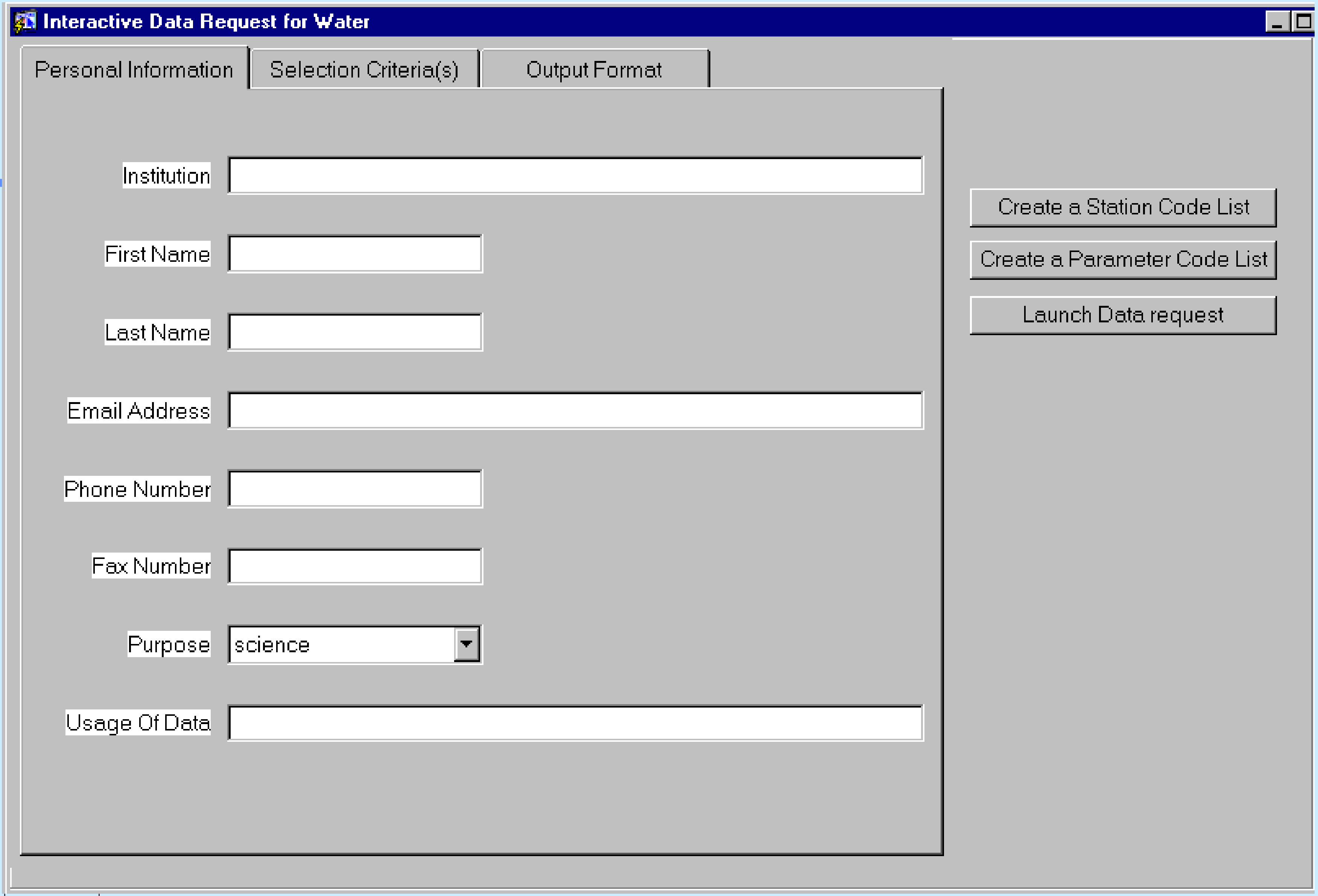
Height	0.08 m

- View details

© MUMM | BMM | UGMM 2001 webmaster@mumm.ac.be
MUMM is a department of the Royal Belgian Institute for Natural Sciences

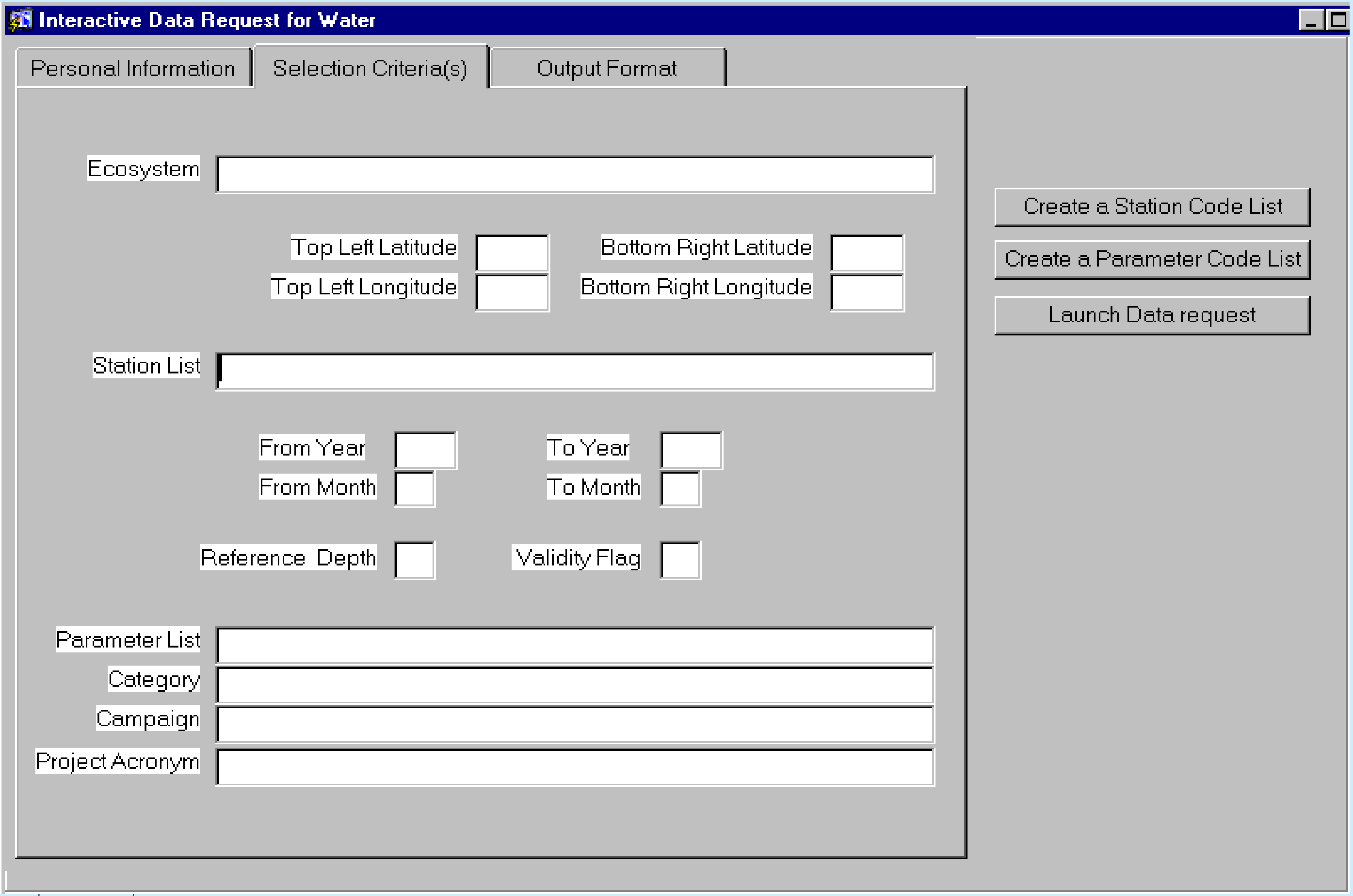
IDOD Database functionalities

- Conditional website access with user friendly interface
- Possible requests based on
  - area type
  - geographical coordinates
  - stations
  - time period
  - parameter / category
  - campaign / project
- Different export formats
- Statistical and Spatial Analysis tools to analyse your results



Purpose: science

IDOD MUMM



Interactive Data Request for Water

Personal Information | Selection Criteria(s) | Output Format

Ecosystem

Top Left Latitude

Top Left Longitude

Bottom Right Latitude

Bottom Right Longitude

Station List

From Year

From Month

To Year

To Month

Reference Depth

Validity Flag

Parameter List

Category

Campaign

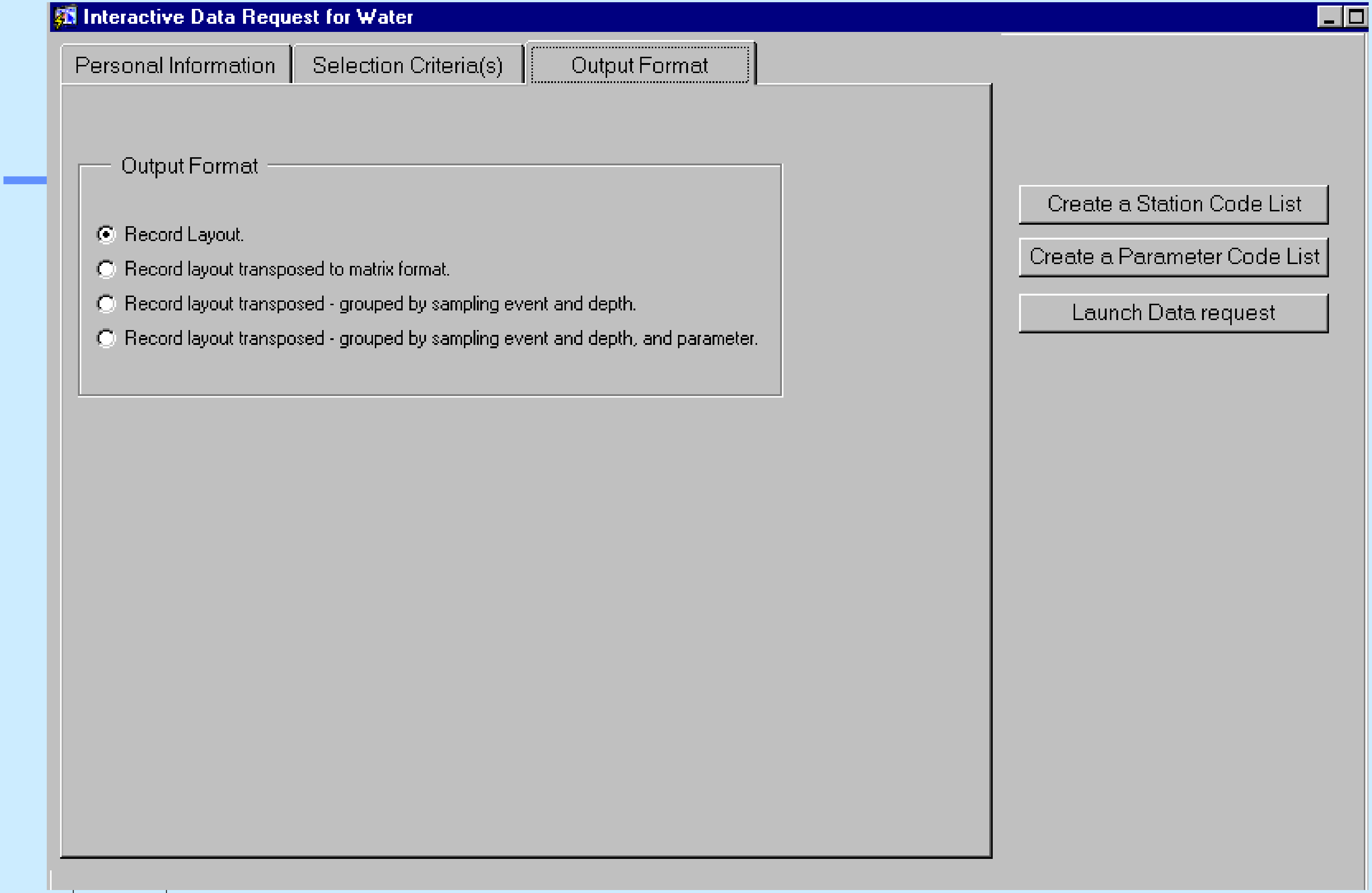
Project Acronym

Create a Station Code List

Create a Parameter Code List

Launch Data request

IDOD MUMM



Interactive Data Request for Water

Output Format

● Record Layout.

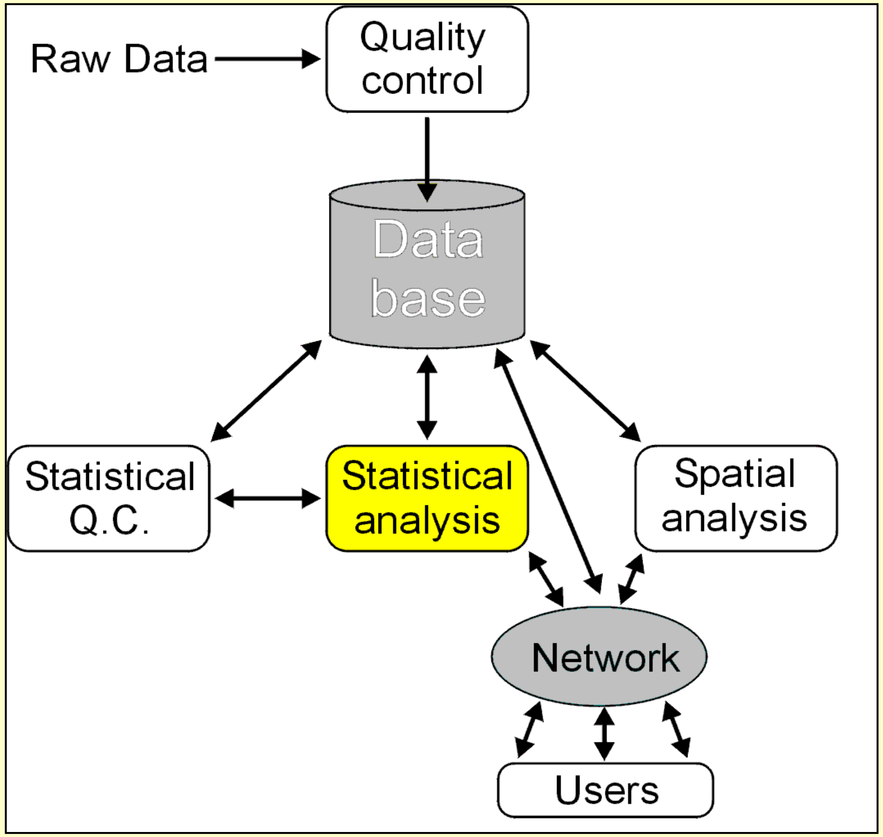
○ Record layout transposed to matrix format.

○ Record layout transposed - grouped by sampling event and depth.

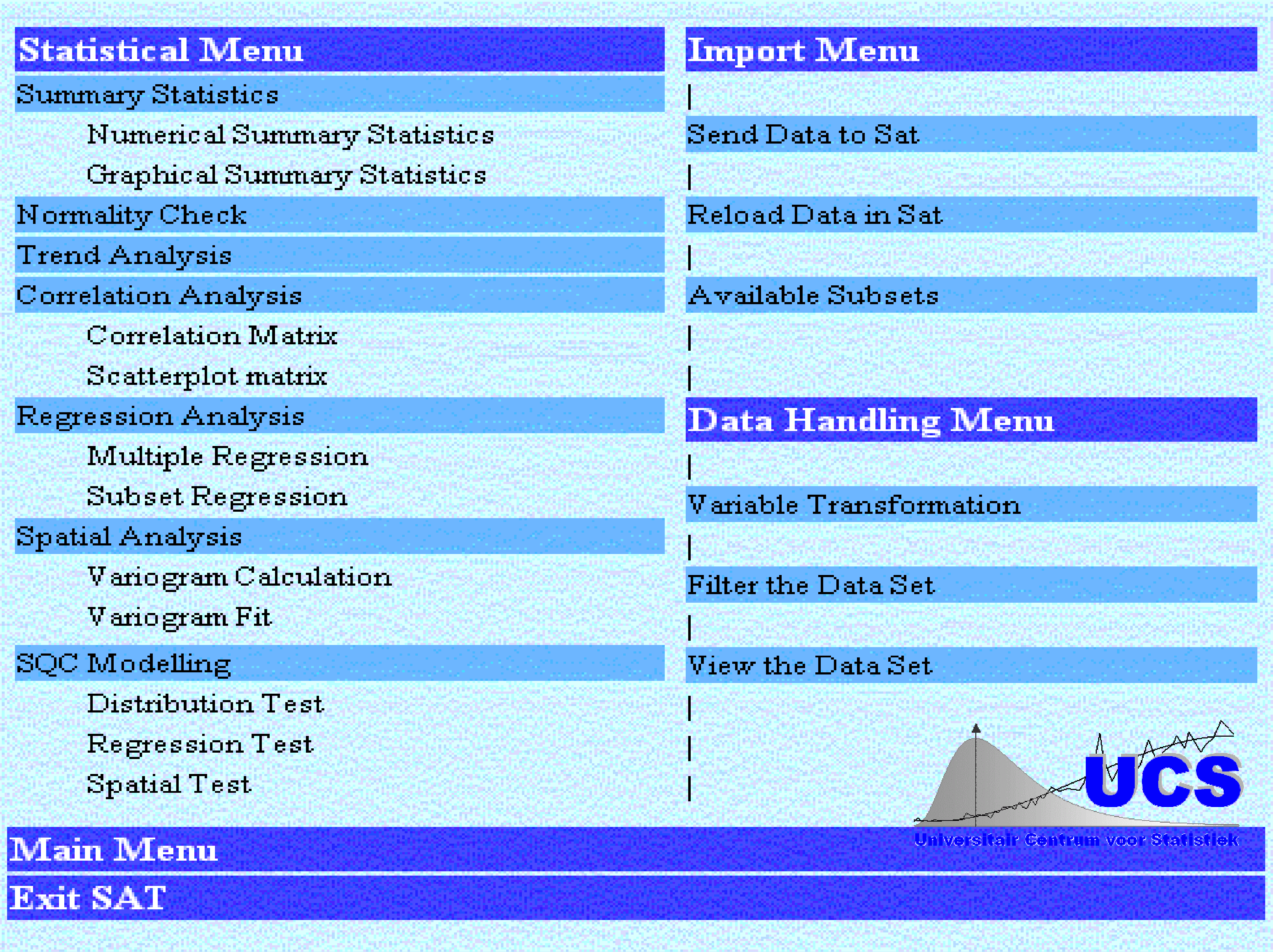
○ Record layout transposed - grouped by sampling event and depth, and parameter.

IDOD MUMM

Statistical Analysis Tool



IDOD MUMM



Statistical Menu

Summary Statistics

Numerical Summary Statistics

Graphical Summary Statistics

Normality Check

Trend Analysis

Correlation Analysis

Correlation Matrix

Scatterplot matrix

Regression Analysis

Multiple Regression

Subset Regression

Spatial Analysis

Variogram Calculation

Variogram Fit

SQC Modelling

Distribution Test

Regression Test

Spatial Test

Import Menu

Send Data to Sat

Reload Data in Sat

Available Subsets

Data Handling Menu

Variable Transformation

Filter the Data Set

View the Data Set

UCS

Universitair Centrum voor Statistiek

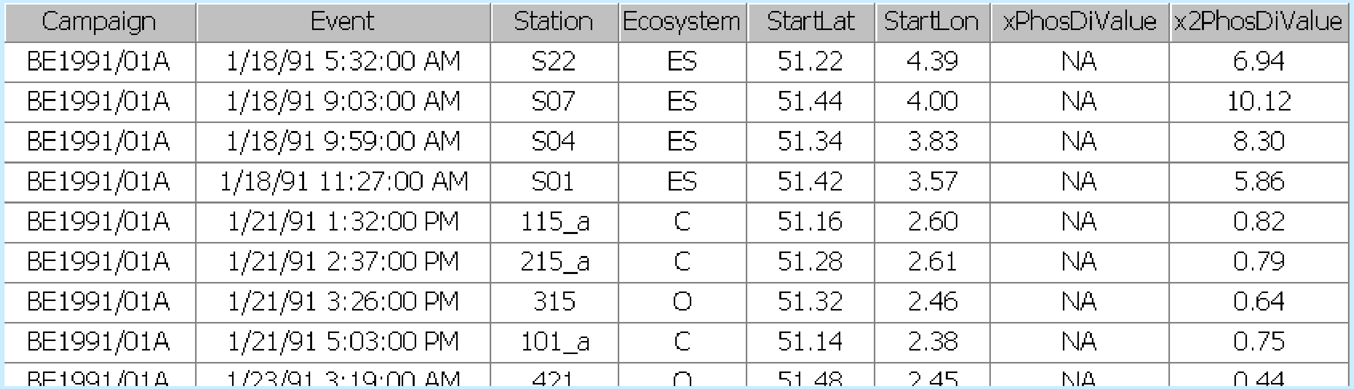
Main Menu

Exit SAT

Universitair Centrum voor Statistiek

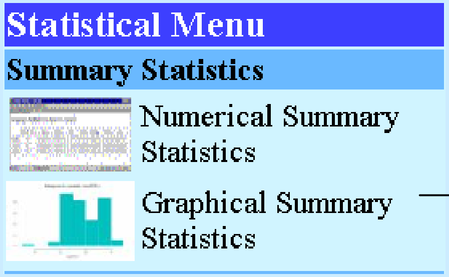
SAT: example phosphorus

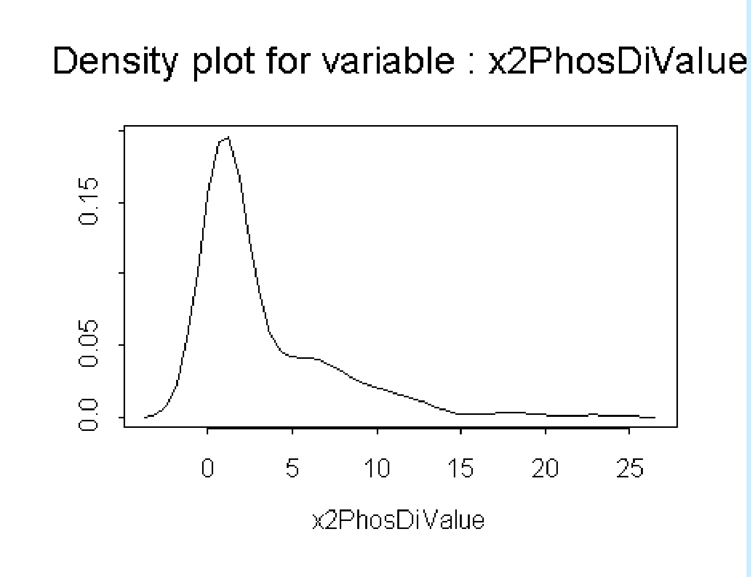
Data set
Wtr44



Campaign	Event	Station	Ecosystem	StartLat	StartLon	xPhosDiValue	x2PhosDiValue
BE1991/01A	1/18/91 5:32:00 AM	S22	ES	51.22	4.39	NA	6.94
BE1991/01A	1/18/91 9:03:00 AM	S07	ES	51.44	4.00	NA	10.12
BE1991/01A	1/18/91 9:59:00 AM	S04	ES	51.34	3.83	NA	8.30
BE1991/01A	1/18/91 11:27:00 AM	S01	ES	51.42	3.57	NA	5.86
BE1991/01A	1/21/91 1:32:00 PM	115_a	C	51.16	2.60	NA	0.82
BE1991/01A	1/21/91 2:37:00 PM	215_a	C	51.28	2.61	NA	0.79
BE1991/01A	1/21/91 3:26:00 PM	315	O	51.32	2.46	NA	0.64
BE1991/01A	1/21/91 5:03:00 PM	101_a	C	51.14	2.38	NA	0.75
BE1991/01A	1/23/91 3:19:00 AM	421	O	51.48	2.45	NA	0.44

Graphical
Summary
Statistics





MUMM

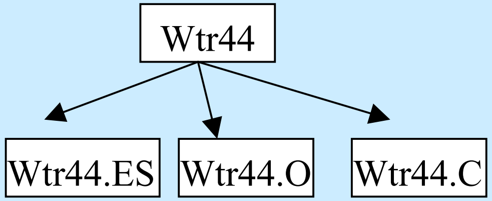
Universitair Centrum voor Statistiek

SAT: example phosphorus

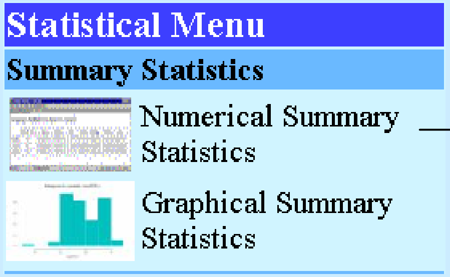
Different Ecosystems

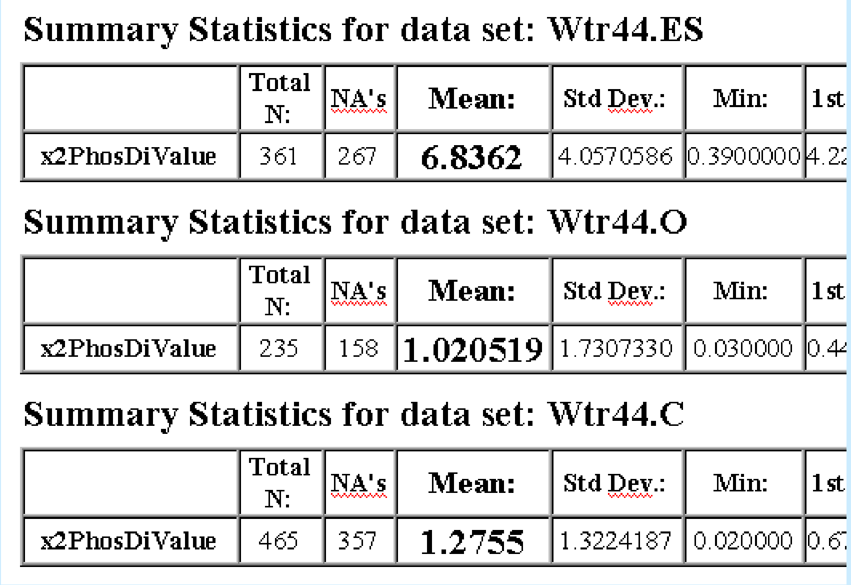
Data Manipulation Menu

- **Variable Transformation**
- **Filter The data Set**
- **View Data Set**



Numerical Summary Statistics





Summary Statistics for data set: Wtr44.ES

	Total N:	NA's	Mean:	Std Dev.:	Min:	1st
x2PhosDiValue	361	267	**6.8362**	4.0570586	0.3900000	4.22

Summary Statistics for data set: Wtr44.O

	Total N:	NA's	Mean:	Std Dev.:	Min:	1st
x2PhosDiValue	235	158	**1.020519**	1.7307330	0.030000	0.44

Summary Statistics for data set: Wtr44.C

	Total N:	NA's	Mean:	Std Dev.:	Min:	1st
x2PhosDiValue	465	357	**1.2755**	1.3224187	0.020000	0.67

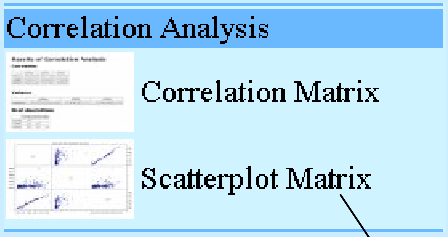
IDOD MUMM

Universitair Centrum voor Statistiek

SAT: example phosphorus

Statistical analysis for ‘ES : Estuary Scheldt

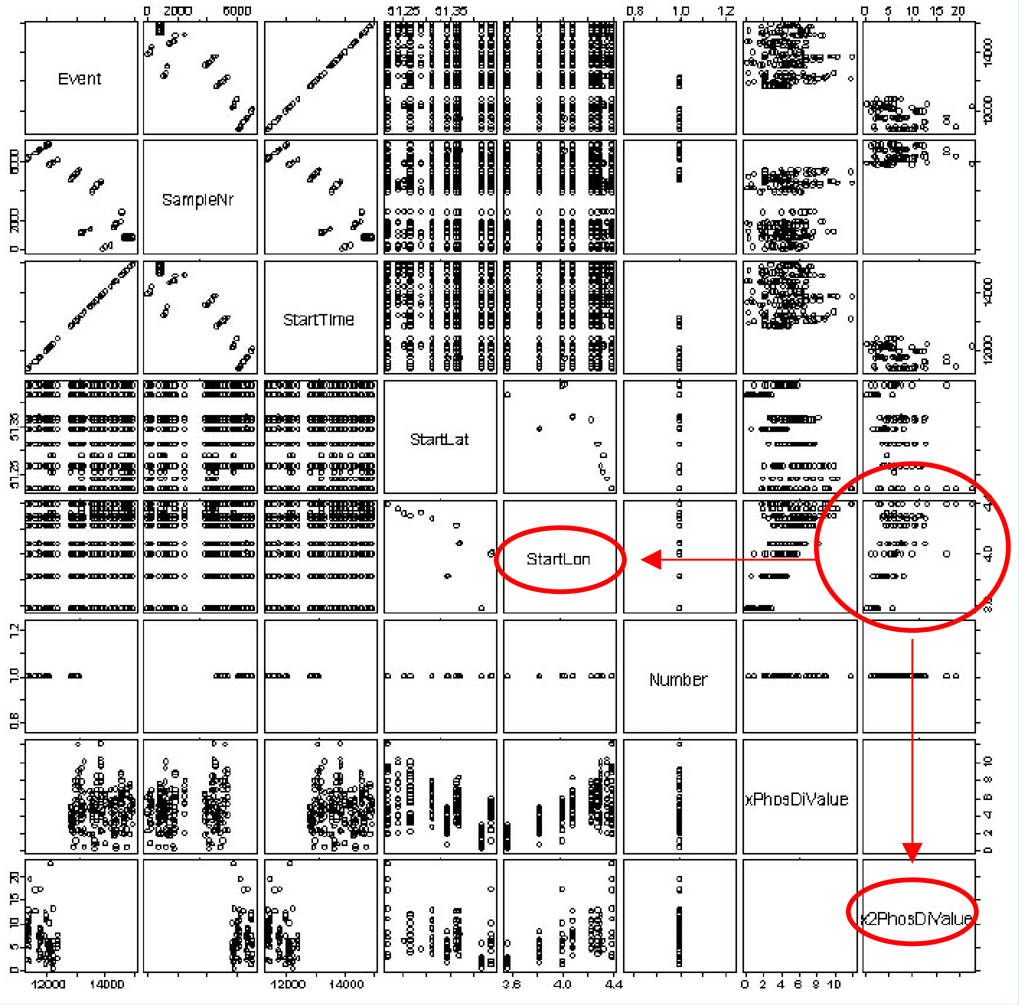
Correlations in the dataset



Correlation Analysis

Correlation Matrix

Scatterplot Matrix



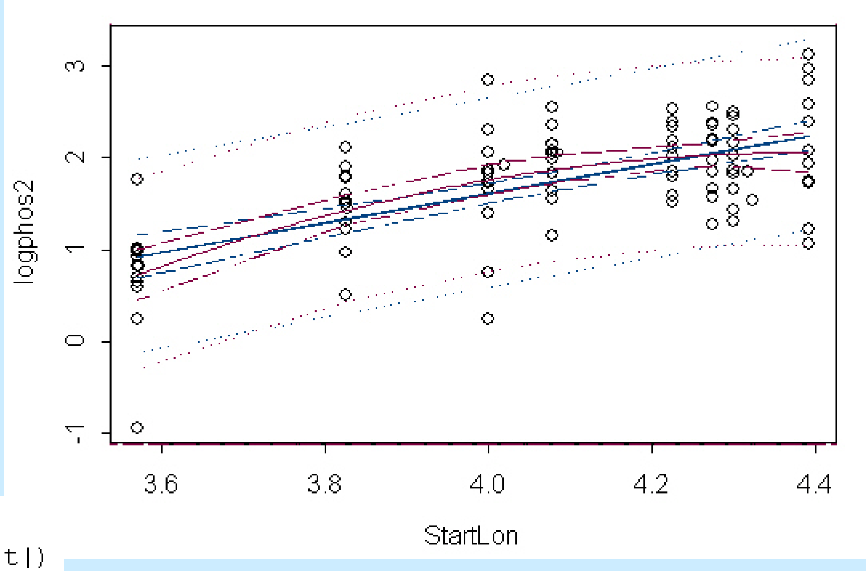
MUMM

Universitair Centrum voor Statistiek

SAT: example phosphorus

Statistical analysis of phosphorus for ‘ES : Estuary Scheldt

Trend Analysis of log(phosph) as a function of Longitude



Linear Regression Results

```
Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept) -4.8138  0.8424    -5.7144  0.0000
       regr  1.6070  0.2062     7.7917  0.0000
```

significant

Quadratic Regression Results

```
              Value Std. Error  t value Pr(>|t|)
(Intercept) -36.8490  13.1012   -2.8127   0.0060
       regr  17.7629   6.5972    2.6925   0.0084
  I(regr^2)  -2.0273   0.8275   -2.4500   0.0162
```

significant

Conclusion : the amount of Phosphorus in the Scheldt changes significantly as the Longitude changes.

Universitair Centrum voor Statistiek

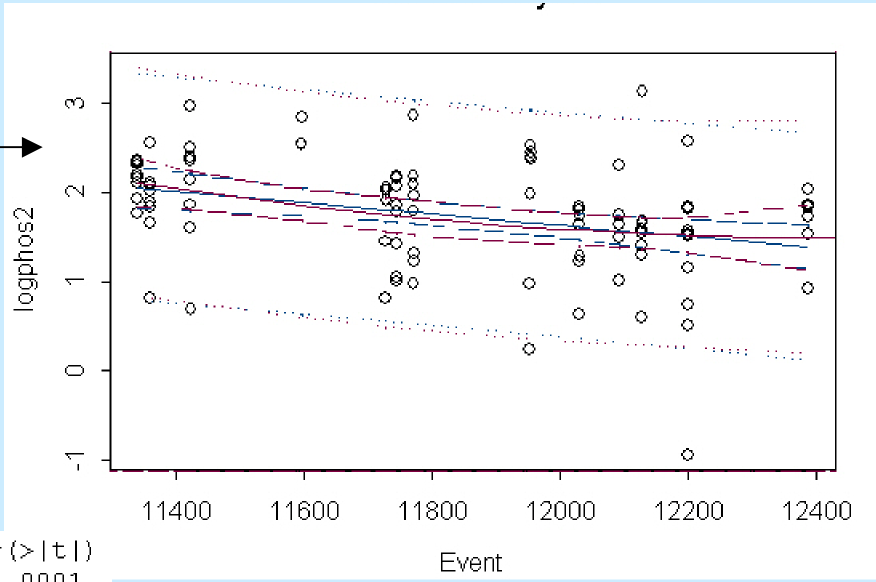
SAT: example phosphorus

Statistical analysis of phosphorus for 'ES : Estuary Scheldt

Trend Analysis of log(phosph) as a function of Event (day+time)

Trend Analysis

Trend Fitting



Linear Regression Results

```
              Value Std. Error t value Pr(>|t|)
(Intercept)  9.3027  2.2864     4.0688  0.0001
       regr -0.0006  0.0002    -3.3105  0.0013
```

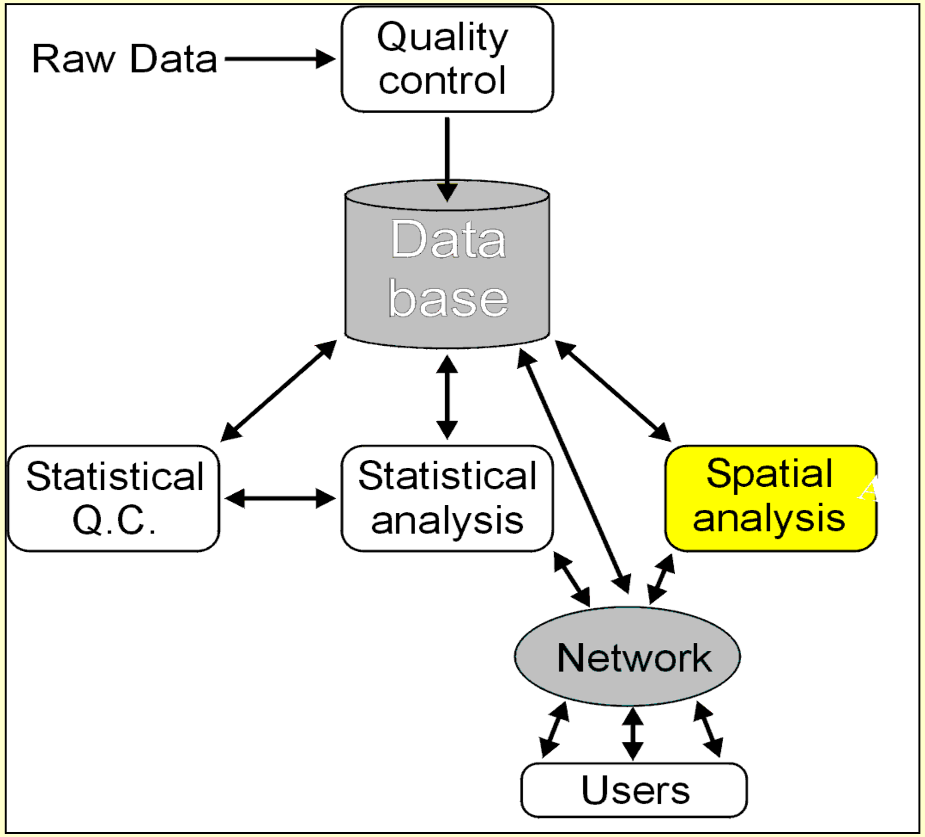
significant

Conclusion : the amount of Phosphorus in the Scheldt diminishes in time. The decrease is small but significant.

Remark : the same analysis for Phosphorus in the open sea and at the coast results not in a significant result.

IDOD MUMM

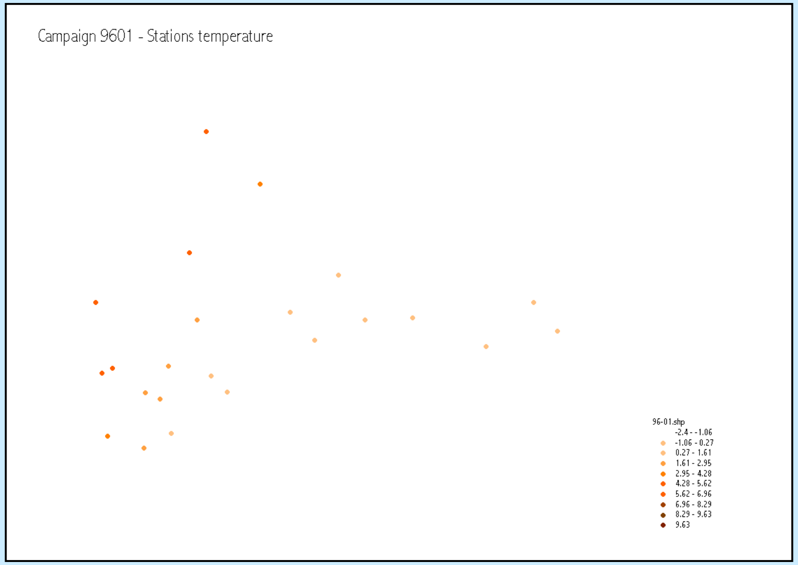
Spatial Analysis Tool



MUMM

GIS Applications

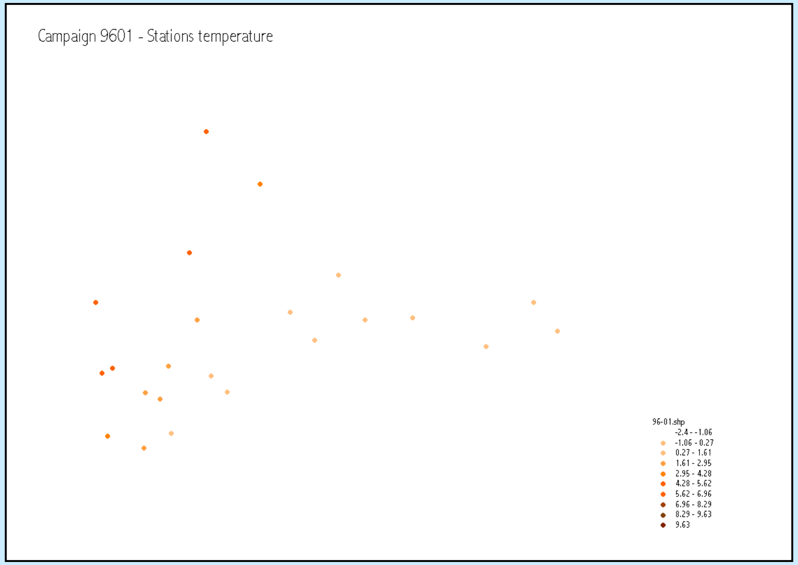
Diversity of geographical objects



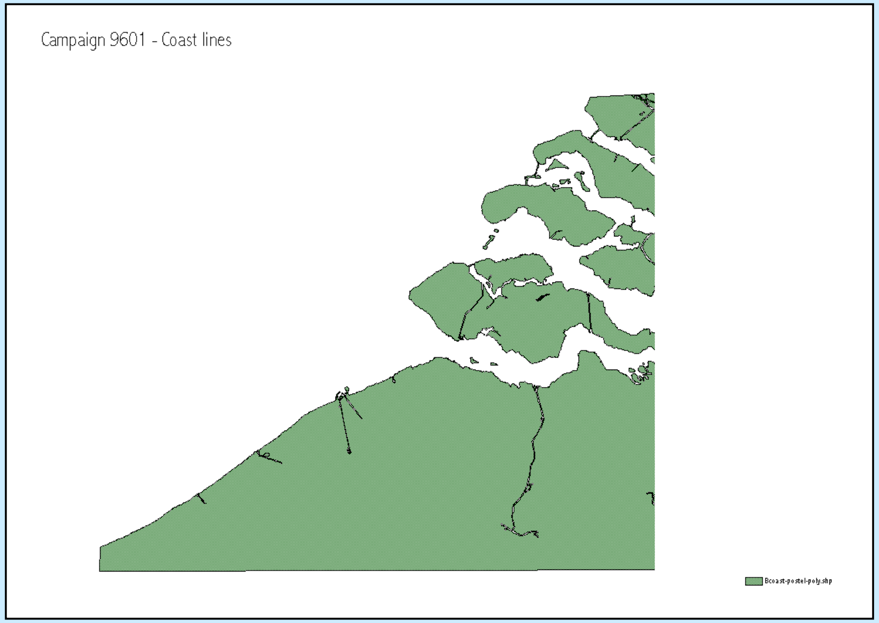
MUMM

GIS Applications

Diversity of geographical objects



Campaign 9601 - Stations temperature



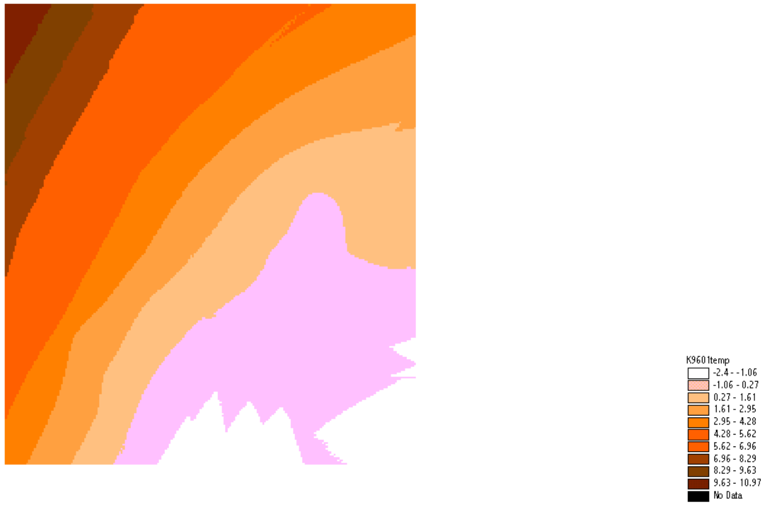
Campaign 9601 - Coast lines

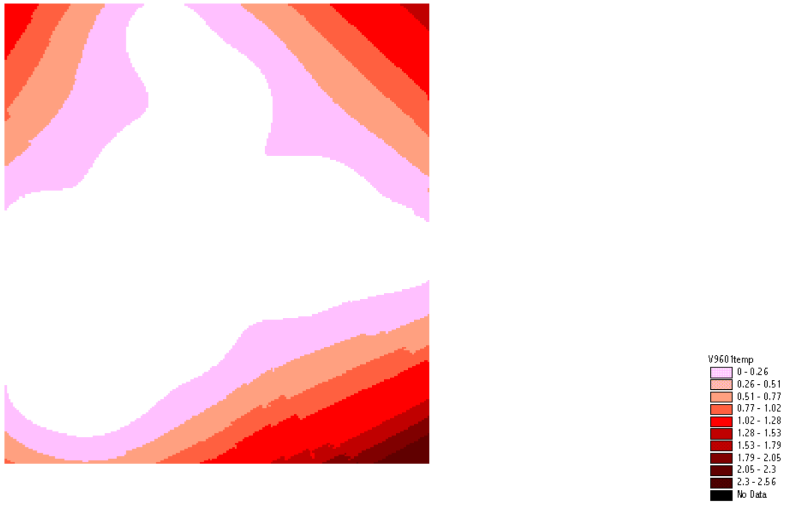
IDOD MUMM

GIS Applications

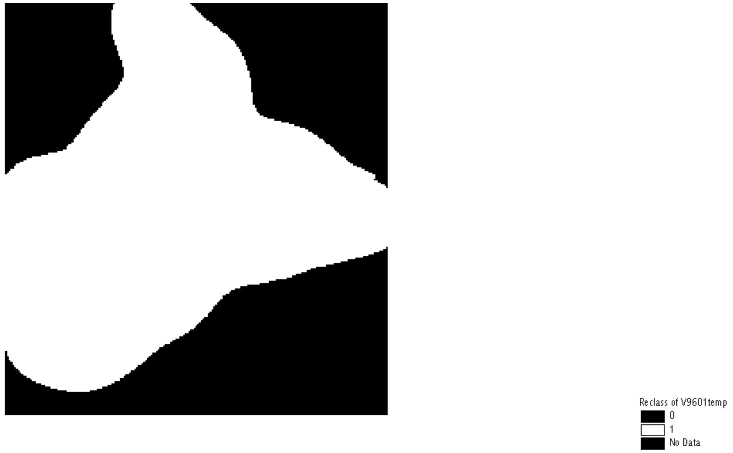
Interpolation results ….

Computed variance





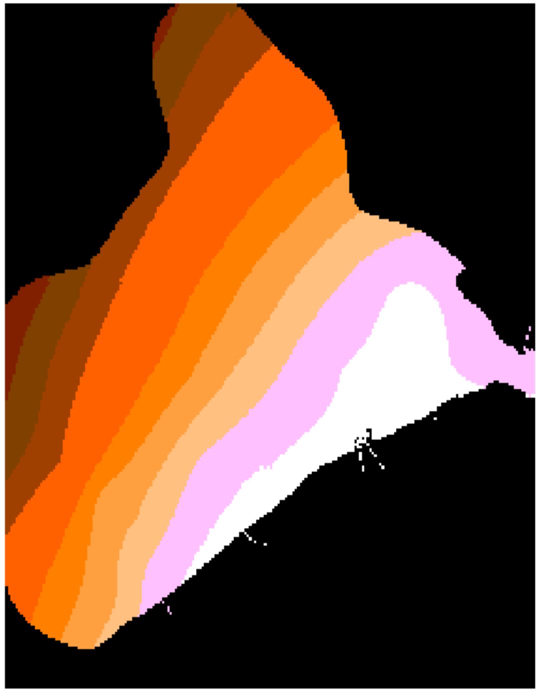
Computed mask to eliminate high variance areas



MUMM

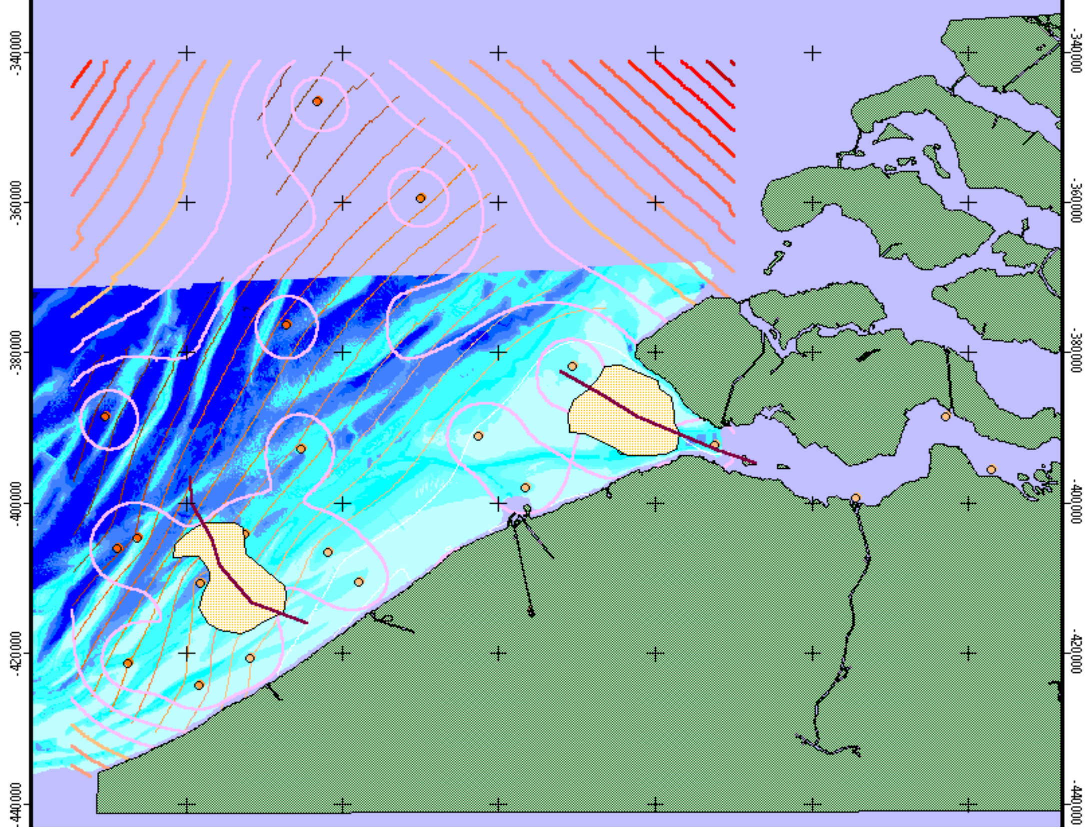
GIS Applications

Computed temperature after application of mask



Superposition of layers (different object types), in the same cartographic reference system

and same accuracy

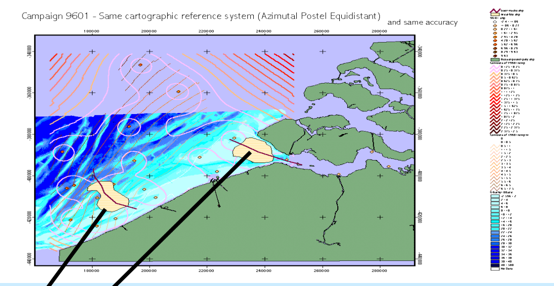


IDOD MUMM

Extraction, processing and aggregation of quantitative and qualitative georeferenced informations

Production of 3 kinds of results (examples)

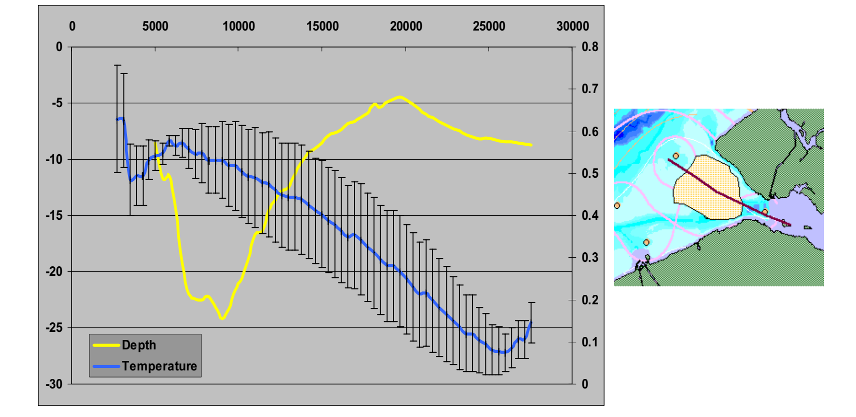
r

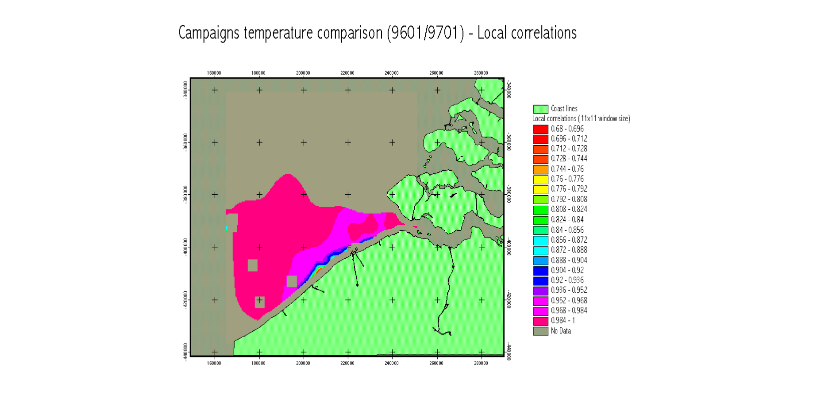


Campaign 9601 - Same cartographic reference system (Azimutal Postel Equidistant) and same accuracy

Aggregation - Zonal statistics - Temperature

Id	Count	Area	Min	Max	Range	Mean	Std	Sum
1	791	126560000	0.9942	3.9606	2.9664	2.2054	0.6792	1744.4856
2	771	123360000	0.1101	0.5658	0.4557	0.3175	0.1196	244.7657





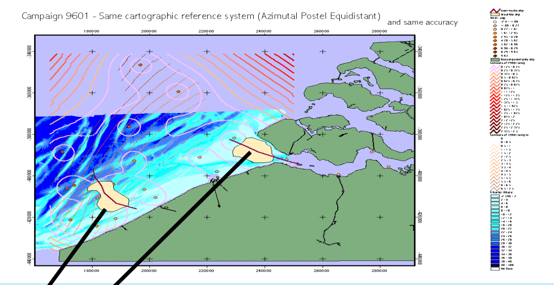
Campaigns temperature comparison (9601/9701) - Local correlations

IDOD MUMM

Extraction, processing and aggregation of quantitative and qualitative georeferenced informations

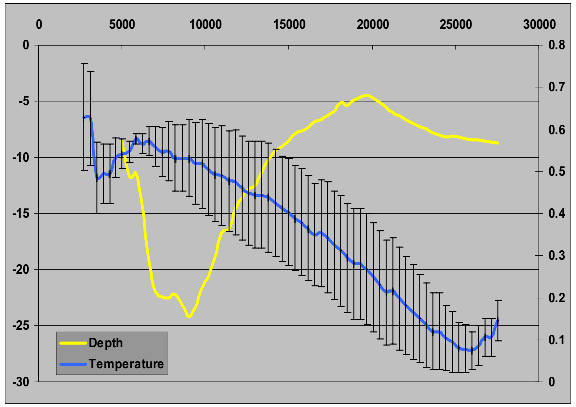
Production of 3 kinds of results (examples)

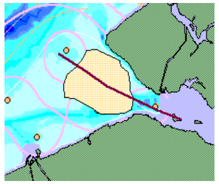
- Tabular

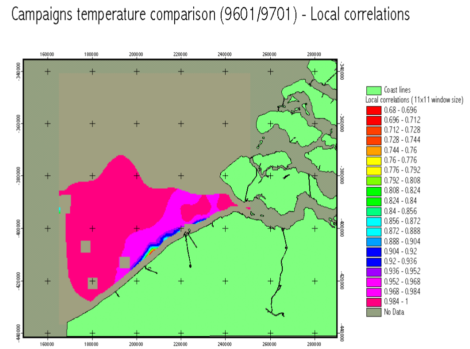

Campaign 9601 - Same cartographic reference system (Azimutal Postel Equidistant) and same accuracy

Aggregation - Zonal statistics - Temperature

Id	Count	Area	Min	Max	Range	Mean	Std	Sum
1	791	126560000	0.9942	3.9606	2.9664	2.2054	0.6792	1744.4856
2	771	123360000	0.1101	0.5658	0.4557	0.3175	0.1196	244.7657





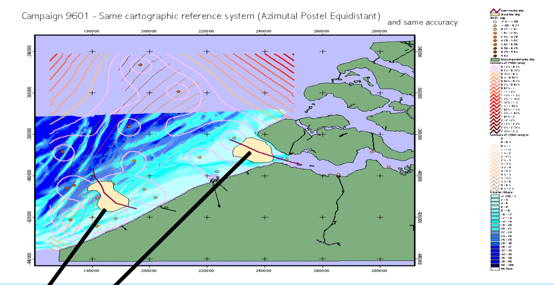


IDOD MUMM

Extraction, processing and aggregation of quantitative and qualitative georeferenced informations

Production of 3 kinds of results (examples)

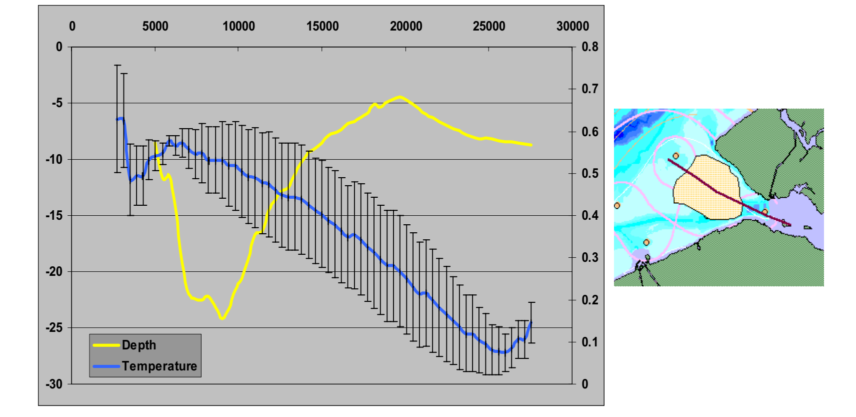
- Tabular


Campaign 9601 - Same cartographic reference system (Azimutal Postel Equidistant) and same accuracy

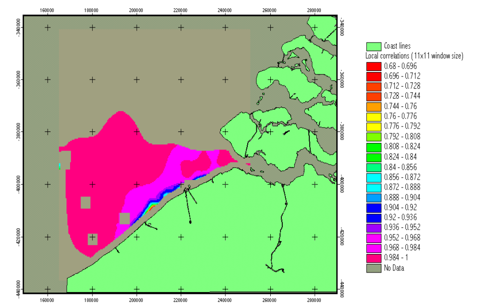
Aggregation - Zonal statistics - Temperature

Id	Count	Area	Min	Max	Range	Mean	Std	Sum
1	791	126560000	0.9942	3.9606	2.9664	2.2054	0.6792	1744.4856
2	771	123360000	0.1101	0.5658	0.4557	0.3175	0.1196	244.7657

- Graphical



Campaigns temperature comparison (9601/9701) - Local correlations

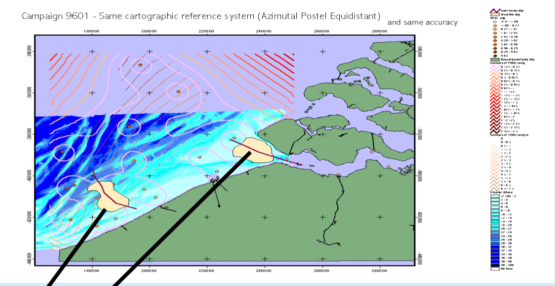


IDOD MUMM

Extraction, processing and aggregation of quantitative and qualitative georeferenced informations

Production of 3 kinds of results (examples)

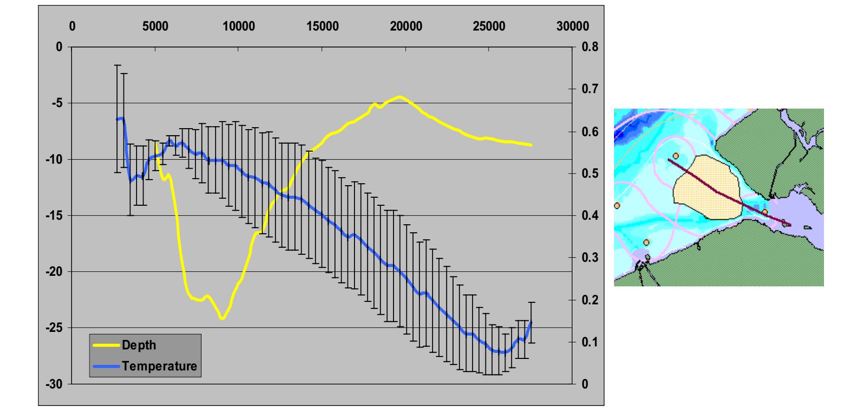
- Tabular


Campaign 9601 - Same cartographic reference system (Azimutal Postel Equidistant)
and same accuracy

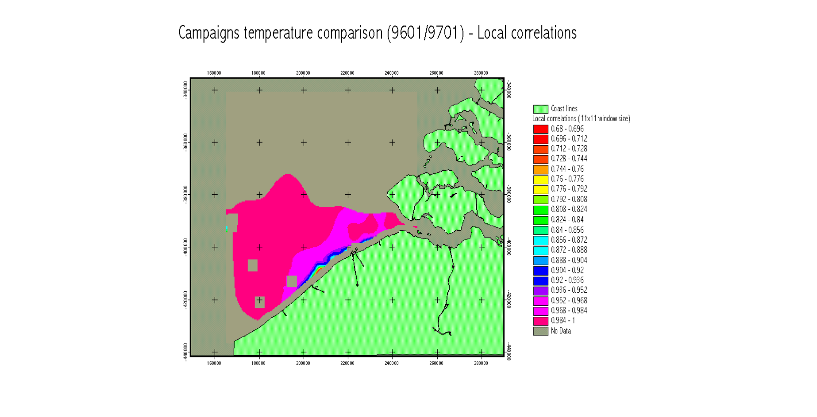
Aggregation - Zonal statistics - Temperature

Id	Count	Area	Min	Max	Range	Mean	Std	Sum
1	791	126560000	0.9942	3.9606	2.9664	2.2054	0.6792	1744.4856
2	771	123360000	0.1101	0.5658	0.4557	0.3175	0.1196	244.7657

- Graphical



- Cartographical

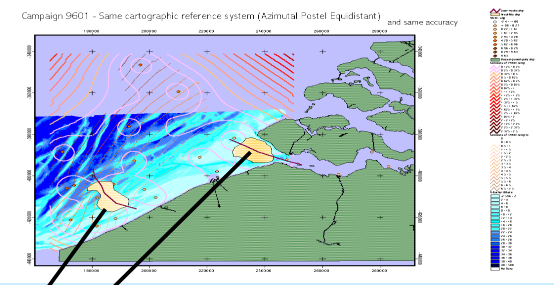

Campaigns temperature comparison (9601/9701) - Local correlations

IDOD MUMM

Extraction, processing and aggregation of quantitative and qualitative georeferenced informations

Production of 3 kinds of results (examples)

- Tabular

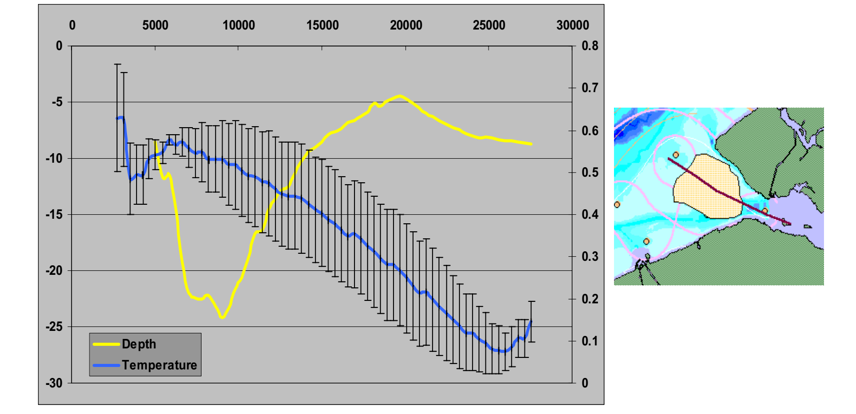

Campaign 9601 - Same cartographic reference system (Azimutal Postel Equidistant)
and same accuracy

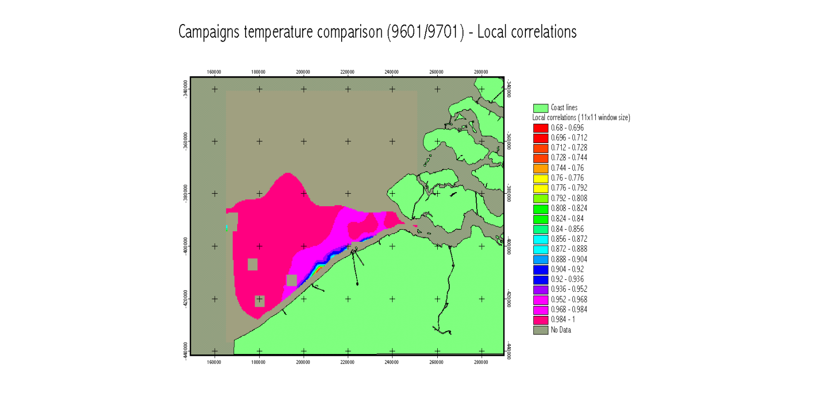
Aggregation - Zonal statistics - Temperature

Id	Count	Area	Min	Max	Range	Mean	Std	Sum
1	791	126560000	0.9942	3.9606	2.9664	2.2054	0.6792	1744.4856
2	771	123360000	0.1101	0.5658	0.4557	0.3175	0.1196	244.7657

- Graphical

- Cartographical



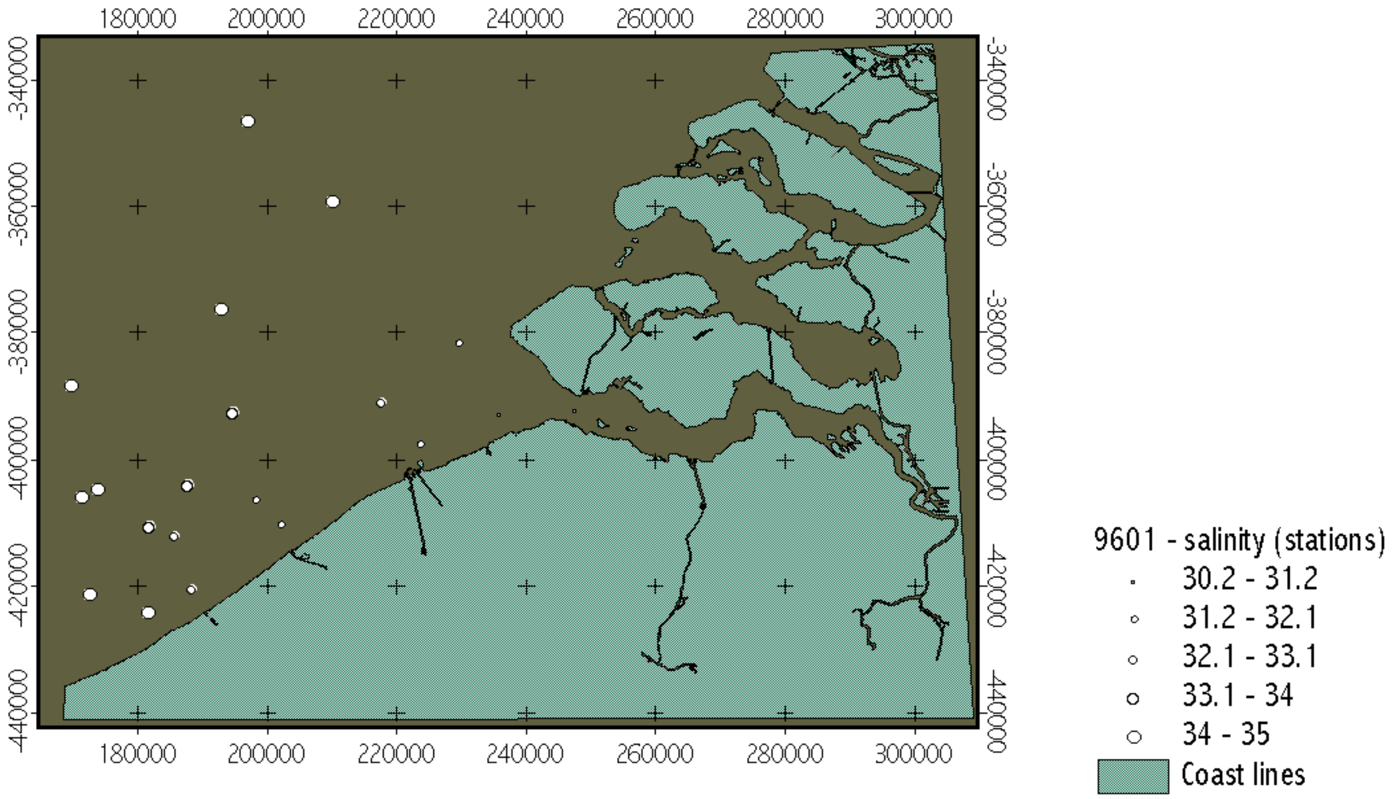

Campaigns temperature comparison (9601/9701) - Local correlations

Interpretation of the user and hypothesis definition

Spatial Analysis

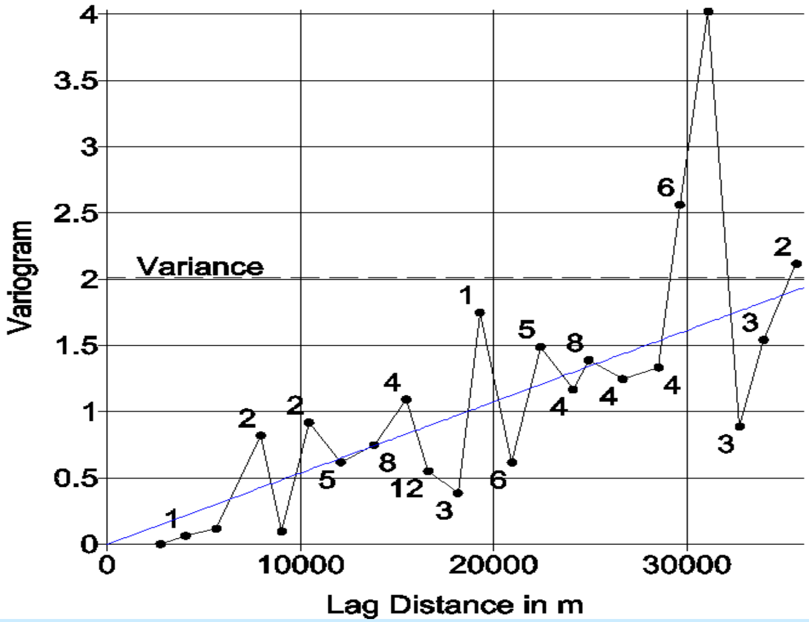
Punctual data set (IDOD)

Campaign 9601 - data - salinity (stations)

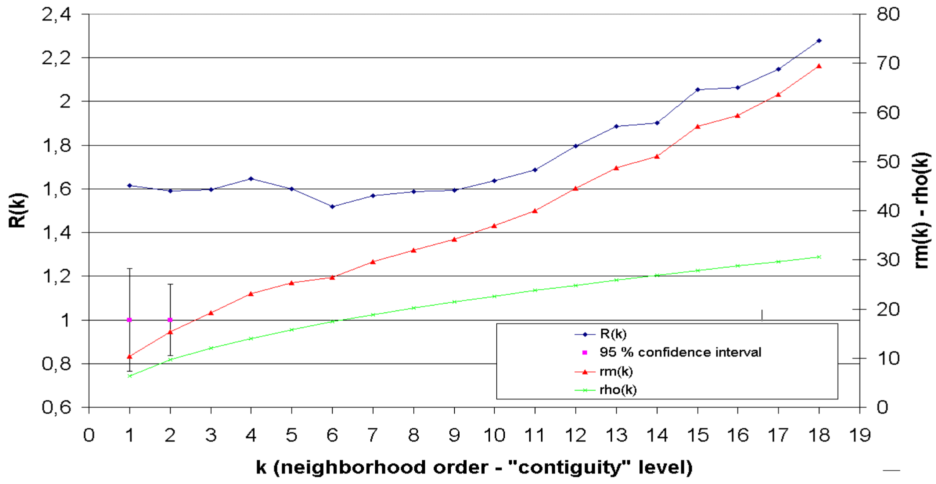


Spatial properties of the data set

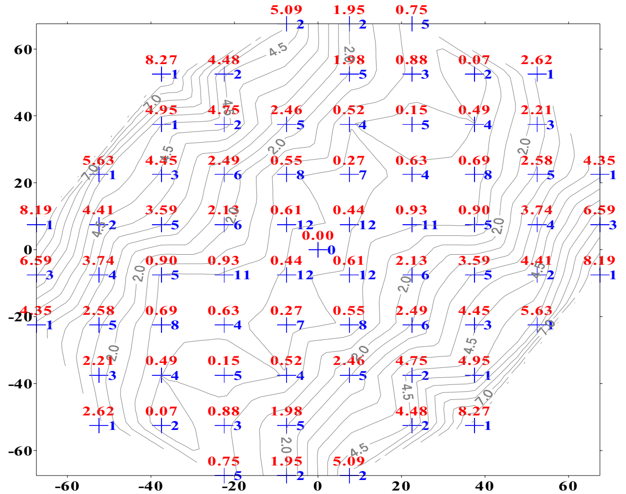
Campaign 9601 - "Psalat31_v"



R-statistics - campaign 9601



Semivariogramme directionnel - Salinité (9601 - stations de mer)

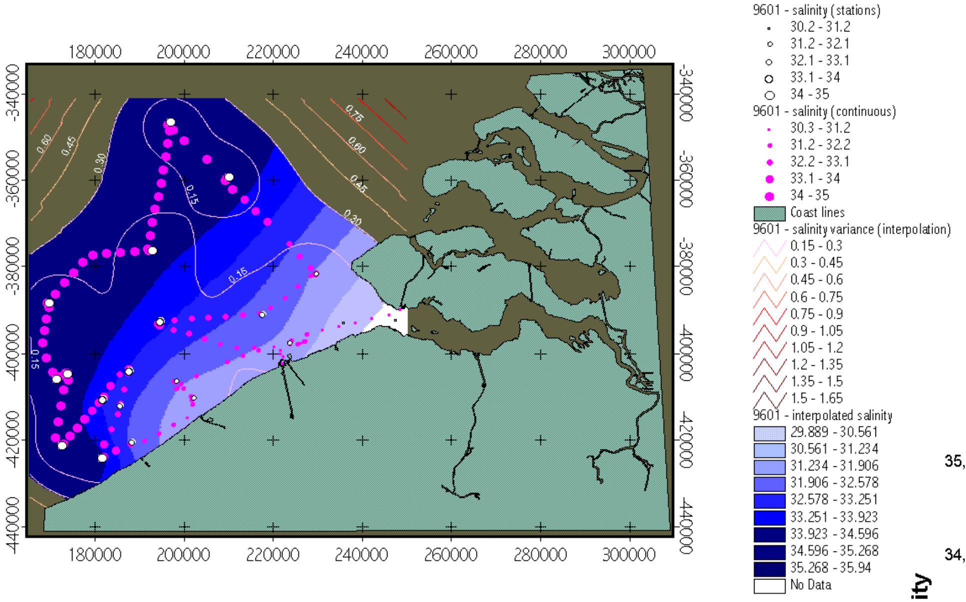


IDOD

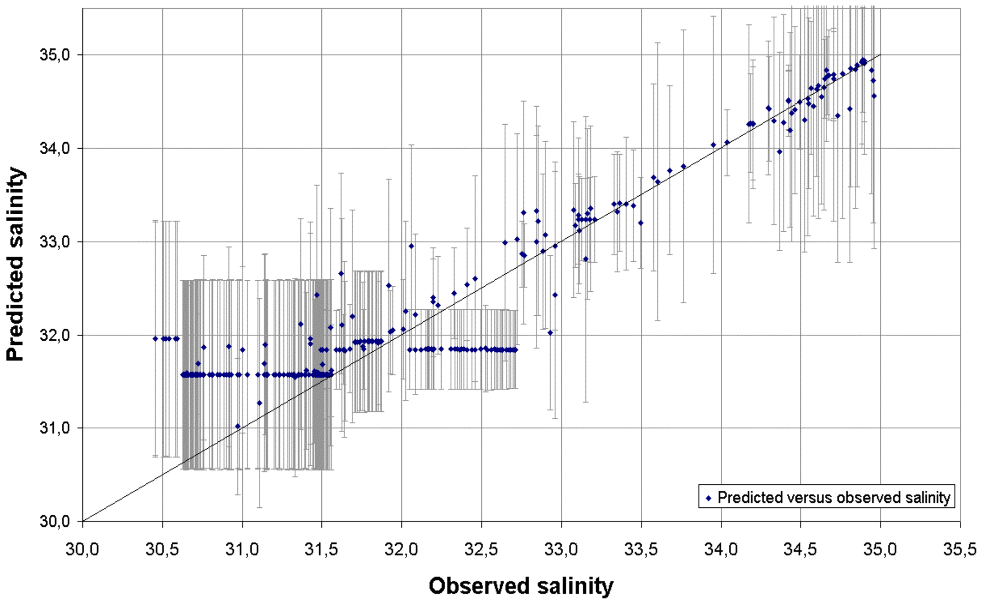
MUMM

Interpolation and validation

Campaign 9601 - interpolation validation using "continuous" salinity measures

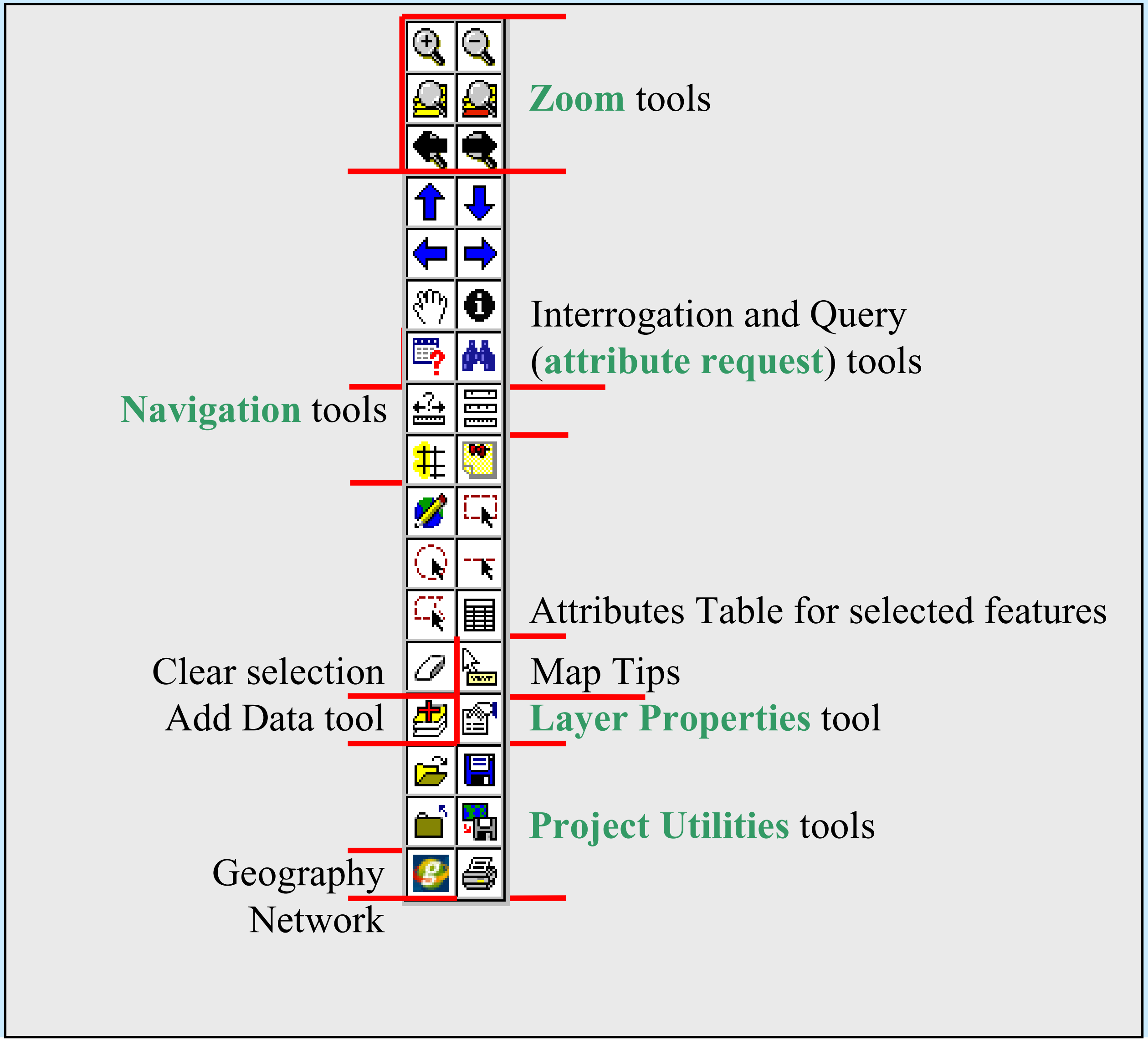


Predicted versus observed salinity and 95% confidence interval



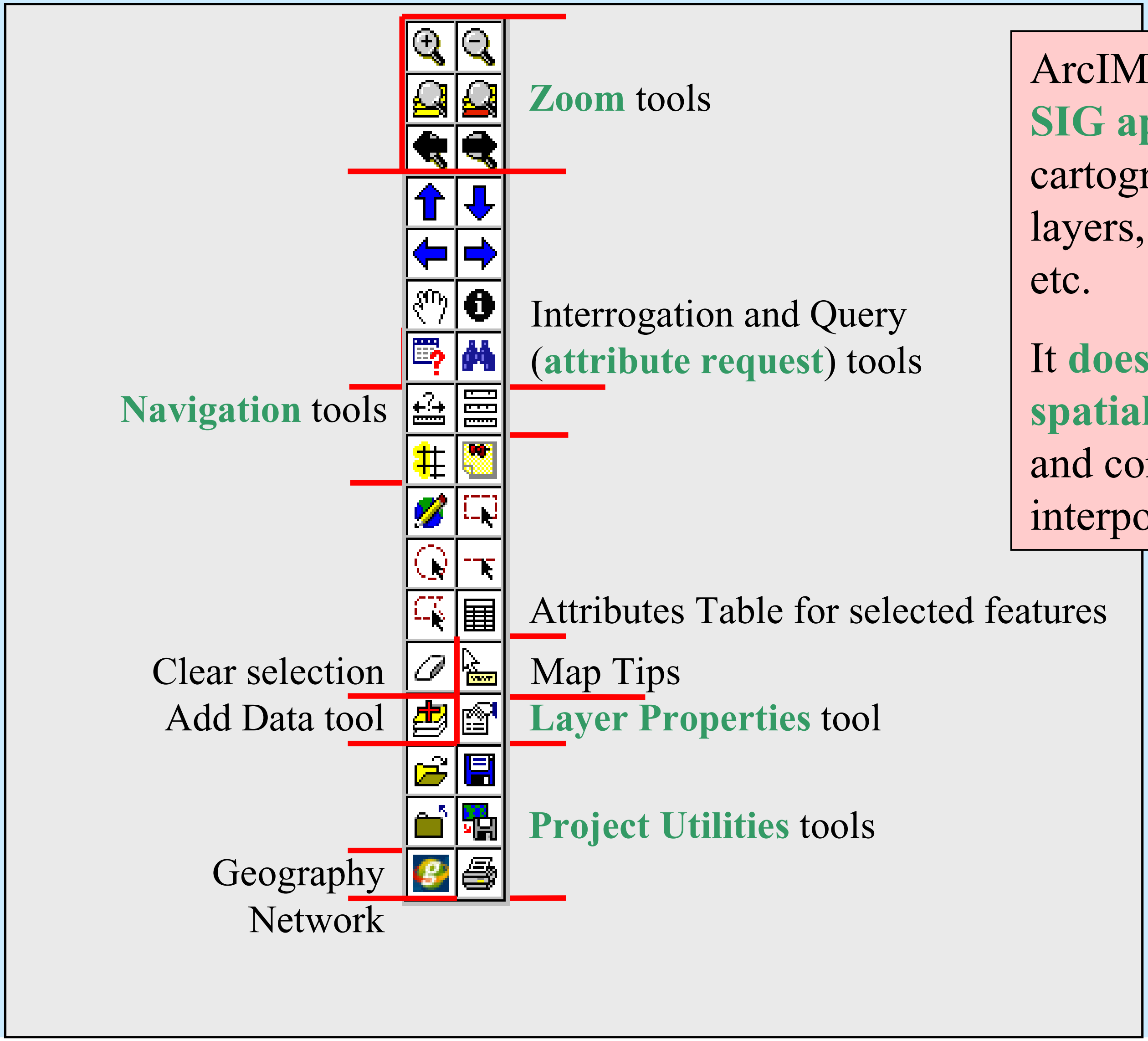
MUMM

Website utilities (ArcIMS)



Zoom tools
Interrogation and Query (**attribute request**) tools
Navigation tools
Attributes Table for selected features
Clear selection
Map Tips
Add Data tool
Layer Properties tool
Project Utilities tools
Geography Network

Website utilities (ArcIMS)


Zoom tools
Interrogation and Query (**attribute request**) tools
Navigation tools
Attributes Table for selected features
Clear selection
Map Tips
Add Data tool
Layer Properties tool
Project Utilities tools
Geography Network

ArcIMS is **sufficient for « light » SIG applications** like cartographying vector and raster layers, performing easy requests, etc.

It **does not fit for complex spatial processes** like combined and complex requests, interpolation, geoprocessing, etc.

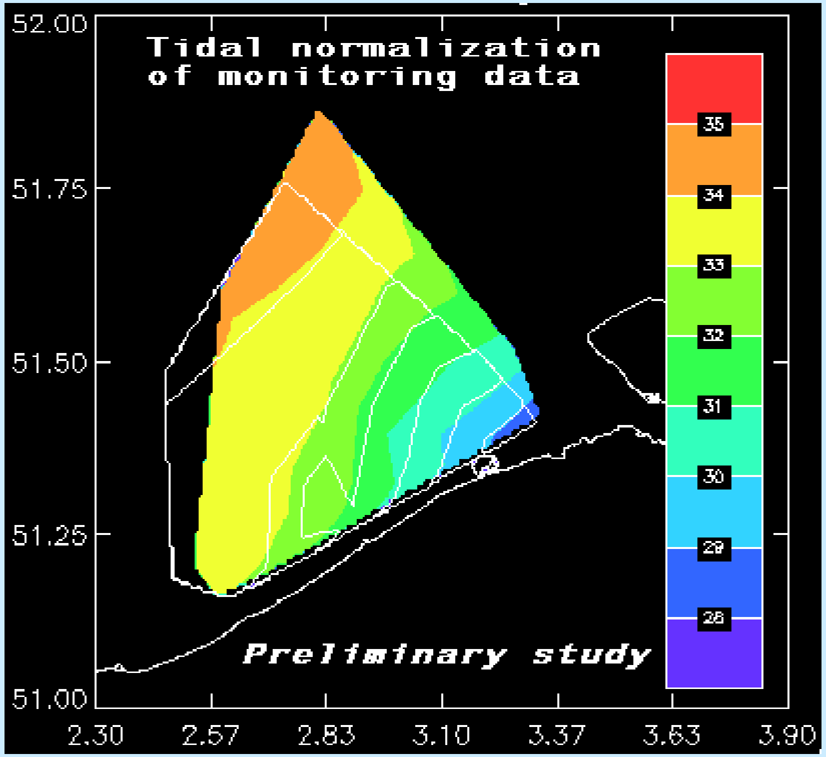
Website possibilities

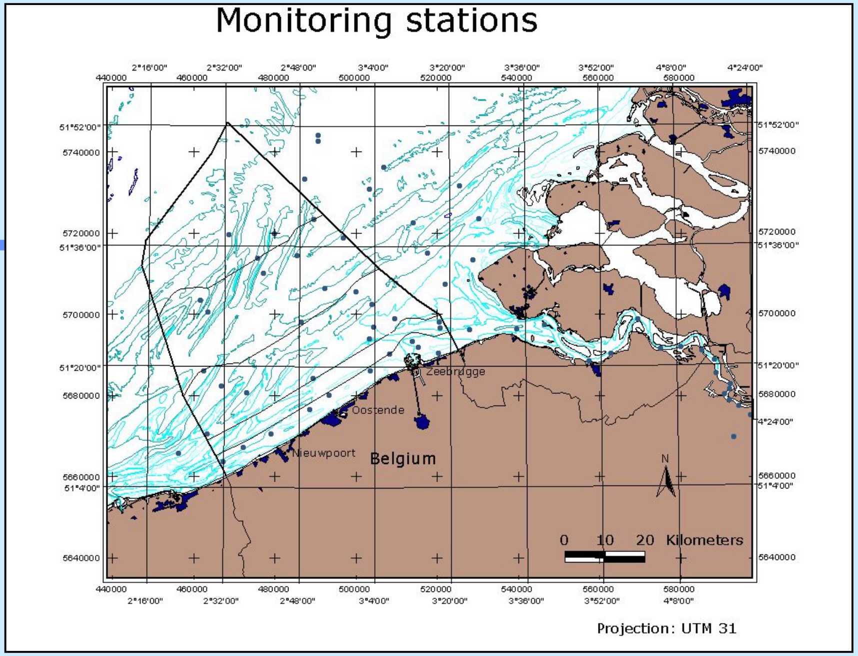
IDOD MUMM

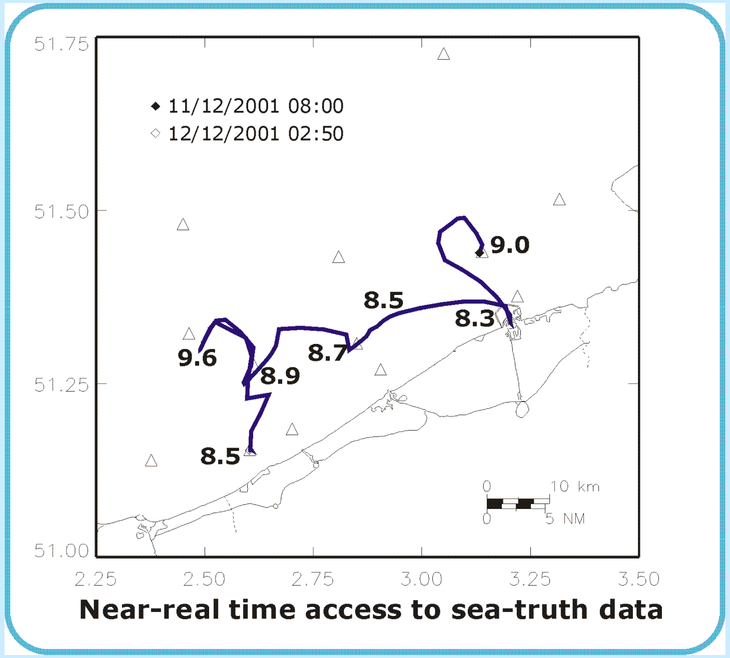
Products and Applications…

IDOD MUMM

Products




Monitoring stations



Near-real time access to sea-truth data

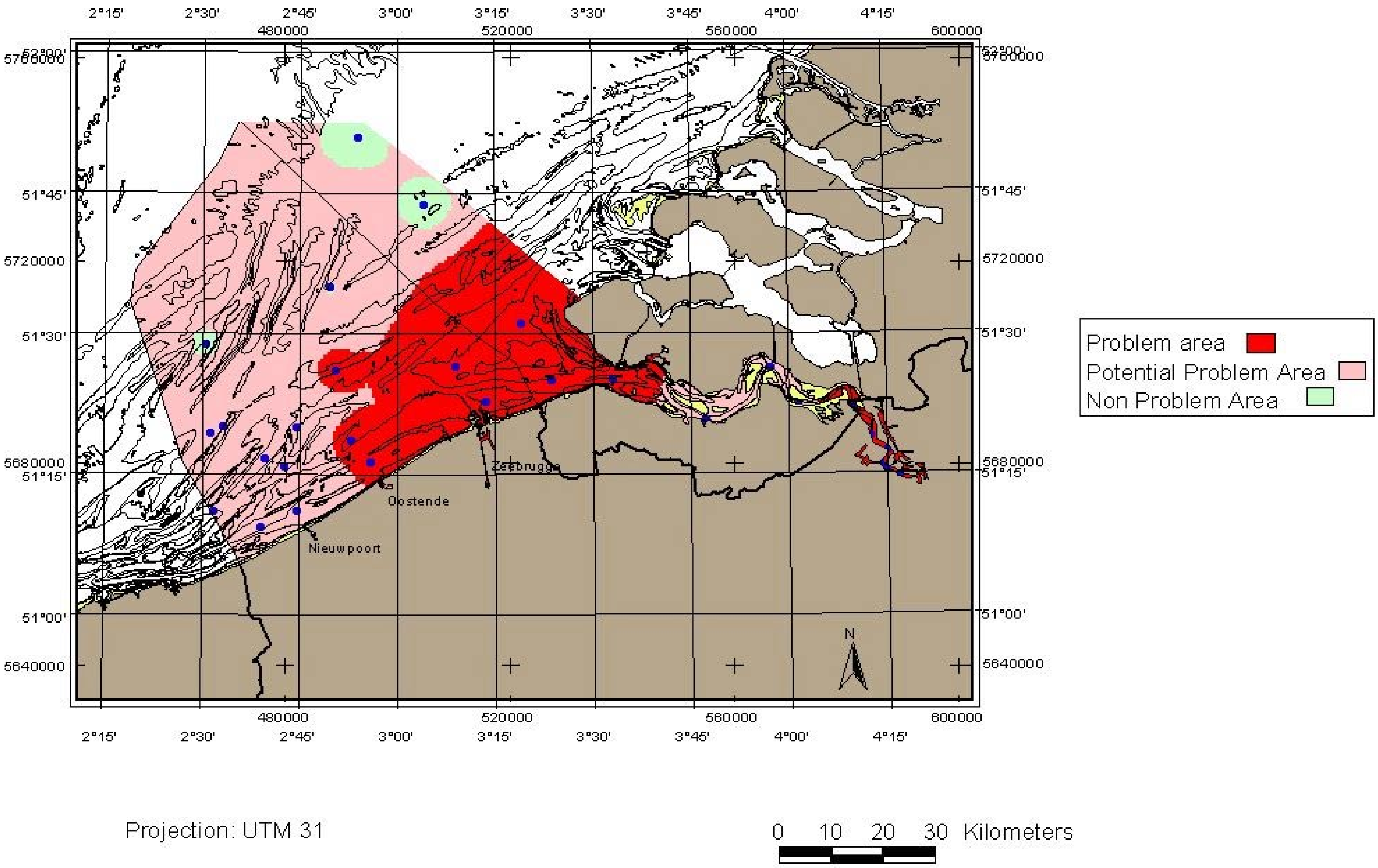
IDOD MUMM

Application

OSPAR Eutrophication assessment criteria

DIN > 15µmol/l AND/OR DIP > 0.8µmol/l (during winter)	CHL A > 15µg/l (during growing season)		O2 < 6mg/l (during growing season)	STATUS
+	+	AND/OR	+	PROBLEM AREA
-	+	AND/OR	+	PROBLEM AREA
+	-		-	POTENTIAL PROBLEM AREA
-	-		-	NON PROBLEM AREA

Eutrophisation assessment 1998



Conclusion

IDOD - Integrated and Dynamical Oceanographic Data Management

- An important experience has been gained…
- Belgian NODC is an efficient project data manager in European projects...

IDOD - Integrated and Dynamical Oceanographic Data Management

BUT IDOD success will be
reached when …
you will use it!

www.mumm.ac.be/datacentre

MUMM

BELGIAN SCIENCE POLICY

WETENSCHAPSSTRAAT 8
RUE DE LA SCIENCE 8
1000 BRUSSELS
www.belspo.be