belspo

BRAIN-be
Belgian Research Action through Interdisciplinary Networks

PIONEER PROJECTS

CHARACTERIZATION OF ACTIVE REGIONS' TIME EVOLUTION IN VIEW OF SOLAR FLARE PREDICTION

CONTRACT - BR/121/PI/PREDISOL

FINAL REPORT

15/12/2016

Promotors
Dr Véronique Delouille (Royal Observatory of Belgium, Avenue Circulaire 3, B-1180 Ukkel)
Prof. Alfred O. Hero (The University of Michigan, 1301 Beal Avenue, Ann Arbor, MI, USA)

Authors
Dr. Véronique Delouille (Royal Observatory of Belgium)
Prof. Alfred O. Hero (The University of Michigan)
Mr. Ruben De Visscher (Royal Observatory of Belgium)
Dr. Kevin Moon (The University of Michigan)
Dr. Raphaël Attié (Royal Observatory of Belgium)
Prof. Pierre Dupont (Université catholique de Louvain)

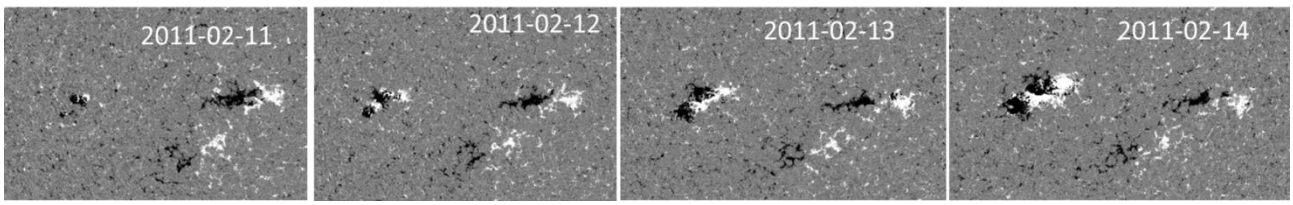

Figure 1 Evolution of two active regions as seen by the magnetogram on board the SDO mission. The AR on the left grew quickly and produced major solar eruptions three days after it appeared.

D/XXXX/XXXX/XX (to complete by Belspo)
Published in 20XX by the Belgian Science Policy
Avenue Louise 231
Louizalaan 231
B-1050 Brussels
Belgium
Tel: +32 (0)2 238 34 11 – Fax: +32 (0)2 230 59 12
http://www.belspo.be

Contact person: XXXXXXXX
+32 (0)2 238 3X XX

V. Delouille, A.O. Hero, R. De Visscher, K. Moon, R. Attié, P. Dupont. ***Characterization of active regions' time evolution in view of solar flare prediction*** Final Report. Brussels : Belgian Science Policy 20XX – xx p. (BRAIN-be - (Belgian Research Action through Interdisciplinary Networks)

TABLE OF CONTENTS

CONTENTS

SUMMARY

Context

Solar flares are the most powerful examples of solar activity. When intense and directed towards the Earth, they may affect the ionosphere and radio communications. Providing a reliable prediction with confidence interval for their onset time is thus crucial for Space Weather applications. In particular it is important to predict as early as possible whether an active region will develop into a flare productive active region. About one out of ten active regions will produce one large flare or more (minority class), whereas nine out of ten will produce no or small flares (majority class), but a higher risk is associated to a wrong prediction of the minority class.

Objectives

The objective of PREDISOL is to characterize the evolution of active regions through quantities computed from magnetogram and continuum images, and to be able to link this evolution to the production of large solar flares. A first objective is to provide an optimal classification between flaring and non-flaring active regions that takes into account the specificities of the problem such as the imbalanced class distribution. The second objective is to consider a set of SOHO-MDI magnetogram and continuum images of active regions, to analyze their correlation structure, and to derive a classification rule to separate 'simple' from 'complex' active regions.

Conclusions

To reach the first objective, we developed a software to automate the extraction of active regions in SOHO-MDI magnetograms and the computation of related predictors of flaring activity. We used support vector machine to separate regions producing at least one X flare from regions producing weaker flares, and were able to classify correctly 3/4th of the active regions producing X-flares. In the second part on the modelling of SOHO-MDI magnetogram and continuum images, we found the active region local correlation structure to be linear and negligible at locations more than 90Mm apart, where 90Mm corresponds to the characteristic lengths of the largest penumbral filaments. We developed a data-driven clustering scheme to separate complex from simple active regions and found overlap between the resulting classification and the Mount Wilson classification.

Keywords

Solar flare, active region, sunspot, prediction, magnetic field, continuum image, supervised classification, unsupervised classification, local correlation analysis

SAMENVATTING

Context

Zonnevlammen zijn de meest krachtige voorbeelden van zonneactiviteit. Wanneer ze intens en naar de Aarde gericht zijn, kunnen ze de ionosfeer en radiocommunicatie beïnvloeden. Het verstrekken van een deugdelijke voorspelling met een betrouwbaarheidsinterval op de voorspelde aanvangstijd is dus cruciaal voor ruimteweertoepassingen. Met name is het van belang zo spoedig mogelijk te voorspellen of een actief gebied zonnevlammen zal produceren. Ongeveer één op de tien actieve gebieden zullen één of meerdere grote zonnevlammen produceren (de minderheidsklasse), terwijl negen van de tien geen of enkel kleine zonnevlammen (de meerderheidsklasse) zal produceren, maar aan een verkeerde voorspelling van de minderheidsklasse is wel een hoger risico verbonden.

Doelstellingen

Het doel van PREDISOL is om de ontwikkeling van actieve gebieden te karakteriseren door de evolutie te volgen van eigenschappen berekend uit magnetogram en continuüm waarnemingen, en om deze evolutie te koppelen aan de productie van grote zonnevlammen. Een eerste doelstelling is om een optimale indeling te voorzien tussen actieve gebieden die wel of niet grote zonnevlammen zullen produceren, rekening houdend met de specifieke kenmerken van het probleem zoals de onevenwichtige verdeling tussen de beide klassen. De tweede doelstelling is om een set van SOHO-MDI-magnetogrammen en continuümbeelden van actieve gebieden te beschouwen, hun correlatiestructuur te analyseren, en een classificatieregel op te stellen om 'eenvoudige' van 'complexe' actieve gebieden te scheiden.

Besluiten

Om de eerste doelstelling te bereiken, ontwikkelden we software om de identificatie van actieve gebieden in SOHO-MDI magnetogrammen te automatiseren, en daaruit voorspellingsparameters voor de berekening van zonnevlamactiviteit. We gebruikten "Support Vector Machine"-technieken om gebieden die ten minste één X-klasse zonnevlam produceren, te scheiden van gebieden die zwakkere zonnevlammen produceren. We waren in staat om ¾ van de actieve gebieden die X-zonnevlammen produceren correct te classificeren. In het tweede deel , wat betreft het modelleren van SOHO-MDI-magnetogrammen en continuümbeelden, vonden we dat de lokale correlatiestructuur van actieve gebieden, lineair is en verwaarloosbaar op plaatsen meer dan 90Mm van elkaar. 90Mm komt overeen met de karakteristieke lengte van de grootste penumbrale filamenten. We ontwikkelden een data-driven clusteringregel om complexe van eenvoudige actieve gebieden te scheiden en vonden een overlap tussen de daaruit voortvloeiende classificatie en de Mount Wilson-classificatie.

Trefwoorden

Zonnevlam, actief gebied, zonnevlek, voorspelling, magnetisch veld, het continuüm, gesuperviseerde classificatie, ongecontroleerde classificatie, lokale correlatieanalyse.

RESUME

Contexte

Les éruptions solaires de type 'flare' sont les manifestations les plus puissantes de l'activité solaire, pouvant aller jusqu'à perturber les communications radio et ionosphériques. En météorologie de l'espace, on doit prédire de manière fiable lorsqu'une telle éruption va se produire. Cela revient à prédire le plus tôt possible si une région active va se développer en produisant des éruptions de grandes ampleurs ou non. Environ une région active sur dix produira lors de son existence une ou plusieurs éruptions de grande ampleur (classe minoritaire), tandis que neuf régions sur dix ne produiront pas d'éruptions ou seulement des éruptions de faible amplitude (classe majoritaire). Cependant, un risque plus élevé est associé à une mauvaise prédiction de la classe minoritaire.

Objectifs

L'objectif de PREDISOL est de caractériser l'évolution des régions actives afin de pouvoir relier cette évolution à la production (ou non) d'éruptions solaires de large amplitude. Cette caractérisation s'effectue via des quantités calculées à partir de magnétogrammes et d'images en lumière blanche. Un premier objectif est de fournir une classification optimale entre les régions actives susceptibles de produire des éruptions de grande ampleur et les régions plus calmes, tout en tenant compte du déséquilibre au niveau du nombre de régions actives appartenant à ces deux catégories. Le second objectif est de considérer un ensemble de régions actives observées par les magnétogrammes et images continues de SOHO-MDI, d'analyser sur la structure de leur corrélation locale, et de dériver une règle de classification pour séparer les régions actives 'simples' des régions 'complexes'.

Conclusions

Pour atteindre le premier objectif, nous avons développé un logiciel qui automatise l'extraction des régions actives dans les magnétogrammes de SOHO-MDI ainsi que le calcul des prédicteurs de l'activité solaire. Nous avons utilisé une machine à vecteur de support (SVM) pour séparer les régions produisant au moins une éruption de type X des régions produisant des éruptions plus faibles. Cette classification SVM a permis de classer correctement les trois quarts des régions actives produisant des X-flares. Dans la deuxième partie concernant la modélisation des images de magnéogramme et en lumière blanche de SOHO-MDI, nous avons établi que la corrélation locale dans les taches solaires était une fonction linéaire et qu'à des distances de 90Mm ou plus elle devenait négligeable, ce qui correspond aux longueurs caractéristiques des plus grands filaments pénumbraux. Nous avons développé un schéma de regroupement pour séparer les régions actives 'complexes' des régions 'simples' et avons mesuré la correspondance entre cette nouvelle classification et la classification de Mount Wilson.

Mots-clés

Eruption solaire de type 'flare', région active, tache solaire, prédiction, champ magnétique, image de continuum, classification supervisée, classification non supervisée, corrélation locale.

1. INTRODUCTION

Solar flares are large and abrupt increases of the photon flux, observed in a wide spectral range of electromagnetic radiation. They result from the sudden release of stored magnetic energy, which may reach up to 10^{26} Joules in a matter of a few hours or even minutes. Most flares occur within so-called Active Regions (ARs), that is, regions with locally increased magnetic flux. Intense flares directed towards the Earth may affect the ionosphere and radio communication.

The main purpose of PREDISOL is to find criteria for differentiating active regions that will produce a strong flare from those who will not.

Towards this goal, a first objective is to characterize the evolution of ARs through quantities computed from magnetograms, and to be able to link this evolution to the production of large solar flares. Two work packages were dedicated to this task: *WP2-Preparation of dataset*, and *WP3-Supervised classification of active regions*. In these, we used three day long photospheric measurements of ARs in their growth phase, in order to enable a prediction up to three days in advance. The work in WP2 parts in two: first we implemented a software framework and processing pipeline that allows building sequences of ARs and the computation of relevant properties, and in a second part we computed novel predictors using a fine-scale tracking method called the 'magnetic balltracking'.

The second objective of this project is to model jointly magnetogram and continuum images of ARs in order to retrieve as much information as possible. This was accomplished in *WP4-Analysis of spatial-temporal process*. We proceeded in two steps. First, we analysed the local and modal correlation structure. Second, we built a data-driven clustering scheme of AR using dictionary learning.

2. METHODOLOGY AND RESULTS

2.1 PREDISOL framework and supervised classification of active regions (WP2 and WP3)

PREDISOL framework

An innovative processing pipeline was developed to automate the extraction of regions of interest (also called "patches") in SOHO-MDI magnetograms (Figure 2). We paid special attention to the definition of active regions, the projection effect, and the selection of active regions in their growing or declining phase.

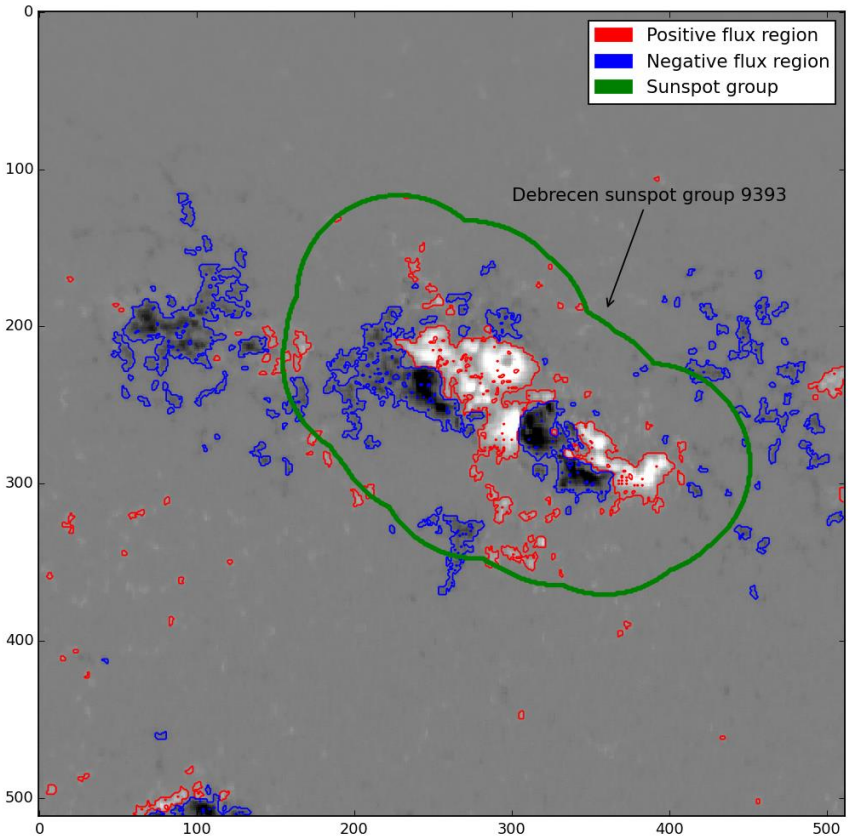

Figure 2 Representation of Debrecen sunspot group 9393 observed on 2001-03-28 at 12:48:30 together with the magnetic fragments detected.

The processing pipeline:

- Imports metadata of the SOHO-MDI magnetograms, as well as active region (NOAA Solar Region Summary and Debrecen sunspots) and flare catalogues (NOAA event catalogue), into an SQL database. This enables advanced querying of and linking the different kinds of data. One can for example query the database for a list of all magnetogram observations of active regions that produced at least one M-class or above flare.
- Produces "patches" of magnetogram data for each active region. These are square or rectangular cutouts of the magnetogram data centered around a given active region. Prior to producing these patches, the extent of the active region must be calculated. For this we use the Debrecen sunspot catalogue (ftp://fenyi.solarobs.unideb.hu/pub/DPD/data/), as it provides the location and size of each observed sunspot in the active region and hence allows us to estimate accurately the size and shape of the active region. These patches are computed using a coordinate transform that corrects for spherical projection effects. More precisely, we transform the images from helioprojective cartesian coordinates to heliographic coordinates. To obtain accurately reprojected images, we implemented the optimized image resampling method described in (DeForest, 2004). This method dynamically adjusts the amount of resampling that is needed based on the coordinate transform

that is being applied and has the option of preserving flux, which is a highly desirable property for our application.

- Detects and segments magnetic fragments, that is, regions of high magnetic flux, in the magnetic AR patches. We use a downhill method to achieve this. The properties of these fragments are used to compute predictors for flaring activity in Step 4. Figure 2 illustrates the fragment detection grouped by polarity on one active region. The center location of each spot reported in the Debrecen catalog was dilated by 64 pixels to obtain the green contour.
- Calculates timelines of predictors of flaring activity based on magnetic patches and fragments.

The main predictors computed with this method are:

- Area and intensity of magnetic flux in the ROI
- R-value: total flux near the polarity inversion line (aka "neutral line")
- Length of the neutral line
- Flux gradients across the neutral line
- Absolute total flux.

An increased absolute flux and area are indicators of a growing active region, and the developed tools allow hence selecting ARs in their growing phase.

The PREDISOL framework also contains a web interface that allows data exploration and visualization. It provides an overview of all active regions, sorted by flaring activity, as well as detailed information on each active region such as flare counts, movies of magnetogram observations, and plots of predictor timelines.

Innovation: The originality of the PREDISOL framework lies in the ability to follow an AR while crossing the solar disk, and in the use of past information up to three days in advance. This is in contrast to most existing methods that use only information at a given time to make prediction on the next 24h or 48h, see e.g. (Ahmed, et al., 2013)

New set of predictors using magnetic ball-tracking (WP2)

In order to enable more in-depth solar physics analyses we developed "stand-alone" modules that do not require running the whole PREDISOL framework as long as we have a series of magnetograms and coordinates of a given AR.

The PREDISOL framework makes it possible to compute advanced physics-based predictors using a new tracking technique called *Magnetic Balltracking* recently developed in (Attié & Innes, 2015). It was originally designed for the tracking of magnetic fragments with low magnetic flux in the quiet photosphere. It was adapted here for application on ARs so we can track the magnetic fragments of each AR almost at pixel-level accuracy, see Figure 3. This

provides improved (with respect to the ones above) and new local measurements in each AR such as:

- Positions and velocity of each magnetic fragment,
- Automatic detection of emerging ARs
- Barycenter of all, or selected groups of magnetic fragments, enabling estimation of induced and mutual helicity
- Lifetime and area of each magnetic fragment
- Univocal identification of the magnetic fragments
- Tracking of the neutral lines, resulting in:
    - Flux-weighted positions of the neutral line
    - Univocal identification of each neutral line in the AR
    - Accurate computation of neutral line length

Innovation: The unprecedented accuracy delivered by the magnetic balltracking method opens the door to the computation of physically relevant predictors such as helicity. Those have not been previously included in comprehensive flare forecasting studies because of the difficulty to compute them. Given the non-trivial projection error present in line of sight magnetograms, it is important to choose the most precise method when computing any predictors. In this aspect, Magnetic Balltracking combined with the automation strategy behind the PREDISOL framework will considerably improve on the accuracy of predictors commonly used in the literature

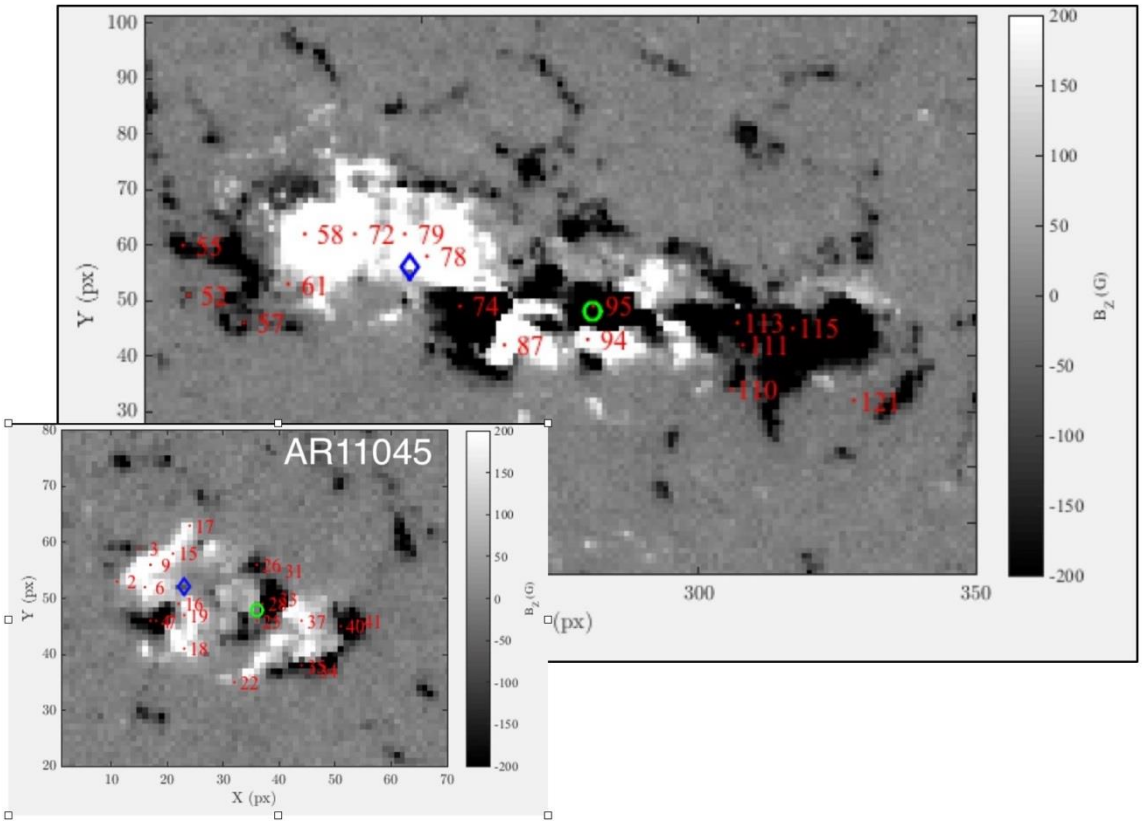

Figure 3 Univocal identification of the tracked magnetic fragments within the early phase of development of an AR observed on February 6, 2010. The bottom left panel is the AR at the early stage of his detection, overlaid on scale over the same AR at a later phase, when the AR has evolved and grown. The blue diamond and green circles are the flux-weighted positions of the fragments of each polarity (blue for positive, green for negative flux). The flux-weighted position of these two barycenters acts as the barycenter for the whole AR, while their relative displacements is used to quantify the average magnetic stress in the AR.

Supervised classification of active regions using time-line of predictors (WP3)

The PREDISOL framework allows us to compute time line of predictors for any active region observed by SOHO-MDI. In order to decide whether an AR was actively producing large flares, we used the *total importance of flares*, computed as a weighted sum of the number of flares that an AR has produced until a certain date. The weights depend on the strength of the flares: it is equal to 1 for C-flare, 10 for M-flare, and 100 for X-flare. We consider an AR as flare-productive when its total importance of flare is larger than 100. For each type of predictors, we computed a maximum of 45 values (3 days at MDI cadence) either just before the total importance of flares produced by this AR reached 100, or before the AR reached its maximum in absolute total flux.

To find a first benchmark for predictor performance, we simplified the problem by:

- Selecting a balanced dataset in order to have a same number of flare-productive and non-flare productive ARs.
- Not taking into account correlation over time of a same predictor measurements

We use Linear Support Vector Machine (LinSVM), a fast classifier for which only one hyper-parameter needs to be estimated. We used a protocol for performance assessment that shuffles the original dataset, splits it into a training set (for hyper-parameter estimation) and a test set (for comparison between predicted and observed value and computation of the confusion matrix), and repeats the operation n times. Estimation of hyperparameters was done using cross-validation by again divided the training set into several sub-sets.

From the confusion matrix, we computed the following:

- True Positive Rate (TPR), the ratio of true positives over total number of flaring ARs; the higher the ratio the better.
- False Positive Rate (FPR), the ratio of false positives over total number of non-flaring ARs; the smaller the ratio the better.
- True Skill Statistics (TSS), which measures the quality of the prediction. Its values range between [-1; 1], 1 being indicative of a perfect classifier, 0 being as good as guessing at random without any analysis. A negative value means the prediction outcome needs to be inversed to have a better guess. The TSS is not influenced by an imbalance in the number of observations in each class (flaring vs non flaring AR).

With $n=5$ runs, we obtained the following results: TPR lying between 60 to 75%, FPR lying between 8 to 20%, TSS lying between 49 to 59%. Those first results are encouraging, and better performance are expected if we consider classifier scheme which takes into account the redundancy in the time series of predictors as well as imbalanced between number of observations in each class.

Innovation: We used a rigorous protocol to evaluate the performance of a supervised classifier scheme in the flare prediction problem. Often when choosing a training and a test set in this context, distinctive years of observations are used (e.g. 2010-2014 for training and 2015 for testing). With such a practice the reshuffling is not entirely random, and the performance measure may include a time effect, reflecting the different phases of the solar cycle considered in the training and test sets.

Supervised classification of solar features using prior information

While the predictors considered here are computed using line of sight magnetogram, it may be useful in the future to add information coming from EUV imagers and featuring the Sun's corona rather than its photosphere. Towards this goal, we need to have an adequate segmentation of AR in the EUV that matches as best as possible the magnetic AR. As a first step, we presented in (De Visscher, et al., 2015) a supervised segmentation method based on a Bayesian classifier and the Maximum A Posteriori rule. The method allows integrating both manually segmented images as well as other type of information such as latitudinal distribution. It was applied on SDO-AIA images to segment them into ARs, coronal holes, and the remaining quiet sun part.

Added value brought by this project

This project enabled a first collaboration between the Royal Observatory of Belgium and the Machine Learning Group, ICTEAM Institute from the Université catholique de Louvain with Prof Dupont from UCL being a subcontractor in this project to deal with issues related to machine learning. As a first step, R. De Visscher and V. Delouille have taken in Fall 2013 the course 'Machine Learning: classification and evaluation' taught by Prof Dupont. Throughout the project Prof. Dupont gave advices and he co-authored one publication (De Visscher, et al., 2015). We expect this collaboration to continue during further development of the PREDISOL framework.

2.2 Analysis of spatial-temporal process (WP4)

Introduction

The flare productivity of an active region is observed to be related to its spatial complexity.
AR complexity is traditionally captured through a categorical classification of sunspots and magnetic AR, such as the Mount Wilson and McIntosh classification. The Mount Wilson classification has four main classes, two for describing 'simple' ARs, and two for more 'complex' ARs. Such categorical classification may not use all the information present in the data, and hinder a systematic study of an AR time evolution.

In WP4 we analysed the spatial complexity of ARs through a quantitative data-driven approach that use optimally the information present at small scales. Two modalities of ARs are jointly studied: SOHO/MDI continuum and SOHO/MDI magnetogram images. Such data-driven approach provided with a new way to look at ARs. We could confirm some commonly accepted fact but also discover new properties.

We proceeded in three parts: 1/ we analysed the spatial and modal correlation, 2/ we devised an unsupervised classification scheme based on Minimum Spanning Tree (MST) for ARs via dimension-reduction techniques, 3/ we provided a performance measure of MST for the task of separating the AR into two classes ('simple' and 'complex'), the reference label being given by the Mount Wilson classes.

Spatial and modal correlation analysis

Our methodology to analyse the local correlation structure relies on a generalization of wavelet analysis, called *dictionary analysis* (Mallat & Zhang., 1993). More specifically, we use a small sized dictionary to find a sparse representation of patches. In this context, a patch is a m x m-pixel neighbourhood, and a patch analysis of a n-pixel image will process the m^2 x n data matrix that collects the overlapping patches, see Figure 4.

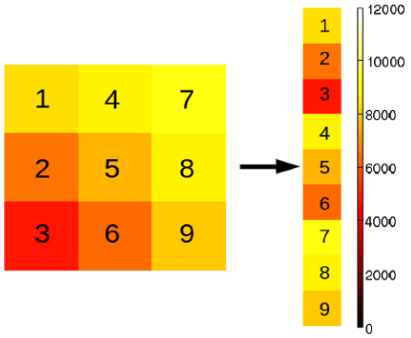

Figure 4 An example of a 3 x 3 pixel neighbourhood ('patch') extracted from the edge of a sunspot in a continuum image and its column representation.

In (Moon, et al., 2016b) we carried out a patch analysis of sunspot masks applied on continuum and magnetogram images. Those images were chosen so as to span the four main Mount Wilson classes, which distinguish 'simple' from more 'complex' AR. We proceeded as follows. First, we determined the number of parameters required to describe the spatial and modal dependencies. Such 'intrinsic dimension' was estimated using both linear and non-linear methods. The aim was to determine whether a linear method was sufficient for describing the correlation structure or not. We found out similar results for linear and non-linear intrinsic dimension estimates, suggesting that linear methods are appropriate.

Second, we analyzed the partial correlation of AR and sunspots, that is, the pixel-to-pixel correlation when the influence of all other pixels has been removed. Our analysis suggests that the images are stationary and follow approximately a third-order nearest neighbour Markov structure in the pixels. Figure 5 displays the partial correlation pattern within sunspots, indicating a stronger correlation in continuum than magnetogram images, and also in the vertical rather than horizontal direction. Our analysis suggests that it is not necessary to take patches larger than 5x5 in order to capture accurately the local spatial dependencies. A *5* x *5* patch on SOHO/MDI represents a square of about *90*Mm x *90*Mm on the Sun and is related to the size of the characteristic length of the largest penumbral filaments (Tiwari, et al., 2013) suggesting that on average the local spatial dependencies within sunspots are minimal beyond this scale.

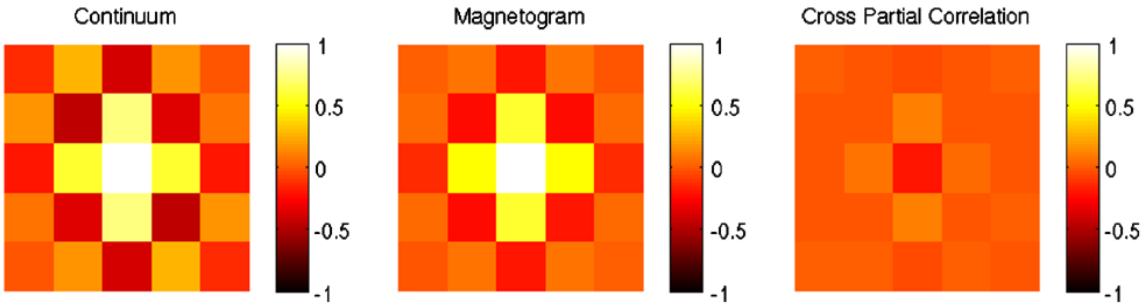

Figure 5 Partial correlation patches from sunspots. The partial correlation is stronger within the continuum.

Finally, we determined the degree of mutual correlation between the two modalities using canonical correlation analysis. Our results published in Moon et al (2016a) suggest that the two modalities are correlated. The correlation is however not perfect and so there may be an advantage to including both modalities in the classification of sunspots and flare prediction.

Clustering of active regions via dimension reduction techniques

The findings published in (Moon, et al., 2016a) can be summarized as follows.

- Linear methods are appropriate to analyse local correlation in AR.
- A 3x3 patch-analysis will capture most of the local correlation.
- Both modalities (continuum and magnetogram) should be used
- Via intrinsic dimension estimation we can choose a reasonable dictionary size,

This knowledge paved the way for building a new classification scheme of AR based on image-patch analysis, see (Moon, et al., 2016b). The first building block is to apply a dimension reduction operation to the AR image patch matrix: we considered the Singular Value Decomposition (SVD), and the Non-negative matrix factorization (NMF). The dimension reduction proceeds via a matrix factorization, and replace the image patch matrix by the product of a matrix of dictionary elements and a matrix of coefficients. Choosing a small number of dictionary elements will result in a dimension reduction. Figure 6 shows dictionary elements obtained from an SVD factorization.

This factorization can be looked at in two ways: we may focus on dissimilarities in the dictionary elements or in the matrix coefficients. In the first case, the matrix of dictionary elements is computed for each AR image, and a pairwise difference between these dictionaries is calculated using a Grassmannian projection metric (Stewart, 1973). In the second case, a single matrix of dictionary elements is computed from the combined collection of image patches from all AR image pairs. The matrix coefficients corresponding to a single AR are treated as samples from a distribution. The pairwise distances between these collections of coefficient samples is calculated using the Hellinger distance (Csiszar, 1967). In both cases, we input the dissimilarity measure into an unsupervised classification (also called clustering) scheme based on MST.

We measured the correspondence between the clusters obtained by MST and the Mount Wilson classification using the adjusted rank index (ARI). This ARI value is positive, indicating some overlap between the two completely different approaches. We also demonstrated that a larger ARI can be obtained by including specifically patches along the polarity inversion line.

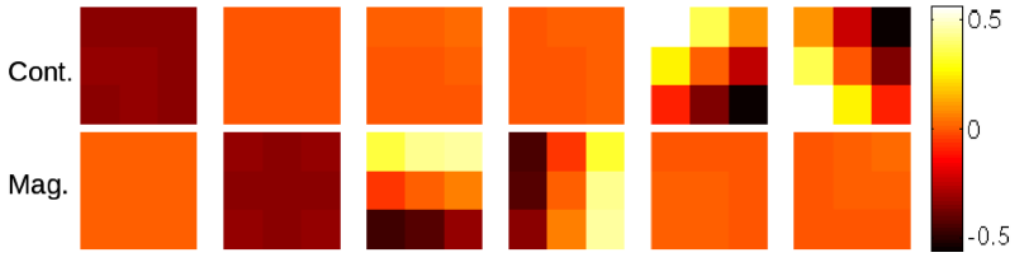

Figure 6 Learned dictionary elements using SVD. The patches consist of uniform patches and gradients in varied directions. Dictionary elements are constrained to be orthonormal, resulting here in magnetogram patches being close to zero when the continuum components are nonzero and vice-versa.

Bounds on performance of classification scheme

In (Moon, et al., 2015), we quantified the ability of the Grassmannian projection metric and Hellinger distance based algorithms to distinghuish simple ARs (i.e. ARs corresponding to the α and β Mount Wislon groups) from complex ARs (corresponding to the $\beta\gamma$ and $\beta\gamma\delta$ groups).

Considering the supervised problem of classifying a feature vector (here represented by the image patches) into one of two classes, it is possible to define the minimum classification error that can be achieved by any classifiers on the feature space. This quantity is called the *Bayes error rate (BER).* Unfortunately, computing the BER is difficult as it requires perfect knowledge of the probability distribution of the feature vectors, which is often unavailable. It is possible however to derive lower and upper bounds on the BER, using for example the Chernoff bound (Chernoff, 1952) or bounds using *f*-divergence functionals (Berisha, et al., 2016). In (Moon, et al., 2015) we used the Grassmannian projection metric previously defined and plug it into a *k*-nearest neighbour estimator to obtain a supervised classification algorithm. More precisely, we used an optimally weighted ensemble estimator using the *k*-nearest neighbour as our base estimator. We estimated upper and lower bounds on the BER using *f*-divergence functions derived in (Berisha, et al., 2016).

Our results indicate that if the goal is to accurately classify ARs into complex or simple ARs based on the Mount Wilson definition, then additional or different features are required. The image patch factorization may still be relevant for other learning tasks such as distinguishing between different local magnetic field structures.

Added value brought by this project

Through this collaboration with U. Michigan, the team at ROB learned about advanced technique for local correlation analysis, classification, and performance estimation of a

classifier scheme. Such advanced techniques had never been used on images of the Sun and brought a new light on these data. Novel ideas and techniques acquired from this collaboration can be extended to other types of images. On the other hand, the team at U. Michigan acquired new experience with the analysis of solar data, learning about their peculiarities and adapting the classification techniques in consequence.

3. DISSEMINATION AND VALORISATION

Publications: The scientific results were published in five peer review research papers: three in the Topical Issue on 'Statistical Challenges in Solar Information Processing' from the *Journal of Space Weather and Space Climate* and two as IEEE conference proceedings (see Section 5 on publications). The decision to have a topical issue on 'Statistical Challenges in Solar Information Processing' was taken at the end of the 7th Solar Information Processing workshop, for which V. Delouille was chair of the Scientific Organizing Committee.

Website: http://sdoatsidc.oma.be/web/sdoatsidc/SoftwarePREDISOL

Software: The resampling scheme proposed in (DeForest, 2004) and used in the PREDISOL framework was initially written in PDL. We have re-implemented this scheme using the Cython extension for Python for high performance. It is being considered for inclusion into AstroPy, a Python library for astrophysics that becomes more and more popular in the community.

Presentations at international conferences: The PREDISOL team gave eight presentations at international conferences: R. De Visscher presented a poster at the 2013 and 2014 editions of the European Space Weather week (ESWW) and two posters at the 7th Solar Information Processing workshop (7th SIPWork) in 2014. K. Moon presented a poster at two IEEE International Conference (one in 2014 and on in 2015) and gave a talk at the 7th SIPWork meeting in 2014. Finally, R. Attié gave a talk at the 2016 edition of the ESWW.

The full list of presentations (chronological order) is as follows:

1) "Predicting Flaring Activity through Supervised Classification on Predictor Variables" De Visscher, R.; Delouille, V.; Dupont, P. ESWW 10, Nov 18-23, 2013, Belgium (poster).

2) "Image patch analysis and clustering of sunspots: a dimensionality reduction approach " Moon, K.; Li, J.; Delouille, V.; Watson, F.; Hero III, A.O., 7th SIPWork, Aug 18-21, 2014, Belgium (oral)

3) "Supervised Classification of Solar Features Using Prior Information" R. De Visscher, V. Delouille, 7th SIPWork, Aug 18–21, 2014, Belgium (poster)

4) "Predicting Flaring Activity through Supervised Classification on Predictor Variables" R. De Visscher, V. Delouille, 7th SIPWork, Aug 18–21, 2014, Belgium (poster)

5) "Image patch analysis and clustering of sunspots: a dimensionality reduction approach" Moon, K.; Li, J.; Delouille, V.; Watson, F.; Hero III, A.O., IEEE International Conference on Image Processing, 28-30 October 2014, France (poster)

6) "Predicting Flaring Activity through Supervised Classification on Predictor Variables" R. De Visscher, V. Delouille, ESWW, 17-21 November, 2014, Belgium (poster)

7) "Meta learning of bounds on the Bayes classifier error" Moon, K.; Delouille, V.; Hero III, A.O. Proceeding of IEEE Signal Processing and SP Education workshop, Aug 9-12 2015, USA (poster).

8) "Characterization of active regions' time evolution in view of solar flare prediction" R. Attie, R. De Visscher, V. Delouille, ESWW 13, 14-18 November 2016, Belgium (talk).

4. PERSPECTIVES

V. Delouille and R. Attié, with advice from P. Dupont, will further develop the project along two axes: V. Delouille will work on tailoring the machine learning algorithms to the peculiarities of the problem whereas R. Attié will implement new predictors within the PREDISOL framework.

V. Delouille will for example account for imbalance in the number of flaring versus non-flaring active regions, by considering a modified version of the SVM algorithm, where the regularization of the loss function contains two penalty terms, that is, one for each class (flaring and non-flaring). Performance metrics may also be improved if we account for the correlation in time of measurements of a same predictor by applying a functional analysis that extracts the main characteristics of the predictor evolution before feeding it into the classifier scheme.

Magnetic ball tracking was developed in MATLAB and C++. The building of an interface in Python is currently under development by R. Attié. This will provide the possibility to add new predictors with interesting properties related to the physics of flares, which

should further improve the performance of flare forecasting using the PREDISOL framework.

5. PUBLICATIONS (PEER REVIEW)

1. K. Moon, J. Li, V. Delouille, F. Watson, A.O. Hero III
   Image patch analysis and clustering of sunspots: a dimensionality reduction approach, *Proceedings of IEEE International Conference on Image Processing*, Paris, France, 2014.

2. De Visscher R, Delouille V, Dupont P, and Deledalle C-A. Supervised classification of solar features using prior information. *Journal of Space Weather and Space Climate*,**5**, A34, 2015, DOI: 10.1051/swsc/2015033.

3. K. Moon, V. Delouille, A.O. Hero III. Meta learning of bounds on the Bayes classifier error, *Proceeding of IEEE Signal Processing and SP Education workshop*, Snowbird UT, 2015.

4. K. Moon, J. Li, V. Delouille, R. De Visscher, F. Watson, A.O. Hero III. Image patch analysis of sunspots and active regions. I. Intrinsic dimension and correlation analysis. *Journal of Space Weather and Space Climate*, 6, A2, 2016, DOI: 10.1051/swsc/2015044

5. K. Moon, V. Delouille, J. Li, R. De Visscher, F. Watson, A.O. Hero III. Image patch analysis of sunspots and active regions. II. Clustering via dictionary learning *Journal of Space Weather and Space Climate*, 6, A3, 2016, DOI: 10.1051/swsc/2015043

6. ACKNOWLEDGEMENTS

V. Delouille acknowledges support from the Belgian Federal Science Policy Office through the ESA-PRODEX program, grant No. 4000103240.

Professor Alfred Hero and research student Kevin Moon gratefully acknowledge funding from US National Science Foundation grant CCF-1217880 that supported Alfred Hero and the US National Science Foundation Graduate Fellowship Program that supported Kevin Moon during his PhD under grant number F031543.

6. REFERENCES

Ahmed, O. et al., 2013. Solar flare prediction using advanced feature extraction, machine learning, and feature selection. *Sol Phys,* Volume 2013, pp. 157-175.

Attié, R. & Innes, D., 2015. Magnetic balltracking: tracking the photospheric magnetic flux. *Astronomy & Astrophysics,* p. id A106.

Berisha, V., Wisler, A., Hero, A. & Spanias, A., 2016. Empirically Estimable Classification Bounds Based on a Nonparametric Divergence Measure. *IEEE Transactions on Signal Processing,* 64(3), pp. 580-591.

Chernoff, H., 1952. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics,* p. 493–507.

Csiszar, I., 1967. Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.,* Volume 2, pp. 299-318.

De Visscher, R., Delouille, V., Dupont, P. & Deledalle, C.-A., 2015. Supervised classification of solar features using prior information. *Journal of Space Weather and Space Climate,* Volume 5, p. A34.

DeForest, C., 2004. On Re-sampling of Solar Images. *Solar Physics,* pp. 3-23.

Elad, M. & Aharon., M., 2006. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Trans. Image Process.,* 15(12), p. 3736–3745.

Mallat, S. & Zhang., Z., 1993. Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Process.,* 41(12), p. 3397–3415.

Moon, K., Delouille, V. & Hero III, A., 2015. Meta learning of bounds on the Bayes classifier error. *Proceedings of the IEEE Signal Processing and SP Education workshop.*

Moon, K. et al., 2016b. Image patch analysis of sunspots and active regions. II. Clustering via dictionary learning. *Journal of Space Weather and Space Climate,* Volume 6, p. A3.

Moon, K. et al., 2016a. Image patch analysis of sunspots and active regions. I. Intrinsic dimension and correlation analysis. *Journal of Space Weather and Space Climate,* Volume 6, p. A2.

Moon, K. et al., 2014. Image patch analysis and clustering of sunspots: a dimensionality reduction approach. *Proceedings of the IEEE International Conference on Image Processing.*

Stewart, G., 1973. Error and perturbation bounds for subspaces associated with certain eigenvalue problems. *SIAM Review,* 15(4), p. 727–764.

Tiwari, S., Noort, M. v., Lagg, A. & Solanki., S., 2013. Structure of sunspot penumbral filaments: a remarkable uniformity of properties. *Astronomy & Astrophysics,* Volume 557, p. A25.