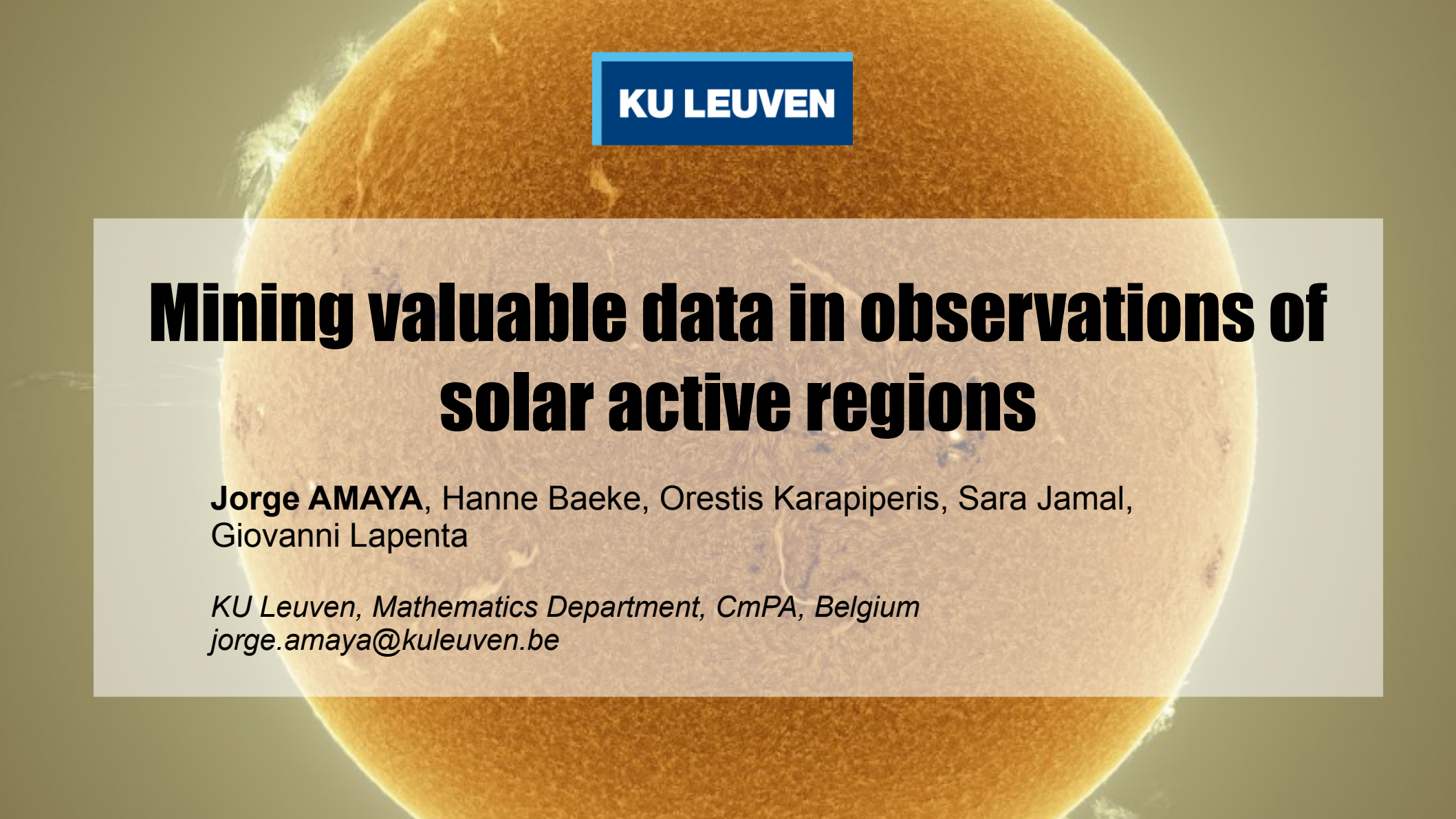


KU LEUVEN

Mining valuable data in observations of solar active regions

Jorge AMAYA, Hanne Baeke, Orestis Karapiperis, Sara Jamal, Giovanni Lapenta

KU Leuven, Mathematics Department, CmPA, Belgium
jorge.amaya@kuleuven.be

OBJECTIVE

Better identification of the properties associated to flaring activity in active regions to perform better flare forecasts

WORK ALREADY DONE

- Multiple papers based on the SHARP magnetic field features The most famous work in Europe is by the Flarecast project [Florios et al., 2018, Massone et al. 2018,...].

- Some papers use machine learning [Benvenuto et al., 2017, Liu et al. 2017, Nishizuka et al. 2017, Chen et al. 2019,...].

  - Others use White light classifications [McCloskey et al., 2018,...].

- A few authors have included image analysis in their forecasting [Jonas et al. 2018,...]

OUR CONTRIBUTION

Project this N-dimensional representation of the ARs into a latent space that is easier to work with: separable, sparse, non-convex.

We do this already in other domains: project the data into an orthogonal base where all the points are linearly dependent on a few number of components.

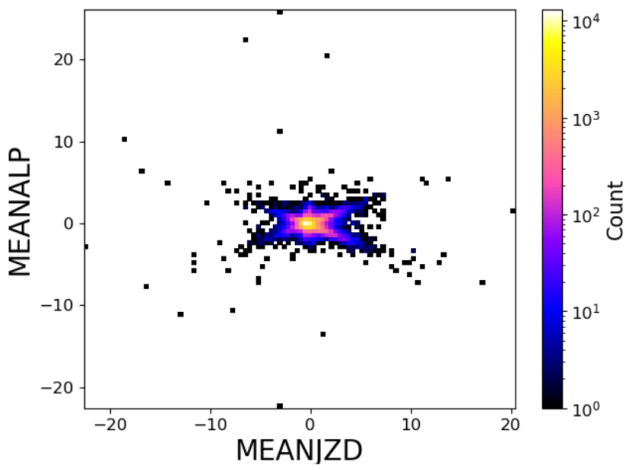
THE DATA

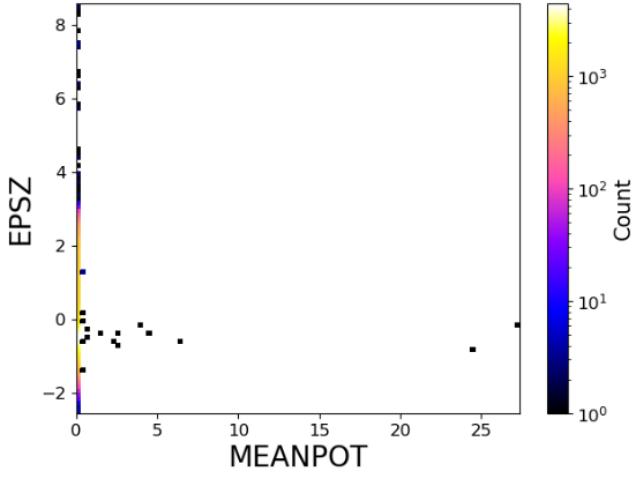
- Multivariate Time Series MVTS extracted from SHARP regions [Angryk et al. 2020]
- 24 parameters from May 2010 – December 2018 (24th solar cycle)

	Complete Data	Data without NaN	Sampled Data
nb data points	2 906 658	2 609 953	57 444
nb No flares		2 602 509	50 000
nb C-flares		6717	6717
nb M-flares		680	680
nb X-flares		47	47

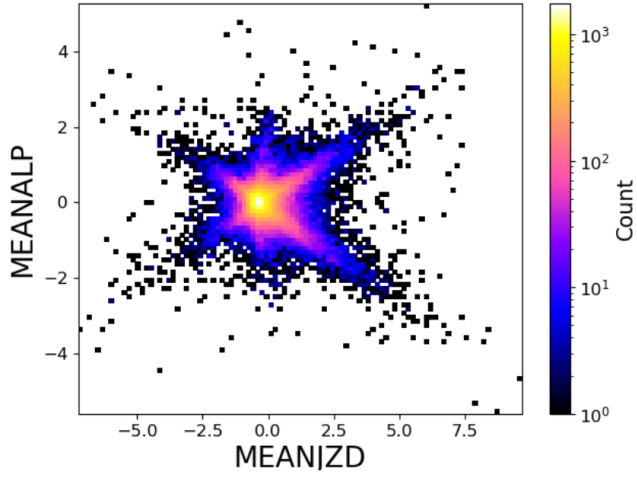
OUTLIER REMOVAL

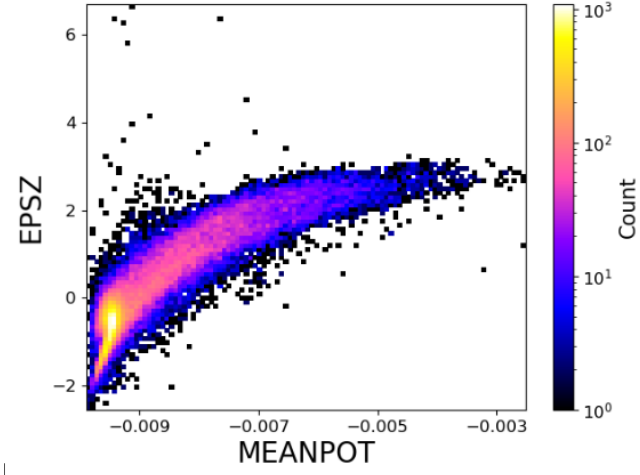
BEFORE





AFTER





CHECKING OUTLIERS

	Standardized data	Outliers HDBSCAN	Outliers MEANPOT	% outliers in data
nb data points	57 444	586	33	1.08 %
nb No-flares	50 000	214	30	0.49 %
nb C-flares	6717	282	2	4.23 %
nb M-flares	680	80	1	11.91 %
nb X-flares	47	10	0	21.28 %

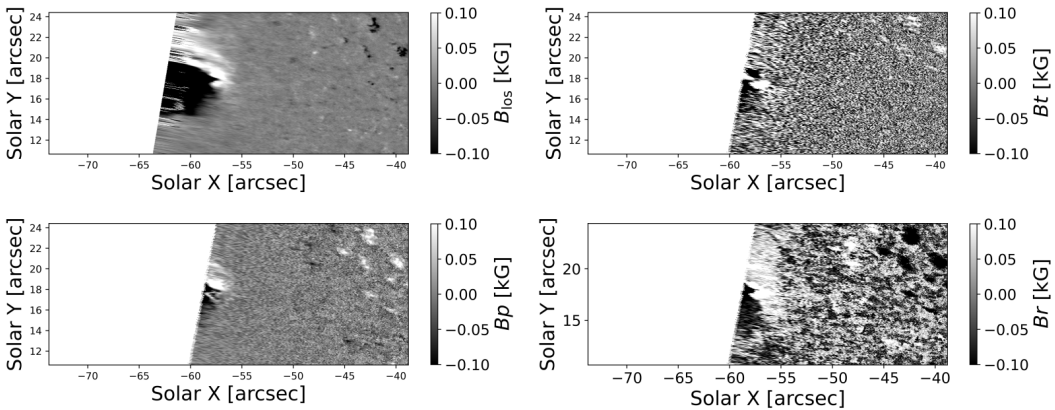
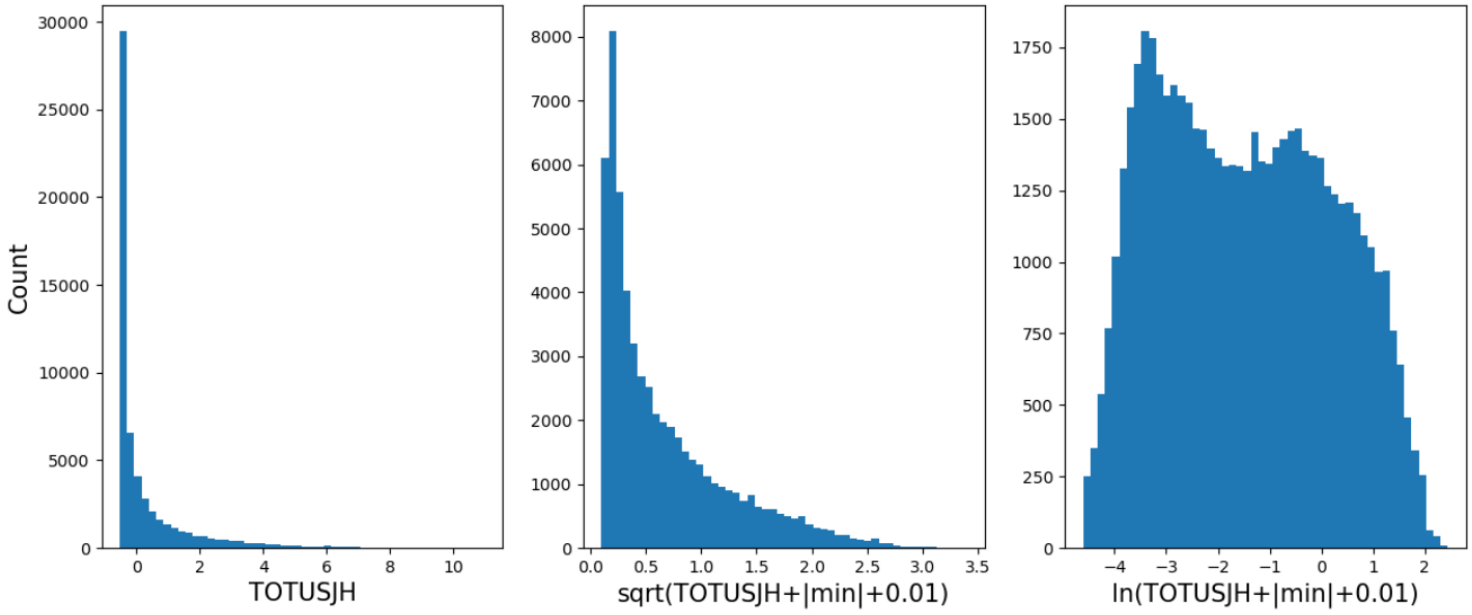
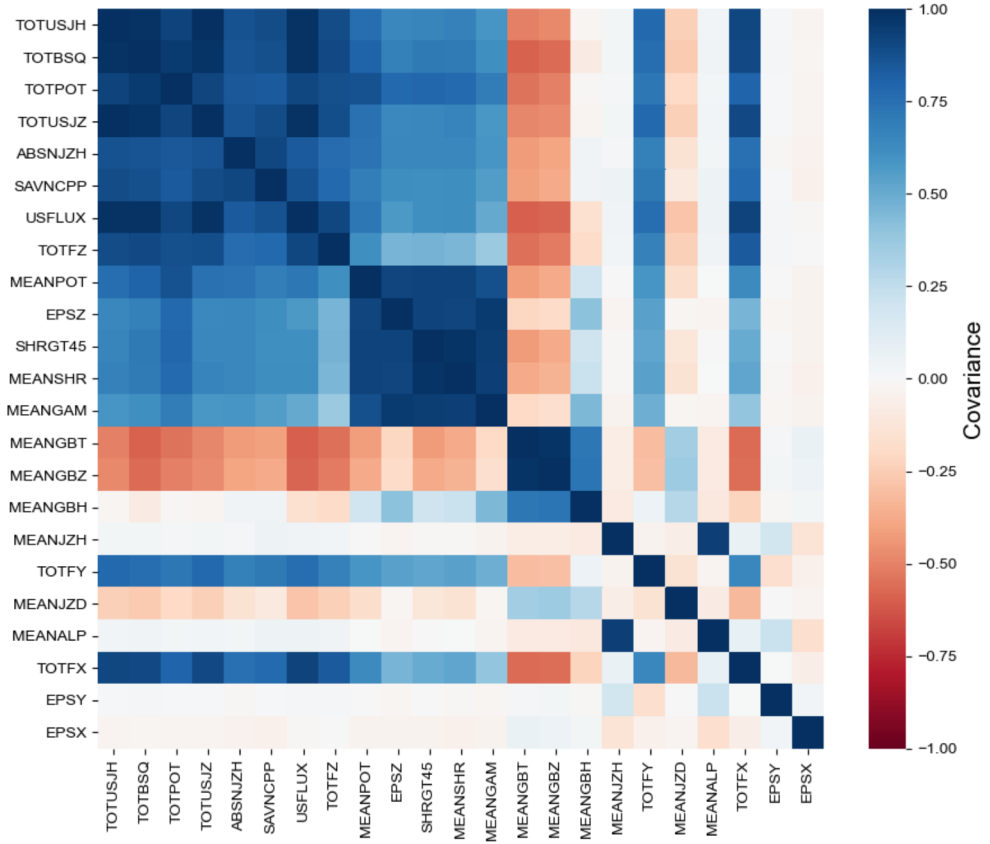


Figure 2.9: Magnetic field components for SHARP with HARPNUM 1449, measured at 2012-03-02, 17:36:00 International Atomic Time (TAI).

DATA DISTRIBUTIONS

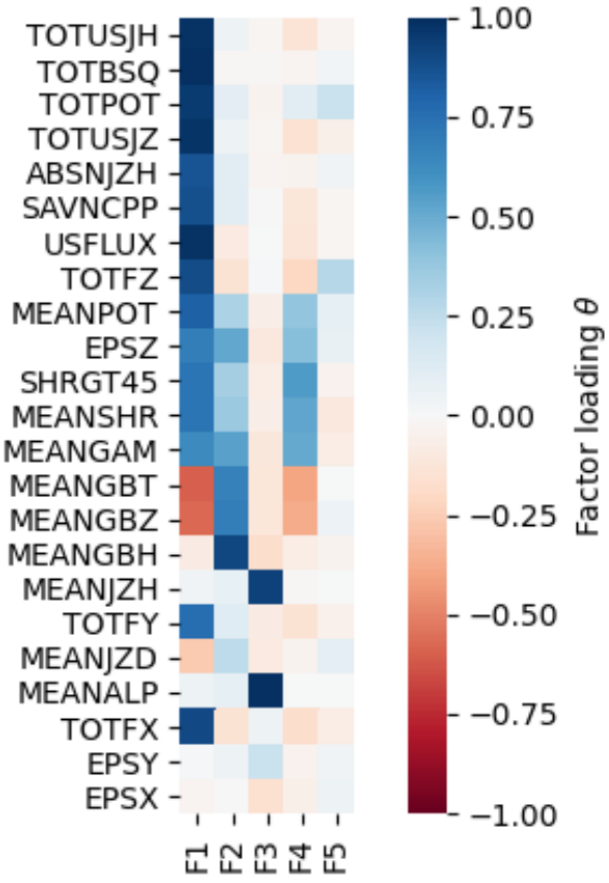


UNNECESSARY REPETITION OF DATA

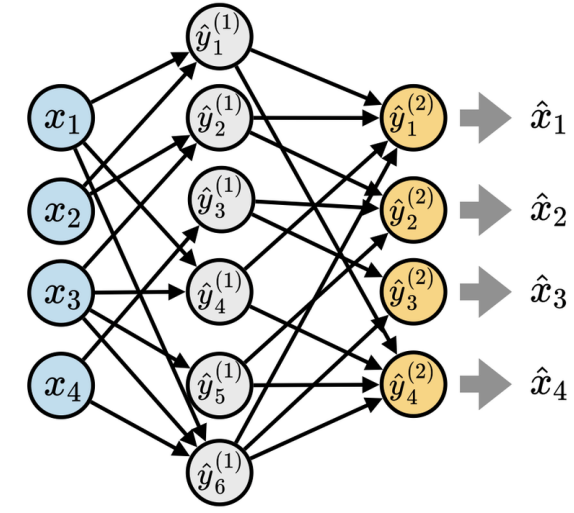


COMMON FACTOR ANALYSIS

$$Y_1 = \theta_{10} + \theta_{11}F_1 + \theta_{12}F_2 + e_1$$
$$Y_2 = \theta_{20} + \theta_{21}F_1 + \theta_{22}F_2 + e_2$$
$$Y_3 = \theta_{30} + \theta_{31}F_1 + \theta_{32}F_2 + e_3$$
$$\dots$$
$$\dots$$
$$Y_{24} = \dots$$

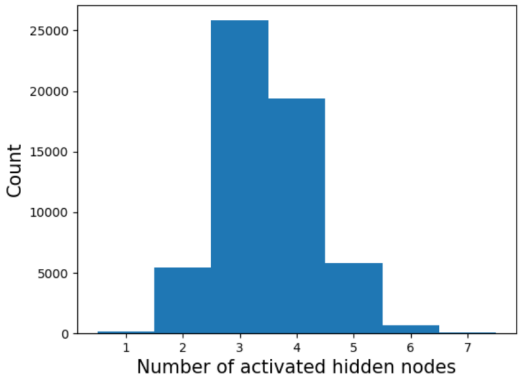


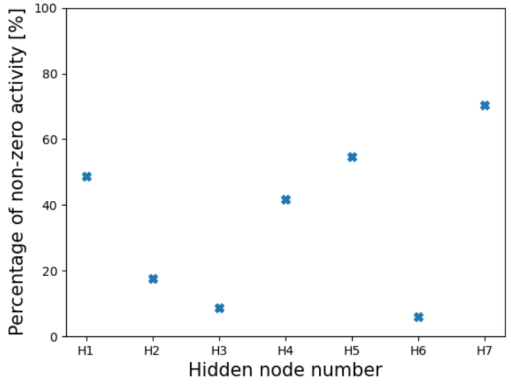
INTRODUCE SPARSITY: SPARSE AUTOENCODERS

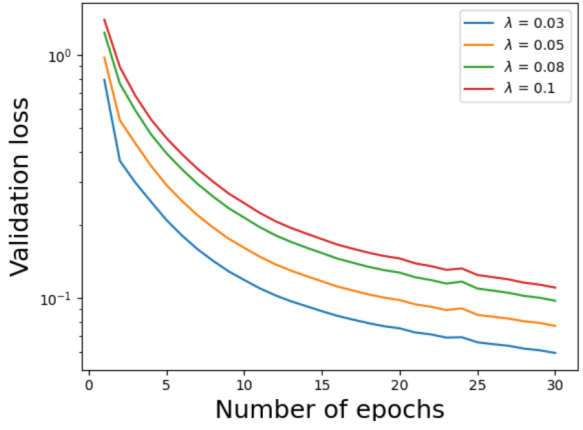


Sparse features

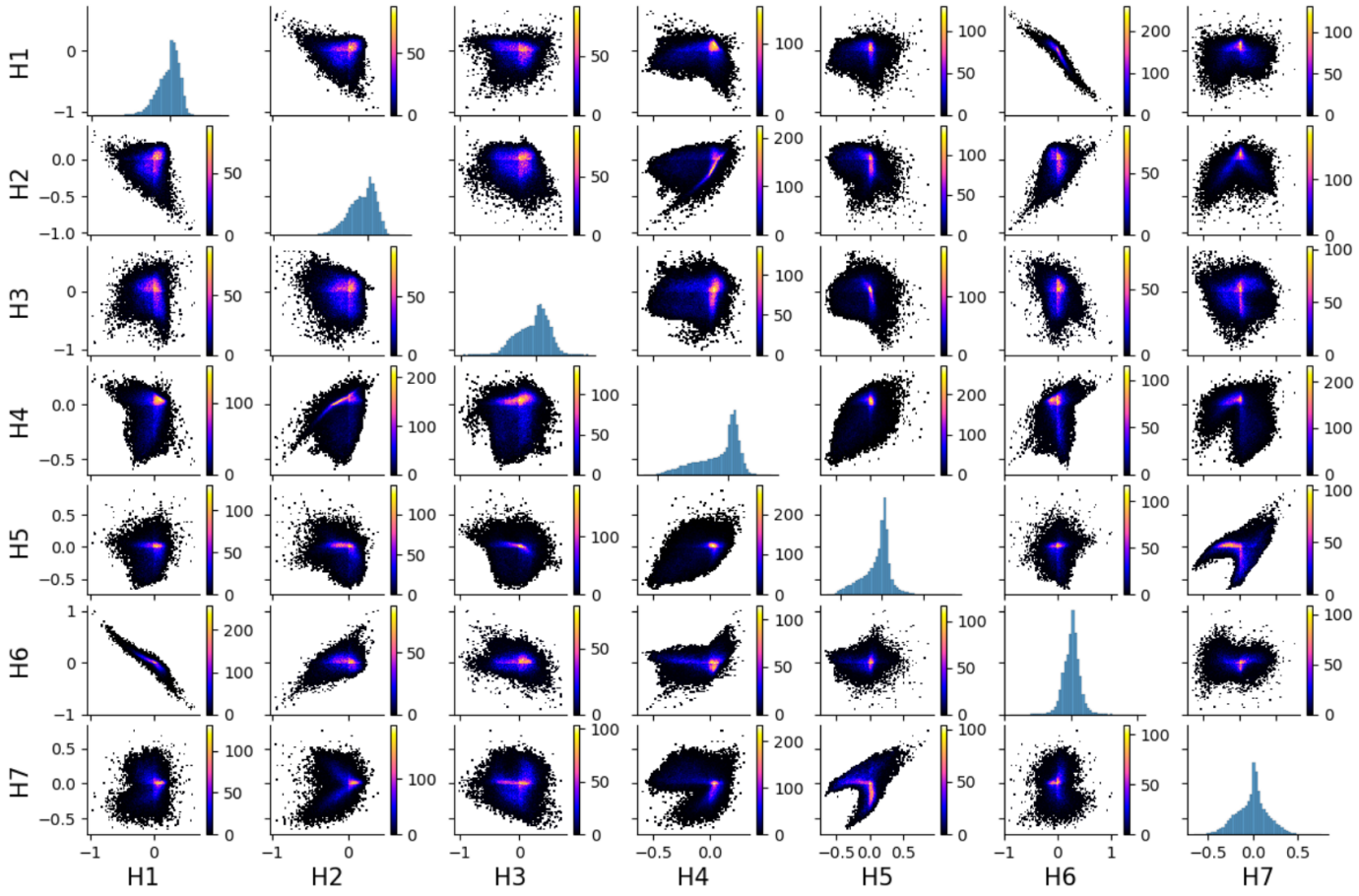
$$L1 = \lambda \sum_i \left| a_i^{(h)} \right|$$



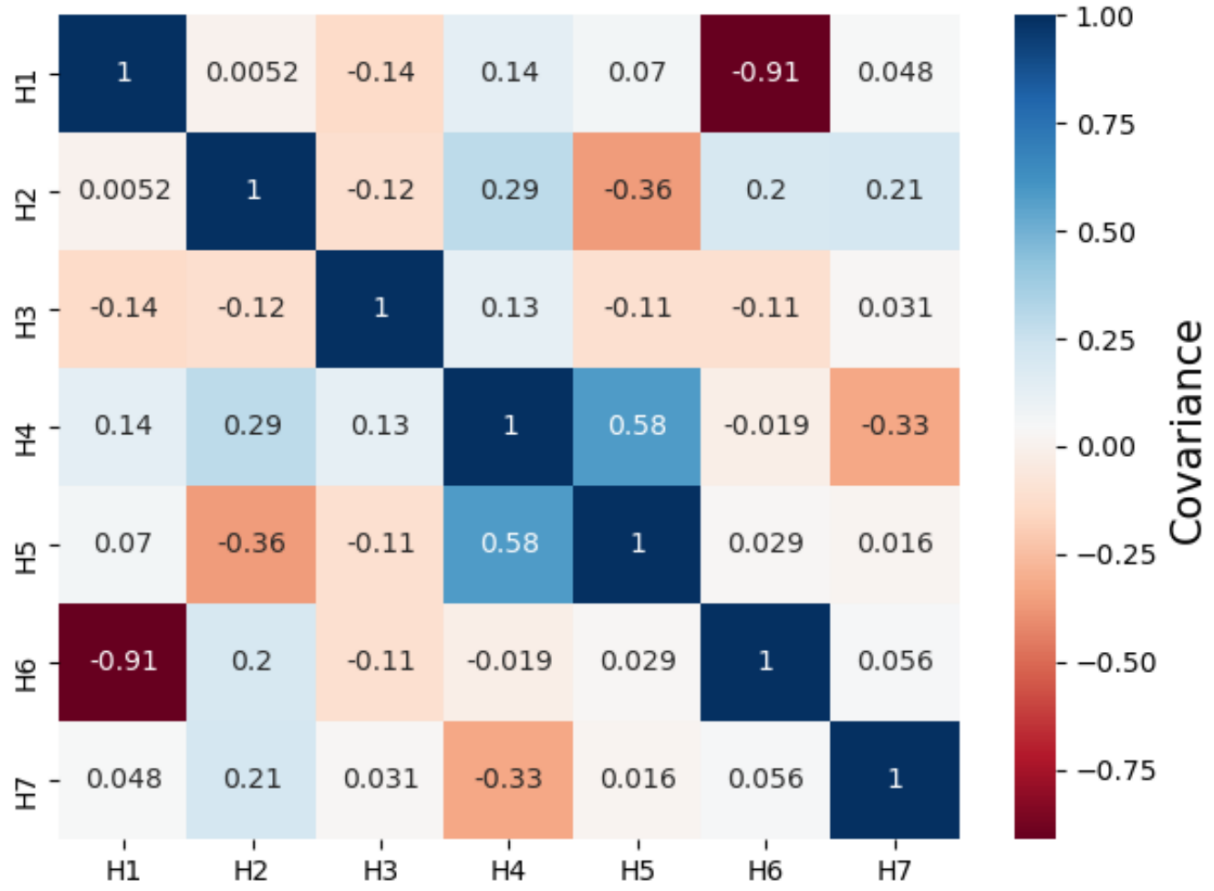




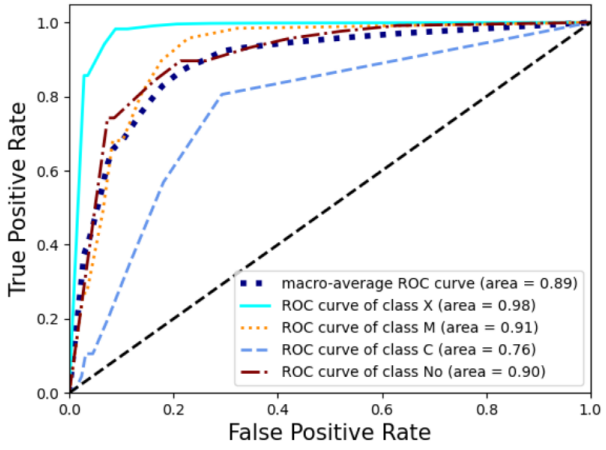
TRANSFORMED DATA DISTRIBUTIONS

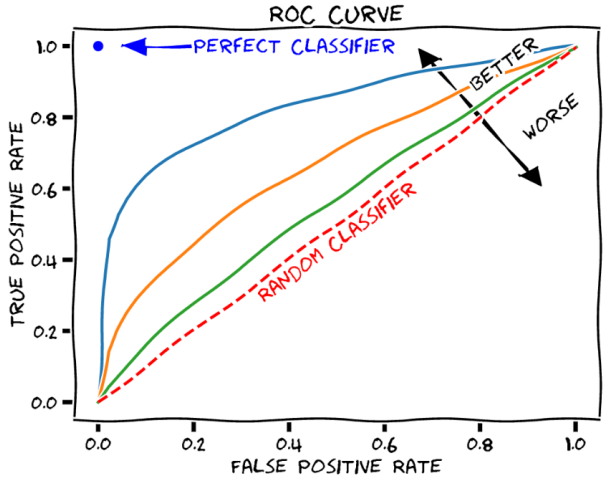


SPARSE FEATURES COVARIANCE MATRIX



SUPERVISED KNN CLUSTERING





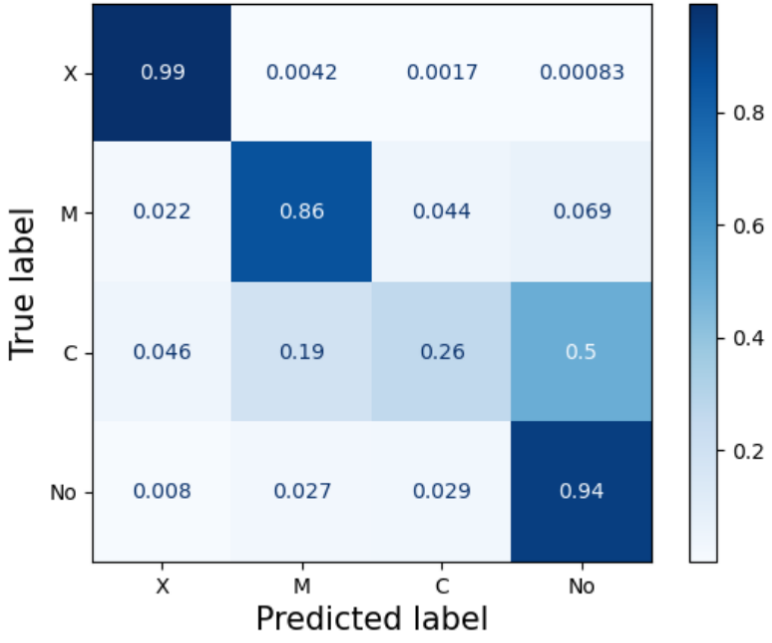


Figure 4.4: Normalized confusion matrix for KNN, for the over-sampled data set with SMOTE to 6000 M- and X-flares.

UNSUPERVISED K-MEANS

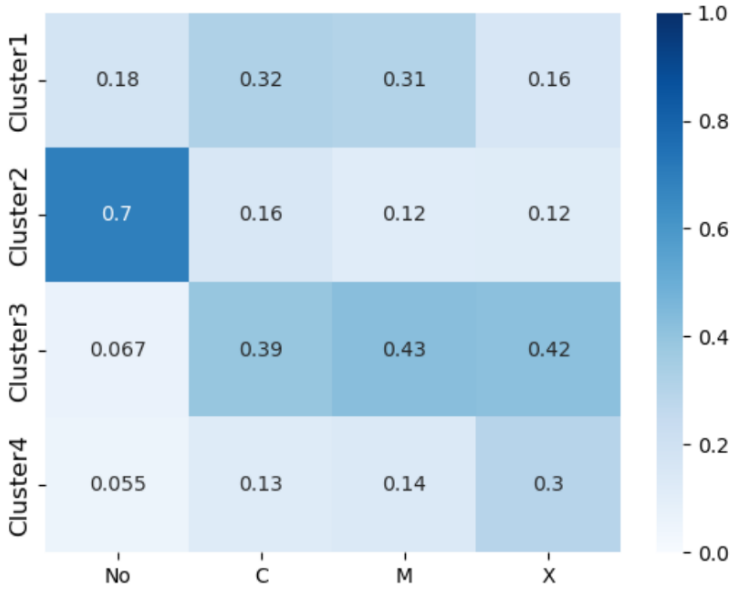
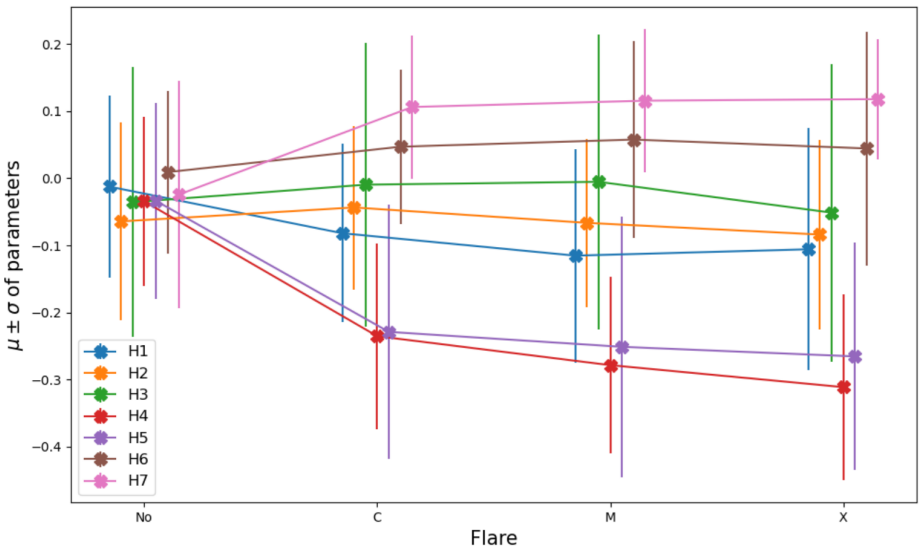


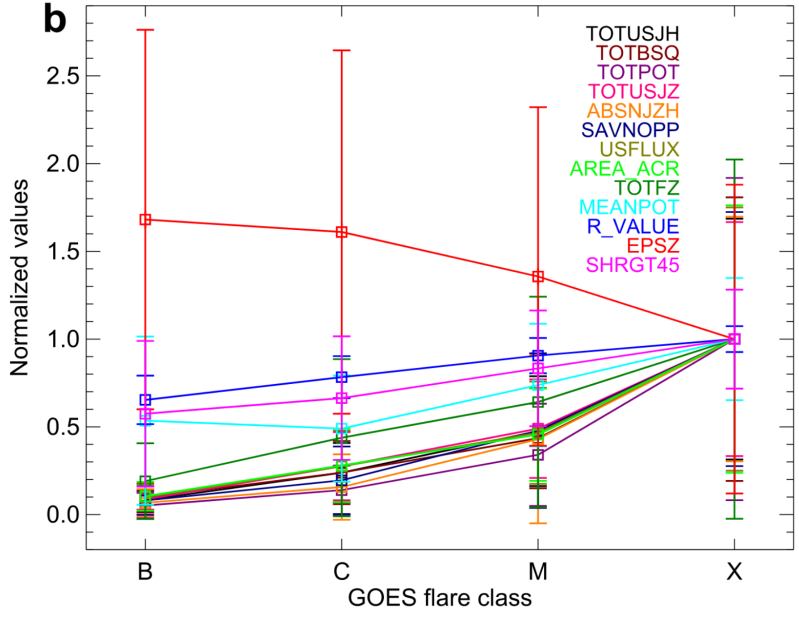
Figure 4.18: Clustering results of K-means, for the randomly under-sampled and SMOTE over-sampled data set to 6000 samples in each class. The percentage of each flare included in the four clusters is shown. Normalization is performed per flare type.

SPARSE FEATURES PER FLARE TYPE

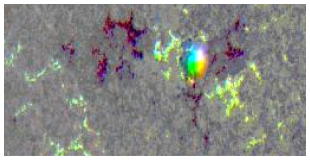
Our work:

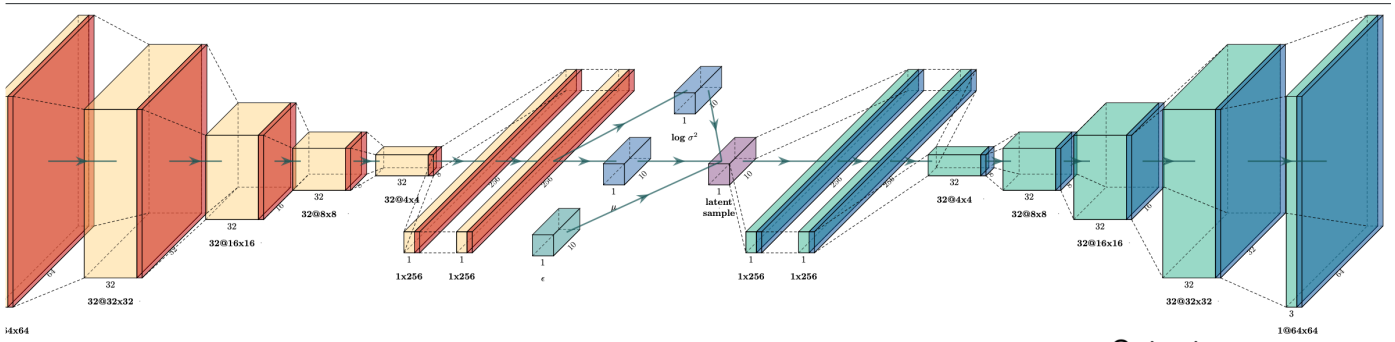


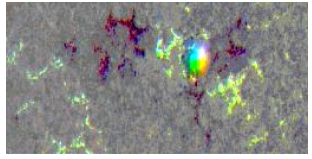
[Liu et al., 2017]

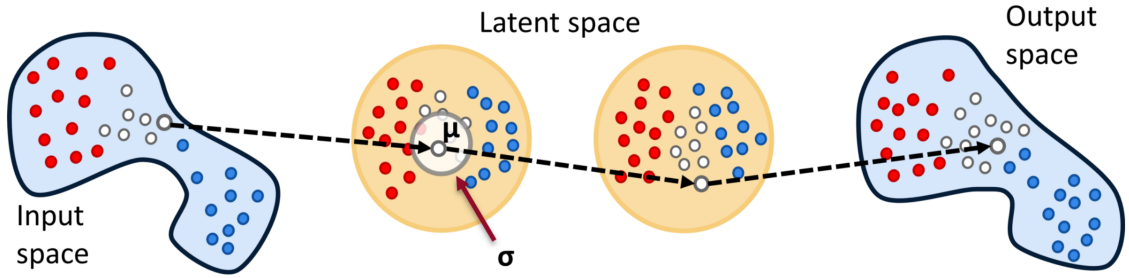


Variational Autoencoders





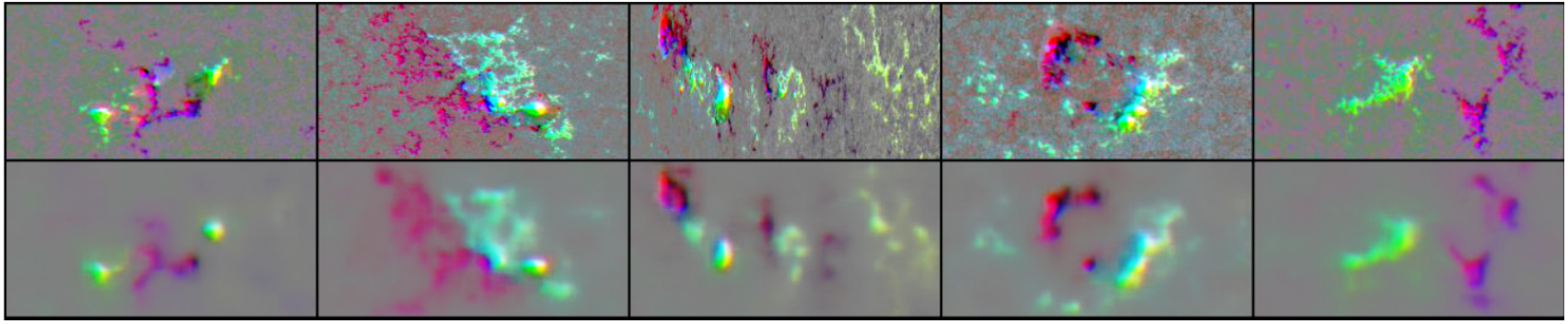




DEMONSTRATION OF RECONSTRUCTIONS

64 hidden units. Input/output image: 128 x 256 x 3

Original AR



VAE reconstruction

GENERATION OF ARTIFICIAL AR