

OMNI2 low-resolution data set

What is the intrinsic dimensionality of the OMNI data? A dimensionality reduction study

Jannis Teunissen
CWI, Amsterdam

Collaborators: R. Dupuis, C. Shneider, E. Camporeale



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776262 (AIDA, www.aida-space.eu)



- 52 variables
 - IMF data
 - Solar wind parameters
 - Geomagnetic and solar activity indices
 - Energetic proton fluxes
- Hourly averages 1975 – 2019
- Non-physical variables removed (e.g. spacecraft ID)

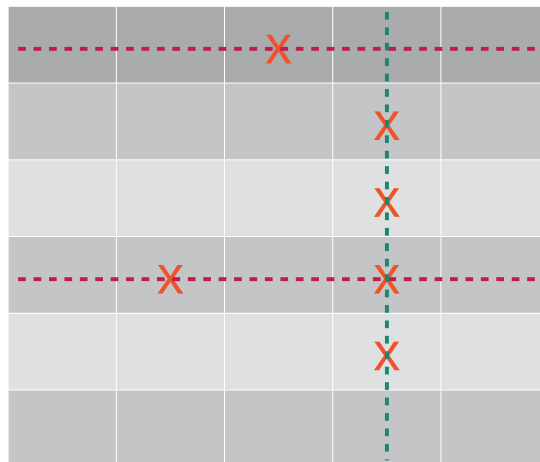
[Credits: GSFC/SPDF and OMNIWeb]

https://spdf.gsfc.nasa.gov/pub/data/omni/low_res_omni/

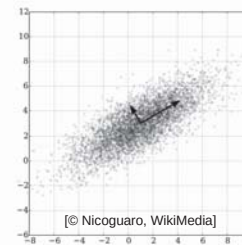
Data preprocessing

Step 1: Remove columns with too many missing values

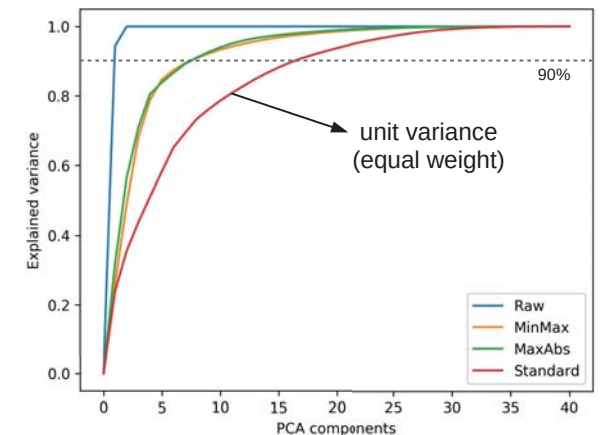
Step 2: Remove rows with any missing value



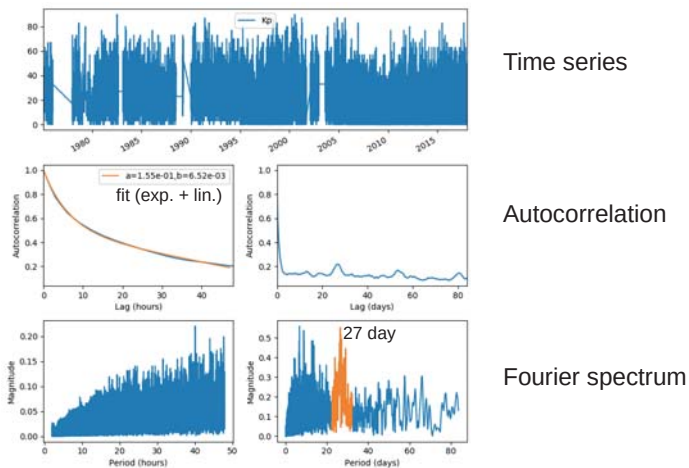
Principal Component Analysis (PCA)



- Linear transformation
- Components in direction of most variance
- Normalization is important!
- Difficult to interpret here

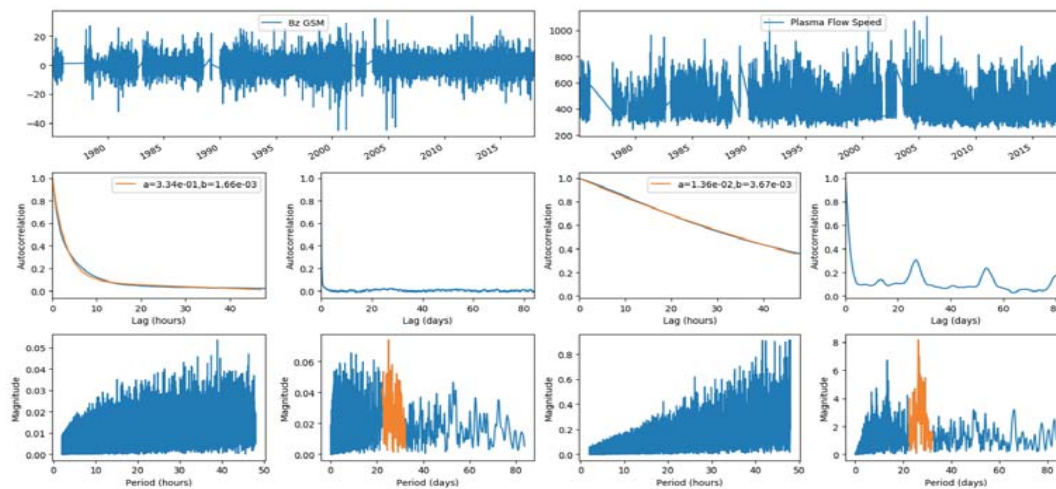


Exploring data – Kp example

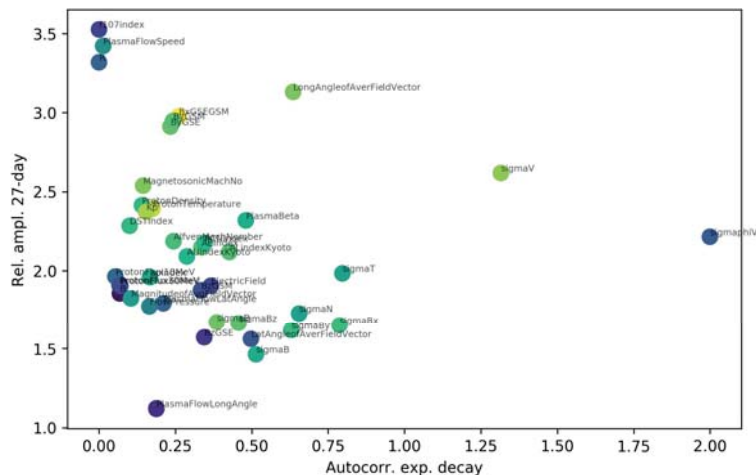


Bz GSM

Plasma Flow Speed



Distribution of variables



Ranking variables by predictive power

- Linear model ($y = Ax + b$)
 - Feature: x_i
 - Target: $\mathbf{X} = (x_1, x_2, \dots, x_N)$
- Measure R^2 score (components or average)

$$R^2 = 1 - \frac{\langle (f(x) - y)^2 \rangle}{\text{Var}(y)}$$

\bar{R}^2	Top 5
0.188	Kp
0.180	AE Index
0.167	ap index
0.162	AL index (Kyoto)
0.161	sigma B
0.160	PC(N) index

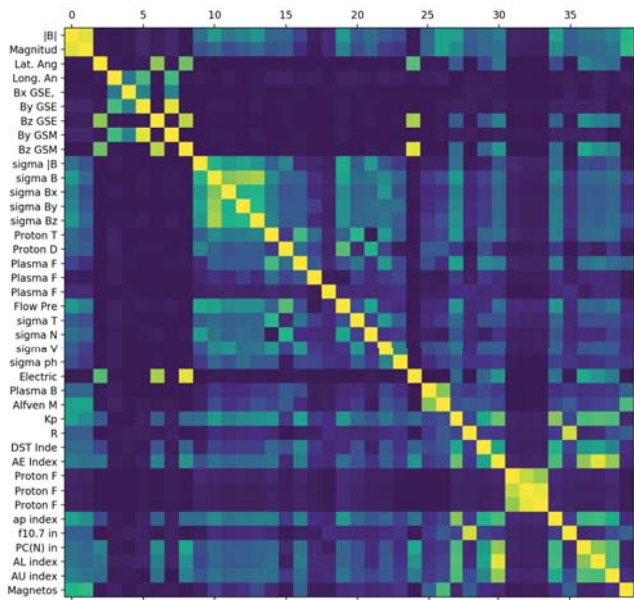
Which variables do they predict?

\bar{R}^2	Bottom 5
0.0608	Plasma Beta
0.0539	sigma phi V
0.0525	Long. Angle of Aver. Field
0.0387	Bx GSE, GSM
0.0361	Plasma Flow Long. Angle
0.0290	Plasma Flow Lat. Angle

Correlation matrix

For a linear model

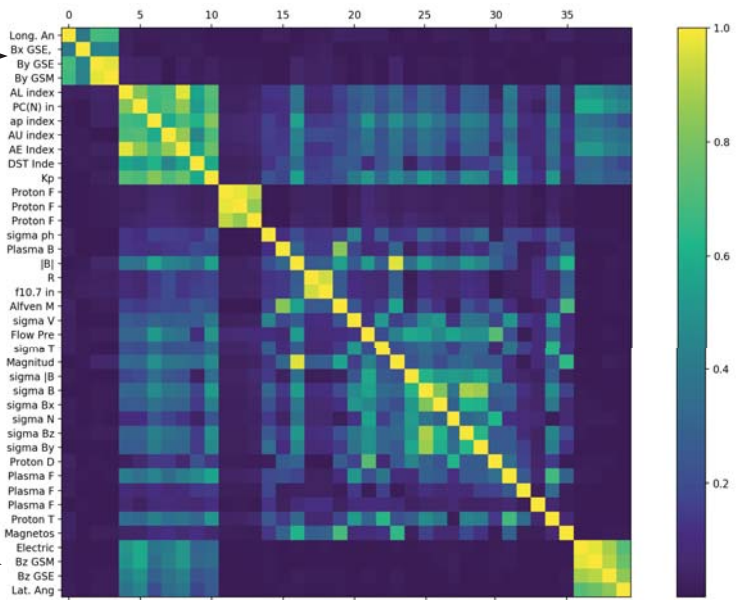
$$\text{corr}(x, y) = \sqrt{R^2}$$



Bx, By, long. angle →
 Al, PC, ap, AU, →
 AE, Dst, Kp →
 Proton fluxes →

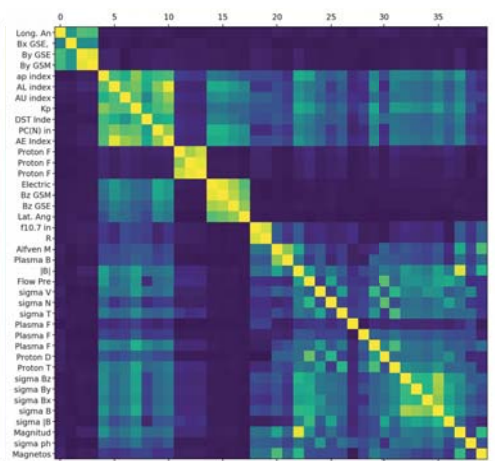
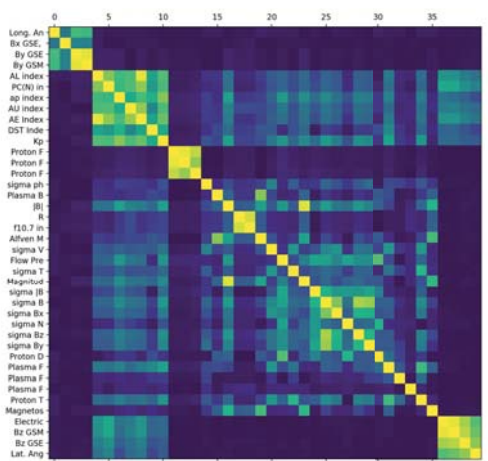
Block-diagonal reordering

E, Bz, lat. angle →



Without history

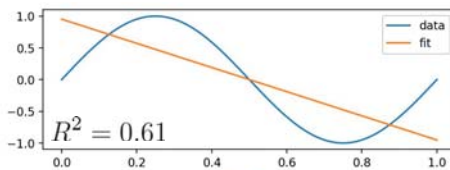
24h history



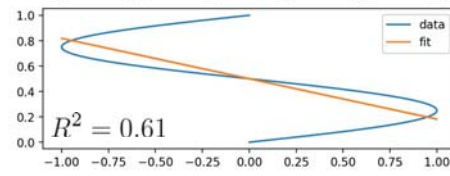
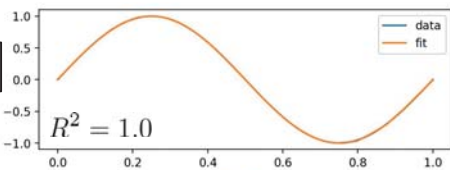
Linear vs non-linear models

Linear model

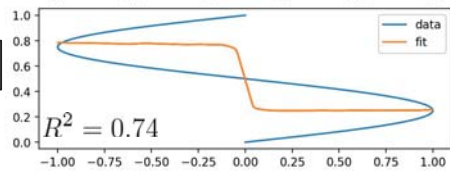
Neural network



$x \rightarrow \sin(x)$

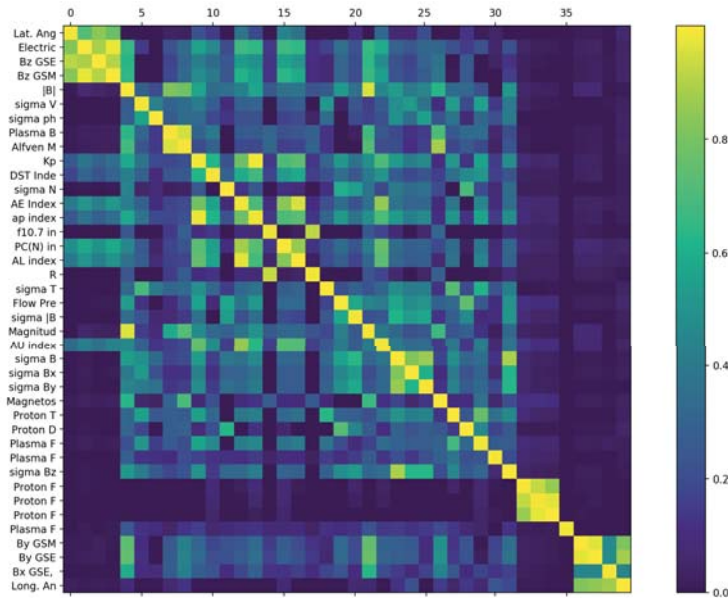


$\sin(x) \rightarrow x$



Neural network
24h history
Shown is $\sqrt{R^2}$

asymmetry linked
to causality?



Summary

- PCA: ~17 variables for 90% of variance (with proper normalization)
- Autocorrelation times and strength of 27-day periodicities are spread out; no obvious clusters
- Correlation matrices and their non-linear generalization help to discover relations between variables



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776262 (AIDA, www.aida-space.eu)

