FILLING THE GAPS

Neural Nets for Space Stats

THE PROBLEM

Telemetry, instrumentation and calibration problems onboard spacecraft mean the data they send back to Earth often has large gaps in it. This has severe implications for many different fields within space science, including heliophysics. This study was motivated by the large amount of missing data in Voyager datasets, from which we need good spectral statistics to understand the evolution of turbulence in the solar wind. Gaps in the data mean that such statistics become noisy and unhelpful for extracting physical information. Unlike previous attempts[1][2] to rectify this, which have focussed on techniques such as linear interpolation and maximum likelihood, this study took a machine learning approach.

THE STATS

A time series of the magnetic field strength of the solar wind often appears chaotic and unpredictable. In order to make sense of this data, we calculate statistics that summarise how the field is changing. In particular, we can examine the scales on which the field fluctuates. Here we focus on the **structure function**, a function of distance between two points that is central to turbulence models. These, in turn, are critical to improve our space weather models.

MAGNETIC FIELD OVER TIME

$\langle |\boldsymbol{B}(t+\tau) - \boldsymbol{B}(t)|^n \rangle$

Mathematical equation

STRUCTURE FUNCTION

Neural network?

DATA

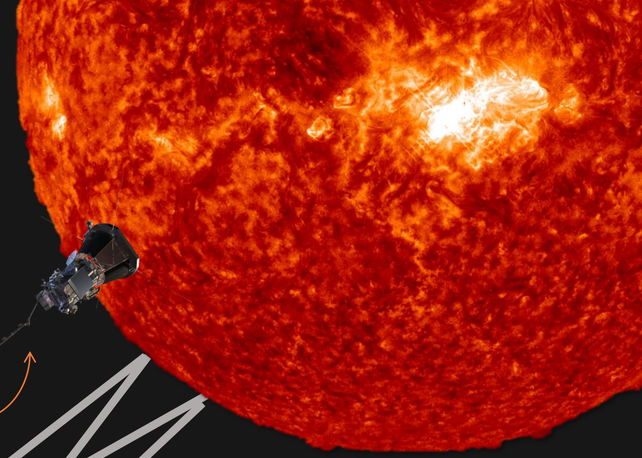
Due to its high frequency measurements and high degree of completeness, an interval from Parker Solar Probe was chosen to train and evaluate a simple feed-forward neural network. A 17-day interval of magnetic field strength from the Fluxgate Magnetometer recorded during November 2018 of Encounter 1 of the spacecraft was chosen. This interval was then split into 195 intervals, each of dimension 3 x 10,000. The second-order vector structure function (1 x 2000) was then calculated for each.

METHOD

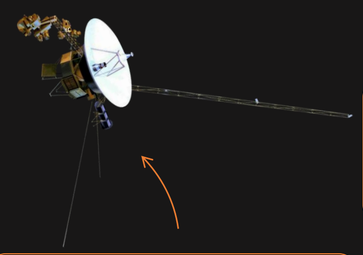
80% of intervals were used for training and 20% for testing. Both sets were duplicated to increase the available examples, after which random gaps were removed to simulate real-word dirty series. The training (and validation) data was then used to find a locally-optimal ANN architecture, which was difficult in that it required not only minimising the loss function, but also visual inspection of a sample of predictions, to ensure overfitting was not occurring.

[1] Gallana *et al.* (2015), *J. Geophys Res.*
[2] Fraternale *et al.* (2019), *ApJ*

Research and poster by **Daniel Wrench**, under the supervision of Dr. Tulasi Parashar and Dr. Ramesh Rayudu
School of Chemical and Physical Sciences, Faculty of Science, Victoria University of Wellington

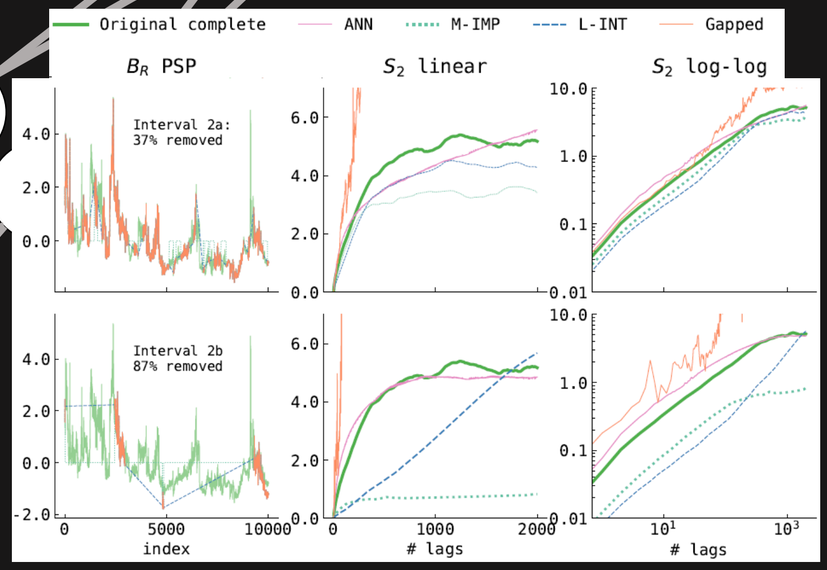


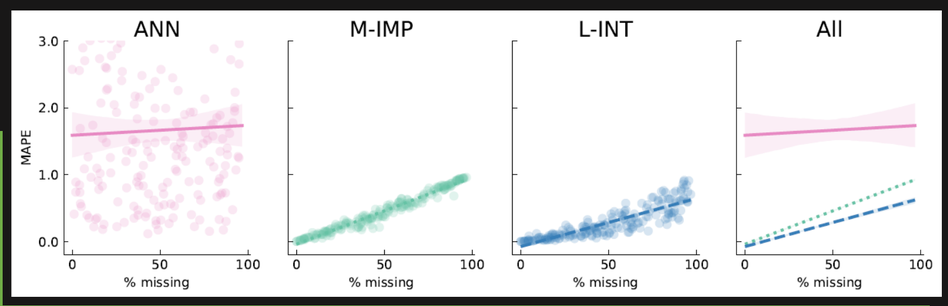
The Sun.
Our star is the ultimate source of data for this project, producing a stream of charged particles every second that hurtle past the Earth in a complex flow.



Voyager.
Currently soaring through interstellar space, the twin Voyager spacecraft are the farthest man-made objects from the Sun. This distance means the data sent back has large gaps in it, hence providing the motivation for this research.

Parker Solar Probe.
This project used magnetic field data from PSP to train its model, which represent the closest measurements of the Sun ever recorded.





RESULTS

Through iteration and inspection the best configuration found was a network with 2 hidden layers, each with 10 nodes. The first column in the top figure shows one original interval gapped in two different ways, producing the red lines. The second and third columns show the predictions made by the ANN of the true structure function for the corresponding clean interval, alongside the structure functions computed normally from the gapped interval, the mean-imputed interval, and the linearly interpolated interval. At low gap %, while the function calculated from the gapped interval is very dirty, these much simpler methods of interpolation still produce a curve which closely matched the shape of the target curve. However, the second row shows these methods diverging much further from the true curve, while the ANN (purple line) produces a prediction that is far closer to the target.The statistical performance of these methods is summarised in the second figure, which shows the mean absolute percentage error, a measure of relative error, of each prediction for each method. The ANN predictions show much greater variability in how close they are to the true curve, whereas the other methods show a much tighter linear relationship with increasing % missing data. Ultimately, the regression lines show that linear interpolation the most reliable approximator, despite edge cases at high % missing where ANNs perform better.

CONCLUSION

This study has demonstrated the utility of simple interpolation methods for dealing with measurements of the solar wind that are plagued with gaps. It has also shown the challenge in optimising and evaluating neural network architectures when trained to fit curves, and approximate their shape, as opposed to simpler scalar regression or classification tasks. Further investigation into optimisation and network architectures suited to sequential data are the next steps, with the ultimate goal of being able to predict a range of turbulence statistics in a variety of heliophysical regimes, thereby greatly increasing our understanding of space weather in the face of incomplete data.