Understanding geomagnetic activity through supervised learning

RUGGERO VASILE[1], IRINA ZHELAVSKAYA[1,2], YURI SHPRITS[1,2,3]

ruggero.vasile@gfz-potsdam.de

[1] GFZ German Research Centre For Geosciences, Potsdam, Germany

[2]University of Potsdam, Potsdam, Germany [3]University of California, Los Angeles, California, USA

1. Introduction

Forecasting geomagnetic activity is crucial for a variety of aspects: to safeguard operations in the near-Earth space, e.g. to protect satellites and space travelers from space radiation, and on the ground, e.g. to ensure correct functioning of power grids.
The Kp geomagnetic index measures the global levels of geomagnetic activity driven by solar particle radiation. It ranges from zero, i.e. variability below noise levels, up to nine, i.e. extreme geomagnetic storm.
It is widely used, for instance, to parametrize atmospheric, thermospheric and magnetospheric models.

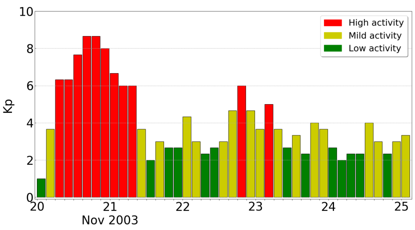

Figure 1: Kp-index time evolution. The standard colors encode the geomagnetic activity levels.

2. Forecast and Data

We use data from spacecraft located around the Lagrange 1 (L1) point. These spacecraft measure solar wind (SW) and interplanetary magnetic field (IMF) variables, such as solar wind bulk speed (V), density of protons (n), magnetic field components (B_x, B_y, B_z). Data is collected from the high resolution OMNI database (1 minute cadance) which contains all the above variables propagated at the position of the bow shock. The full dataset spans a period of nearly two solar cycles (1997-current).
The Kp index is obtained from GFZ Section 2.3, has a cadence of 3 hours and it is available starting from 1932.

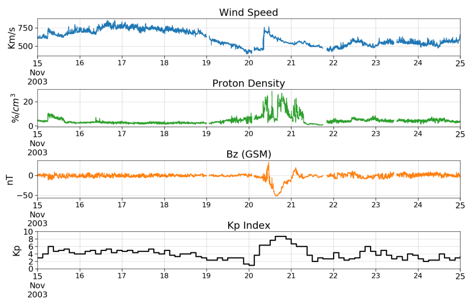

Figure 2: Example of OMNI solar wind and IMF measurements for the period 13 November - 25 November 2003) as well as Kp index.

The forecast model looks for a relation between the target variable $y = Kp(t_0 + h)$ and the input variables $x(t_0^-)$ up the current time t_0. The forecast horizon h ranges from zero (nowcast) up to several days ahead.
The data processing pipeline consists of the following independent and ordered steps:

- Time series feature engineering.
- Dimensionality reduction.
- Resampling algorithms (e.g. RANDOM or SMOTE)

Features are obtained by aggregation and statistical analysis. Each time a raw feature is selected together with a time window. In this range we evaluate the average, minimum or maximum values (e.g. $\langle V\rangle_{[t_0-1hour,t_0]}$ or $min(n)_{[t_0-3hour,t_0-1hour]}$). Through this method we build the following data sets:

- Solar wind: solar wind variables before current time t_0 and up to 9 hours before.
- Historical Kp: Kp index time series up to 9 hours before current time t_0.
- Recurrence: Kp and dolar wind variables 27 days (1 solar rotation) and 54 days (2 solar rotations) before the forecast time $t_0 + h$
- Full Model: features of all the previous 3 models combined.

Each of these data sets, if no resampling is applied, contains about $55k$ instances. The number of input features depends on the data set and ranges from few tens to few hundreds [1,2].

3. Artificial Neural Networks

We use feed-forward artificial neural network regression to forecast the Kp index at a later time. For this specific application a shallow neural network with only one hidden layer and one output layer is sufficient.

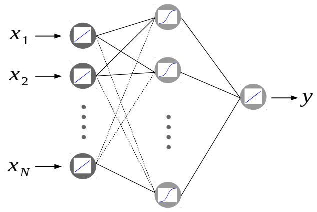

Figure 3: Schematics of a multilayer perceptron with one hidden layer. Activation functions are drawn for each neuron. The x_i are the input features, y is the target, i.e. $Kp(t_0 + h)$.

We use the Deep Learning Matlab toolbox with the Levenberg-Marquardt optimization to train the NNs, which converges faster and it more accurate than first order approaches. We train one network for each forecast horizon h.

4. Long term forecast results

We present a comprehensive picture of the Kp forecast results using the data sets described in Sec. 2, without resampling or dimensionality reduction. We compare with baseline models based on persistence, Kp average and solar cycle average [1]. The metric used is the cross-validation (CV) RMSE.

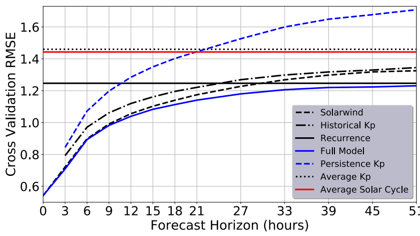

Figure 4: CV RMSE as a function of forecast horizon in various data models.

- Solar wind model performs best for short term forecast.
- Recurrence model kicks in for long term forecast. Both outperform the baseline models.
- The Full Model combines the power of both the Solar wind and Recurrence models.

5. Resampling Kp

The distribution of Kp is unbalanced towards small values.

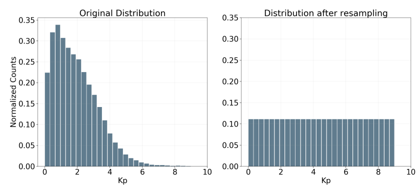

Figure 5: Kp distribution before (left) and after (right) full rebalancing.

The Neural Network adapts to perform better for small Kp, ignoring the higher less frequent values. Strategies:

- Weight the neural network error function during training by penalizing more the errors on high Kp.
- Find more discriminative features (Work in Progress)
- Rebalance (Oversample or Undersample) the data set to introduce more instances for high Kp.

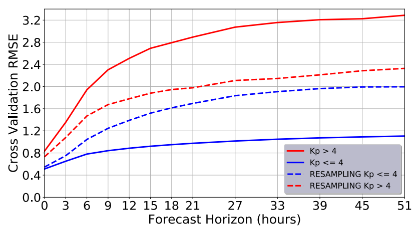

Figure 6: Effect of oversample rebalancing on Solarwind dataset model in the performance for quiet times and storm times as a function of forecast horizon.

The rebalanced model increases the forecast accuracy for high Kp.

6. Feature Ranking

Machine Learning algorithms allow the user to feed the model with many input features, since a model internally could perform feature selection. However understanding which feature actually contributes the most to the forecast can lead to a better understanding of the physical processes in the game. We compare four different approaches:

- Mutual Information Maximimization (MIM).
- Maximum Relevance Minimum Redundance (MRMR).
- Random forest (RF) regression feature importance.
- Fast Function eXtraction (FFX) algorithm coupled with cross-validation [2,3]. (proposed methodology)

The FFX algorithm builds symbolic expressions $\hat{y} = f(x_1, ...x_N)$ by combining functions of given functional space basis. We set up a 50-fold cross validation routine. For each partition we obtain the best regression function $y = f_i(x_1, ...x_N)$. Not every input appears in each expression. We keep only the features present in all 50 expressions.

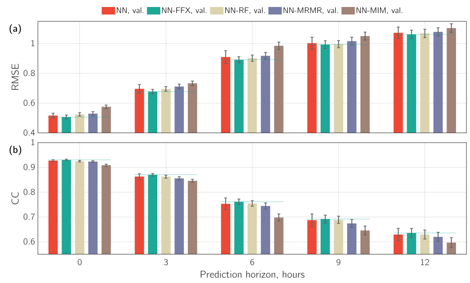

Figure 7: Performance comparison on validation sets using neural networks trained with selected features by different methods. The FFX feature selection perfoms marginally better.

7. Uncertainty estimation (Preliminary)

Models of Kp forecast output the best guess for future value of the index. However forecast models usually also provide a quantification the uncertainty of the estimated value. As our first attempt, we train two ensembles of 10 NNs each. The first uses bootstrapped versions of the training set without replacement, the second uses replacement. We evaluate the RMSE of the average model of both ensembles and compare it with the standard CV RMSE without bootstrapping. Additionally we attempt to measure the accuracy of the uncertainty by checking if estimated and real values are, in average, within an ensemble standard deviation to one another. To do so we measure the distance between real Kp and estimated Kp plus/minus one standard deviation.

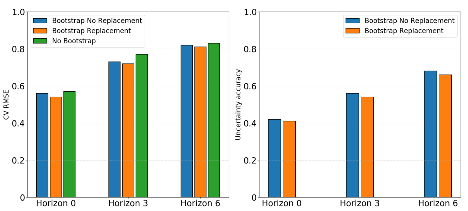

Figure 8: (Left) comparison between ensemble average RMSE for bootstrapped and CV RMSE. (Right) Accuracy of uncertainty estimation in bootstrapped ensembles.

We also implemented MC dropout and Bayesian regularization, however without a clear advantage to the above approach.

Bibliography

1 Y. Y. Shprits, R. Vasile, and I. S. Zhelavskaya (2019), *Nowcasting and Predicting the Kp Index Using Historical Values and Real-Time Observations*, Space Weather,17. https://doi.org/10.1029/2018SW002141

2 I. S. Zhelavskaya, R. Vasile, Y. Y. Shprits, C. Stolle, and J. Matzka (2019), *Systematic Analysis of Machine Learning and Feature Selection Techniques for Prediction of the Kp Index*, Accepted for publication in Space Weather

3 McConaghy, T. (2011). *FFX: Fast, scalable, deterministic symbolic regression technology*, Genetic Programming theory and practice ix(pp 235-260). New York, NY, Springer New York.

Acknowledgement

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776287, SWAMI. The Kp index was provided by GFZ Section 2.3 and can be found at *ftp://ftp.gfz-potsdam.de/pub/home/obs/kp-ap/wdc/*. The OMNI data has been downloaded from *ftp://spdf.gsfc.nasa.gov/pub/data/omni/high_res_omni/*.