Developing a Machine Learning Algorithm to Accurately Predict Geomagnetic Storms Early

Introduction

Geomagnetic storms (GS) occur when solar wind shock waves disrupt Earth's magnetosphere [4]. GS can cause severe damage to satellites, power grids, and communication infrastructures. Estimates of direct economic impacts of a large scale GS exceeds $40 billion a day in the U.S. [8]. A GS can cause electrical blackouts and internet outages on massive scales that may not be repaired for months.

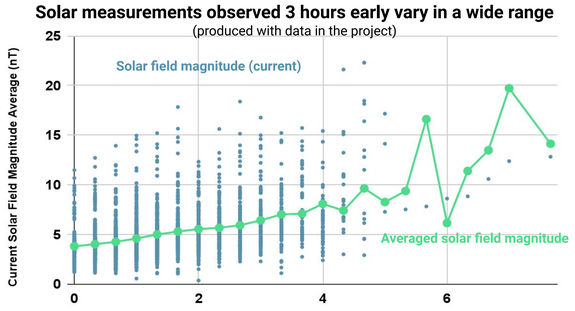
Early prediction is critical in minimizing the hazards. However, current models either don't predict the all types of GS storms, or only make predictions within one hour of the occurrence. One claimed to predict 6 hours in advance but failed to identify all types of geospace storms [9], and another lacked justification and sufficient detail [13].
This project aims to predict all geomagnetic storms reliably and as early as possible using machine learning.

Technical Challenges

Early prediction of geomagnetic storms has been attempted [] but further advance is extremely challenging for the following reasons:

- There is low correlation between current solar measurements and geomagnetic disturbance (measured by Kp index) in the future, making it challenging for early detection; Similarly, many other measurements of solar activity contain little information about the Kp index 3 hours later [2].

Solar measurements observed 3 hours early vary in a wide range
(produced with data in the project)

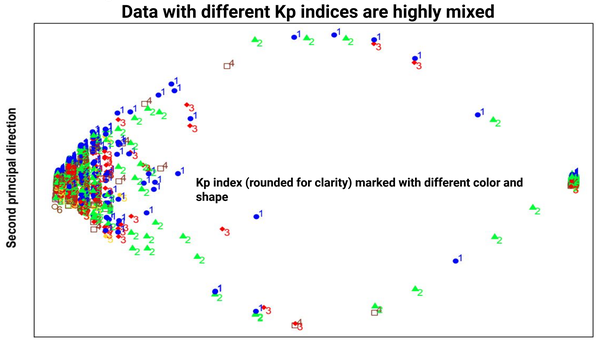


Kp Index 3 Hours Later

Kp index: ranging from 0-9, is commonly used to measure severity of geomagnetic disturbances; an index of 0 indicates no geomagnetic activity.

- Major storms (with high Kp indices) occur rarely [1], causing the data to be highly imbalanced. This can lead to the masking phenomenon, where classes with few data instances would be ignored by algorithm.
- Factors like unfavorable weather conditions or interferences from electromagnetic signals in space can introduce substantial noise or distortion to solar measurements when transmitted to ground station by satellites.
- Solar data is highly mixed for points with different Kp values, increasing difficulty of detection [2].

Data with different Kp indices are highly mixed



2D visualization of data used in training (originally 100 dimensions). The two dimensions in the graph are top two principal components (directions that data stretches the most). Figure shows how points with different Kp indices can be close, or similar feature conditions can lead to very different Kp indices. It also shows the highly nonlinear pattern between Kp indices and solar activity data.

- This project aimed to overcome the above challenges by learning highly nonlinear patterns using Random Forests regression, and using techniques including downsampling, feature selection, and combining data from multiple sources.

Method

- Random Forests regression, a machine learning model, is an ensemble of decision trees, and each tree recursively narrows down which Kp index instances should fall into. The final decision of a forest is taken from the average of decisions from individual trees
- Downsampling to discard data from larger class sizes. This creates a more balanced distribution of data across different classes which can often improve accuracy (Towards Data Science)
- Feature selection to include only the most important variables used in learning algorithm while discarding those weak or noisy variables. This helps reduce wrong generalizations deduced from data.

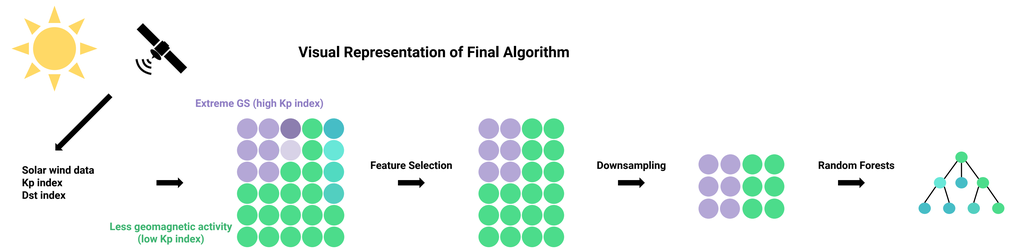
Data from multiple sources were used:
1) OMNIWeb solar wind data (wind speed, proton density, temperature, field magnitude average)
2) GFZ German Research Centre for Geosciences historical Kp indices
3) World Data Center for Geomagnetism Kyoto historical Dst indices

1. R program was used for programming. Solar wind data (measured every 5 minutes) up to 9 hours before prediction, Kp indices (measured every 3 hours) up to 24 hours, and Dst indices (measured every 1 hour) collected up to 3 hours before prediction were used and trained by Random Forests regression.

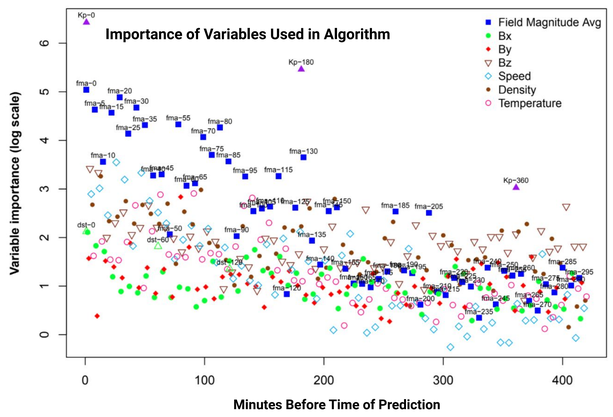
2. Feature selection was applied. The importance variables were determined and ranked. Less important variables were discarded.
3. Downsampling of data with lower Kp indices. This reduces the chance of masking of rare extreme storms and increases the accuracy.

4. Algorithms were tested with past data and results averaged over 100 runs. The accuracy was determined by finding the percentage of predicted Kp indices within 1.000 of the actual index.

Visual Representation of Final Algorithm



Results



Kp indices, FMAs, and other solar activity measurements taken at different time are arranged by the time they were measured, with more recent measurements displayed first.

- The Kp indices before the time of prediction and the field magnitude average (FMA) were the most important features as determined by Random Forests
- The importances decrease very quickly with time. As the Kp indices are measured every 3 hours, this also indicates the difficulty of early prediction at an even early time window (i.e., 6 hours ahead); less informative variables would likely not lead to satisfactory results. Thus, unless one can measure Kp indices more frequently, this algorithm's 3 hours in advance prediction has achieved the practical limit.
- Less important variables were discarded for improved performance.

Accuracy in Predicting Kp Index 3 Hours In Advance

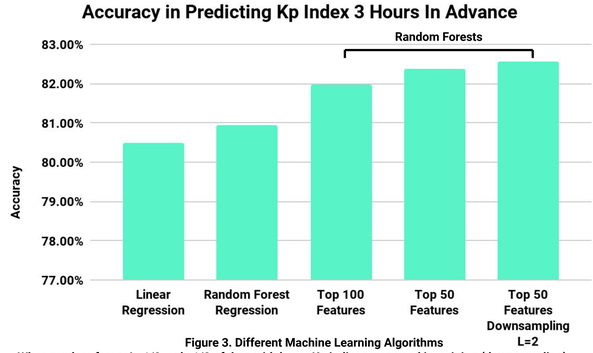
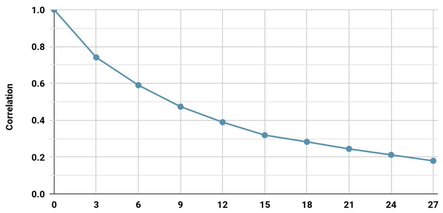


Figure 3. Different Machine Learning Algorithms
When random forest L=1/2, only 1/2 of data with lower Kp indices was used in training (downsampling).

- Random Forests regression outperforms competing algorithms
- Random Forests substantially outperforms linear regression, which is consistent with the nonlinear pattern underlying the solar activity as shown in the 2D visualization
- Feature selection of top 50 features improved Random Forests accuracy
- Downsampling with top 50 features was able to achieve 82.55% accuracy
- For the first time, the Kp index was successfully predicted 3 hours ahead
- The results are statistically significant, with $p < 10^{-3}$.

Results (cont.)

Correlation Between Predicted Kp Index and Kp Index n Hours Before



Correlation Between Predicted Kp Index and Kp Index n Hours Before

- **As the time between the two Kp indices increases (or n increases), the correlation significantly decreases.**
- **A good correlation between Kp index at 3 hours in advance ensure the reliable prediction of geomagnetic storms.**

Conclusions

- For the first time, the algorithm successfully predicted geomagnetic storms in terms of global index, Kp index, 3 hours ahead
- The proposed algorithm enables predictions three hours ahead of time at an 82.55% accuracy, which outperforms competing algorithms by ~3%
- By keeping only the most informative features, feature selection substantially improves the performance of the algorithm
- Downsampling data instances that occur more often (i.e., lower Kp indices) further improves the performance in prediction by effectively overcoming the masking phenomenon.
- The importance of variables from solar measurements decreases very quickly as the time before the prediction increases. The 6 hours in advance measurements become barely informative, indicating that the 3 hours in advance prediction as we achieved has reached the practical limit, unless the Kp indices can be measured more frequently.

Implications

- Predicting geomagnetic storms accurately in advance will provide a sufficient warning time window for preparation
- This will greatly decrease the amount of damage caused to satellites, power grids, and communication infrastructure.
- Future explorations include combining more algorithm models or to extend the algorithm to predict Kp indices 6 hours ahead

Reference

[1] Chapman, S. C., and R. B. Horne. "Using the aa Index Over the Last 14 Solar Cycles to Characterize Extreme Geomagnetic Activity." Geophysical Research Letters, https://doi.org/10.1029/2019GL086524. Accessed 26 Oct. 2021.
[2] Elliott, Heather A., et al. "The Kp index and Solar Wind Speed Relationship: Insights for Improving Space Weather Forecasts." Space Weather, vol. 11, no. 6, 2013, pp. 339–349., https://doi.org/10.1002/swe.20053.
[3] "Final Dst Index." WDC for Geomagnetism, Kyoto, https://doi.org/10.17593/14515-74000.
[4] "Geomagnetic Storms." *Geomagnetic Storms | NOAA / NWS Space Weather Prediction Center*, https://www.swpc.noaa.gov/phenomena/geomagnetic-storms.
[5] "Having an Imbalanced Dataset? Here Is How You Can Fix It." Towards Data Science, towardsdatascience.com/having-an-imbalanced-dataset-here-is-how-you-can-solve-it-1640568947eb. Accessed 26 Jan. 2022.
[6] "Kp Index." The Helmholtz Centre Potsdam - GFZ German Research Centre for Geosciences, https://www.gfz-potsdam.de/en/kp-index/.
[7] Mitsa, T. "How Do You Know You Have Enough Training Data?" Towards Data Science, towardsdatascience.com/how-do-you-know-you-have-enough-training-data-ad9b1fd679ee.
[8] Oughton, Edward J., et al. "Quantifying the Daily Economic Impact of Extreme Space Weather Due to Failure in Electricity Transmission Infrastructure." Space Weather, vol. 15, no. 1, 2017, pp. 65–83., https://doi.org/10.1002/2016sw001491.
[9] Podladchikova, T. V., and A. A. Petrukovich. "Extended geomagnetic storm forecast ahead of available solar wind measurements." Space Weather, https://doi.org/10.1029/2012SW000786. Accessed 26 Oct. 2021.
[10] "Real Time Solar Wind." *Real Time Solar Wind | NOAA / NWS Space Weather Prediction Center*, https://www.swpc.noaa.gov/products/real-time-solar-wind.
[11] Shprits, Yuri Y., et al. "Nowcasting and Predicting the Kp Index Using Historical Values and Real-Time Observations." Space Weather, vol. 17, no. 8, 2019, pp. 1219–1229., https://doi.org/10.1029/2018sw002141.
[12] "SPDF - Omniweb Service." *NASA*, NASA, https://omniweb.gsfc.nasa.gov/index.html. https://www.swpc.noaa.gov/products/planetary-k-index
[13] S. Wing, et.al., "Kp forecast modele". J. Geophysical Res. Vol. 10, Issue A4. https://doi.org/10.1029/2004JA010500