# Problem statement and hypothesis

Using observations and measurements of solar activity (sun spots, solar flares, coronal mass ejections) can we predict geomagnetic activity on Earth?

My hypothesis is that the intensity of geomagnetic storms on Earth increases when solar activity prior to a major event such as a coronal mass ejection is elevated. If my hypothesis is correct, the strongest geomagnetic storms on Earth will be foreshadowed by a period of higher than mild solar activity culminating in a large burst of energy from the Sun.

Glossary

|  |  |
| --- | --- |
| K indices | K indices isolate solar particle effects on the earth's magnetic field. Each activity level relates almost logarithmically to its corresponding disturbance amplitude. K indices range in 28 steps from 0 (quiet) to 9 (greatly disturbed) with fractional parts expressed in thirds of a unit. A K-value equal to 27, for example, means 2 and 2/3 or 3-; a K-value equal to 30 means 3 and 0/3 or 3 exactly; and a K-value equal to 33 means 3 and 1/3 or 3+. Kp is measure over every 3 hour period and then averaged. |
| Ap indices | The Ap-index ranges from 0 to 400 and represents a K-value converted to a linear scale in gammas (nanoTeslas)--a scale that measures equivalent disturbance amplitude of a station at which K=9 has a lower limit of 400 gammas. |
| Cp | Cp or PLANETARY DAILY CHARACTER FIGURE--a qualitative estimate of overall level of magnetic activity for the day determined from the sum of the eight ap amplitudes. Cp ranges, in steps of one-tenth, from 0 (quiet) to 2.5 (highly disturbed). |
| C9 | A conversion of the 0-to-2.5 range of the Cp index to one digit between 0 and 9. |
| Solar Flare Classifications  X, M, C and S | <http://spaceweather.com/glossary/flareclasses.html>  “A solar flare is an explosion on the Sun that happens when energy stored in twisted magnetic fields (usually above sunspots) is suddenly released. Flares produce a burst of radiation across the electromagnetic spectrum, from radio waves to x-rays and gamma-rays.  Scientists classify solar flares according to their x-ray brightness in the wavelength range 1 to 8 Angstroms. There are 3 categories: X-class flares are big; they are major events that can trigger planet-wide radio blackouts and long-lasting radiation storms. M-class flares are medium-sized; they can cause brief radio blackouts that affect Earth's polar regions. Minor radiation storms sometimes follow an M-class flare. Compared to X- and M-class events, C-class flares are small with few noticeable consequences here on Earth.”  The S classification is an older type based on a reading in a different spectrum called Hydrogen-alpha. |
| International Sunspot Number (ISN) | A quantity that measures the number of sunspots and groups of sunspots present on the surface of the sun. <https://en.wikipedia.org/wiki/Wolf_number> |
| Coronal Mass Ejection | <https://en.wikipedia.org/wiki/Coronal_mass_ejection>  “A coronal mass ejection (CME) is an unusually-large release of plasma from the solar corona. They often follow solar flares and are always present during a solar filament eruption. The plasma is released into the solar wind, and can be observed in coronagraph imagery.  Coronal mass ejections are often associated with other forms of solar activity, most notably solar flares or filament eruptions, but a broadly accepted theoretical understanding of these relationships has not been established. CMEs most often originate from active regions on the Sun's surface, such as groupings of sunspots associated with frequent flares. Near solar maxima, the Sun produces about three CMEs every day, whereas near solar minima, there is about one CME every five days.” |
| Proton Flux | A measure of how many protons per sq cm by energy level measured in millions of electron volts MeV. |
| Kp\_max | The maximum K index achieved during a single day. |
| C9-n or C9+n, where n is {1, 2, 3, 4, 5} | Represents the C9 value n days before or after the current date |
| RadioFlux | The background radiation from the sun in the radio frequency spectrum |
| P1MeV, P10MeV, P100MeV | Proton flux counts at 3 different energy levels. Protons/cm3 |
| E08MeV, E2MeV | Electron flux counts at 3 different energy levels. Electrons/cm3 |
| Proton\_density | Protons / cm3 |
| Proton\_temp | The temperature of protons in degrees Kelvin |
| Proton\_speed | The magnitude of the velocity of the protons in km/s |
| He4toproton | The ratio of alpha particles to protons |
| LinearSpeed / 20R\_max | The measured speed of a CME at the edge of the sun’s disk and at 20 solar radii. i.e. two measurements separated in time. |

# Summary of Approach

Figure

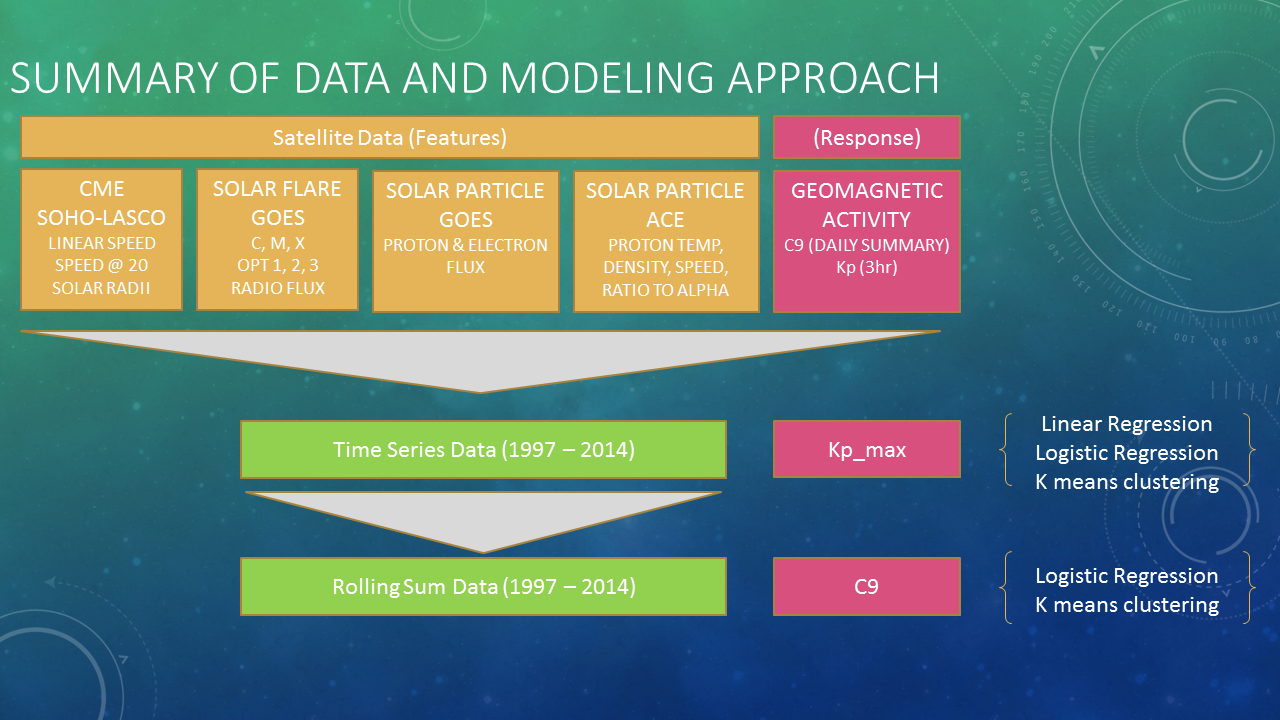


Figure 1 is an overview of the data gathering, summarization and modeling approaches employed. Satellite data from several sources was combined along with terrestrial geomagnetic data from earth based observations. The resulting dataset is about 6,000 rows. I worked with two variations on the response variable. The maximum value of the K index for the day as well as the average (C9). I initially started with the second tier evaluation on the rolling sum data and later backed into the Kp\_max analysis on the simpler dataset.

# Description of your data set and how it was obtained

The data was obtained from the National Geophysical Data Center which records and aggregates information from several observatory stations on Earth. Satellite observations of solar activity are from NASA’s SOHO LASCO CME CATALOG. Both sources of information are in multiple .csv files.

The Selected Geomagnetic and Solar Activity Indices from the National Geophysical Data Center is located at this [url](ftp://ftp.ngdc.noaa.gov/STP/GEOMAGNETIC_DATA/INDICES/KP_AP/%23kp_ap.fmt%23) and the data dictionary is located [here](ftp://ftp.ngdc.noaa.gov/STP/GEOMAGNETIC_DATA/INDICES/KP_AP/%23kp_ap.fmt%23) There is one fixed width text file per year. The relevant data in the file is the **C9 field** which as cumulative measure of geomagnetic disturbance for the day. C9 = 0 is a very quiet day on Earth, C9 = 9 is an extremely active day. The **ISN field** is the International Sunspot Number and is a relative measure of the number of sunspots identified on the surface of the Sun on that day.

I later added another NGDC data set for daily solar activity located [here](ftp://ftp.swpc.noaa.gov/pub/warehouse/). There is an ftp folder for each year and a file called YYYY\_DSD.text for the specific year. DSD stands for Dailey Solar Data. The relevant data in these files are the counts of x-ray flares by energy category: S, C, M, X. As well as optical flares by energy category: 1, 2, and 3.

In addition, I added another NGDC data set for daily solar particle activity located with the daily solar activity linked to above. The relevant data in these files are flux measurements, particles per square centimeter of proton flows at 3 different energy levels.

In addition, I added solar wind data from the [ACE satellite](http://www.srl.caltech.edu/ACE/ASC/level2/lvl2DATA_SWEPAM.html). That data required summarizing multiple measurements in one day into a single daily value. In this case, I was interested in the maximum values for the day in order to correlate them to the maximum Kp value observed for that day.

NASA’s SOHO LASCO CME catalog is located at this [url](http://cdaw.gsfc.nasa.gov/CME_list/UNIVERSAL/text_ver/). The file **univ\_all.txt** contains the entire catalog and the data dictionary is located [here](http://cdaw.gsfc.nasa.gov/CME_list/catalog_description.htm). The relevant data in the file is the **Linear Speed** and **Speed at 20 Solar Radii**.

I also explored many catalogs of information in the hope of getting specific energy levels for the solar flare counts in the DSD files. SpaceWeatherLive.com has a fantastic searchable database but it was not suitable for downloading a series of events. I reached out the site admin and received a reply pointing me back to the NGDC data. I’m still looking for the correct files for this information.

# Description of any pre-processing steps you took

All data sets are in a fixed file format with footnote indicators embedded in the data. It was necessary to separate out those indicators into separate fields for interpretation.

The file formats for the NGDC DSD data were consistent except for the number of header rows at the top of the files. For some years there was an extra header row. Rather than building a routine to catalog which had the extra rows, I made a local copy of the data and deleted the extra row. For years 2014 and 2015, the headers were repeated in the body of the file and I removed them from my local copies. For years 1995 and 1996, the date format started out as DD MMM YY but in the later files the format is YYYY MM DD. Rather than deal with both formats, I omitted using 1995 and 1996.

The solar indices data is compete except for the ISN numbers for 2015, so I omitted that year from the analysis.

The NASA’s SOHO LASCO CME catalog has a number of data issues in terms of missing data and data flagged as unreliable. I focused my attention on the fields where the quality was not in question. One adjustment I needed to make to this data was to aggregate multiple measurements for a single day. So that it can be matched to the daily geomagnetic and solar indices from the other data sets. I used the maximum value of Linear Speed and Speed at 20 solar radii as the aggregation value. My thinking here is that the dominant effect would be by the largest CME for that day. In the cases where there were no CME’s on that day, I set the speeds to 0.

The most complex part of the pre-processing involves summarizing prior history and adding it as features to each row. For example, I wanted to get a count of solar flares of a certain type over the past week and add it as feature for that day. In addition, I wanted to know if prior C9 values was a good predictor of the current value. I used the .shift method on the datetime index to shift the data by n days and then appending the features of interest to my dataset. For example, after shifting the date by 1 day all the measurements for 12-31-13 are now associated with 1-1-14 in a new dataframe, I can then join the dataframes on date and pick up the historical features. Pretty slick! I also added some code to summarize some feature over a historical date range (e.g. the prior week leading up to that day) and add it using the technique above.

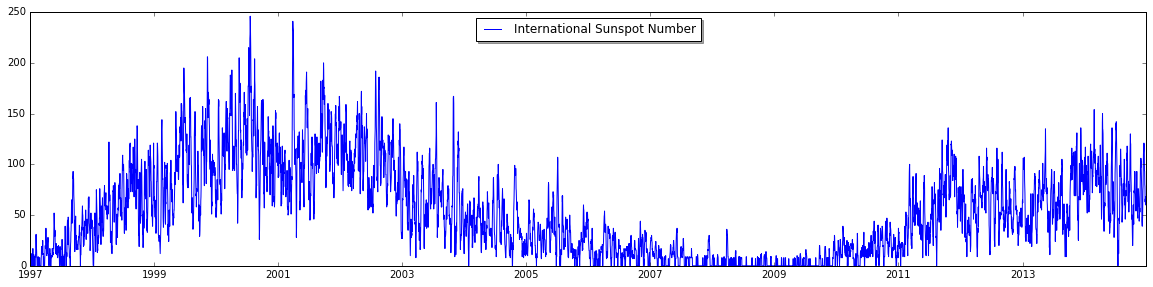
The solar wind data from the ACE satellite posed a challenge in that the date of the measurement was given in days since January first of that year so a conversion we necessary to convert it into an actual date. The data itself was composed of measurements taken every 64 seconds, so it was necessary to do a considerable amount of aggregation for this analysis.

I decided to limit my analysis from 1997 to 2014 based on the quality of the data in those years.

# What you learned from exploring the data, including visualizations

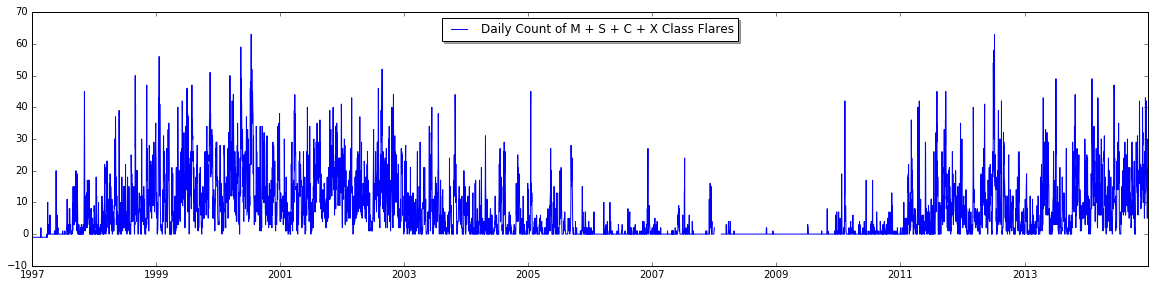
Some expected patterns are present in the data. For example, our sun is on an 11 year solar cycle where the number of sunspots displays a minimum and maximum over this period.

Figure



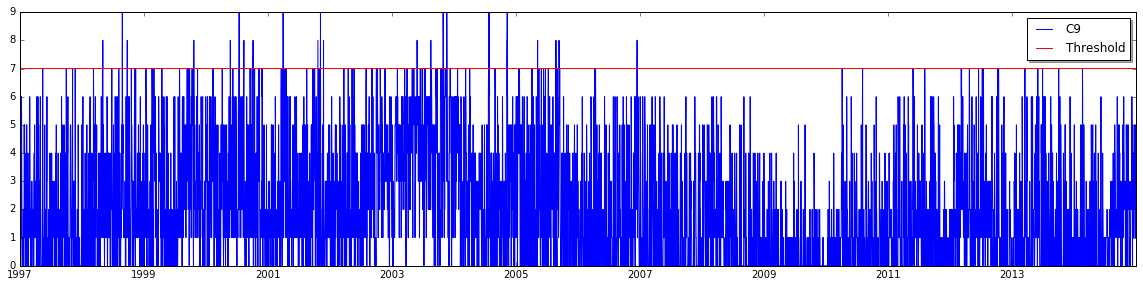
The occurrence of solar flares follows a similar pattern as expected

Figure



However, the occurrence of extreme geomagnetic activity on the Earth is a bit less regular

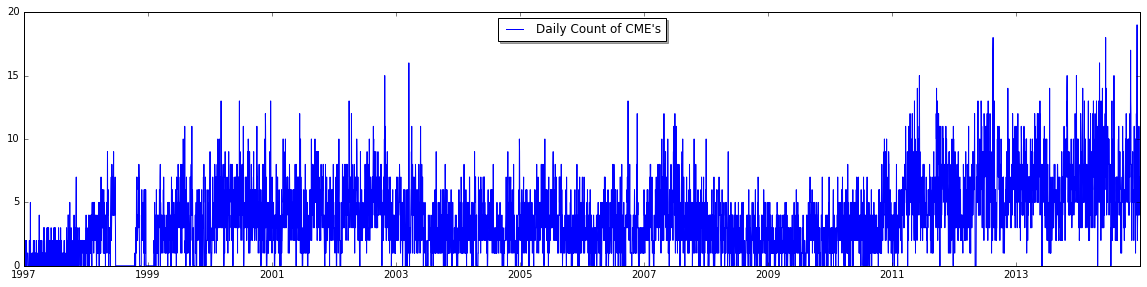
Figure 3



The points above the red threshold line are the extreme events of particular interest for this analysis

In addition, the pattern of CME’s observed during this period was different with some of the most active years later in the graph, but the most extreme geomagnetic days in the earlier years.

Figure 4



It was clear from the simple visualizations that the relationship between events on the sun and those on the Earth is not a simple relationship.

# How you chose which features to use in your analysis

Initially, I chose features that show some initial correlation to the C9 (geomagnetic activity) data: International Sunspot Number and Solar Flare Counts. I also used the CME data thinking that it must have some correlation to since some solar flares are also accompanied by a CME. I later expanded the number of features to include aspects of the solar wind.

# Details of your modeling process, including how you selected your models and validated them

Link to the Jupyter notebook analysis is in this [script](https://github.com/wayneheller/GA-Student-Project/blob/master/ipythonscripts/Intial_Analysis_Cp_and_CME.ipynb). It has all the analysis leading up the Kmeans clustering. I created a second notebook for some logistic regression models [here](https://github.com/wayneheller/GA-Student-Project/blob/master/ipythonscripts/Second_Analysis.ipynb). I should probably combine these two at some point.

## Linear Regression

Figure 5 below is a correlation matrix for all the features in this analysis. Several of the features are correlated based on the natures of their definitions:

* C9, KpSum, Kp\_max all of these are measure of the same geomagnetic activity
* ISN and SECS\_SSN are two standards for measuring sunspot number
* Proton speed and temperature are different measures that are correlated to the energy of the solar wind
* LinearSpeed and 20R\_max are two measurements of the same phenomena (CME)

Figure 5

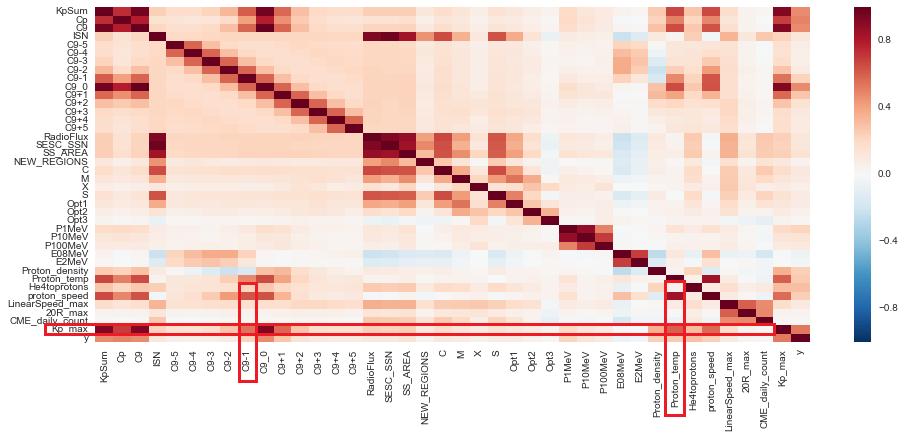
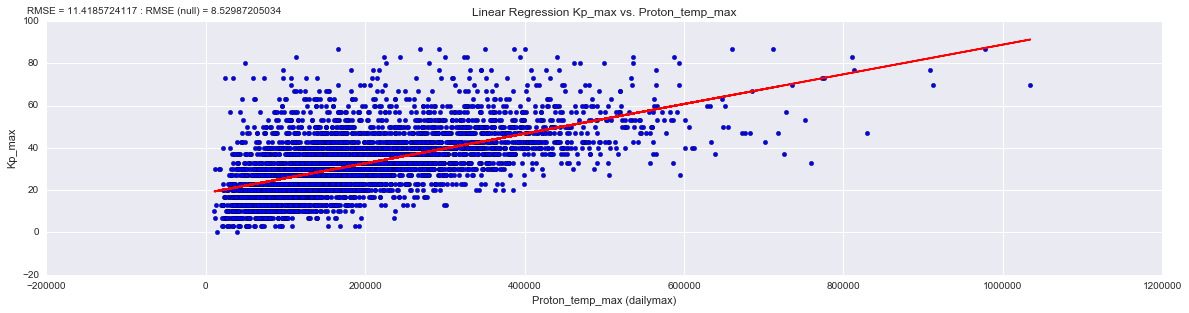


Figure 6 below is representative of all the regression models on the correlated features. **Although there is a noticeable linear relationship to the data, the root mean square error of the model is great than of the null hypothesis using the mean of Kp\_max.**

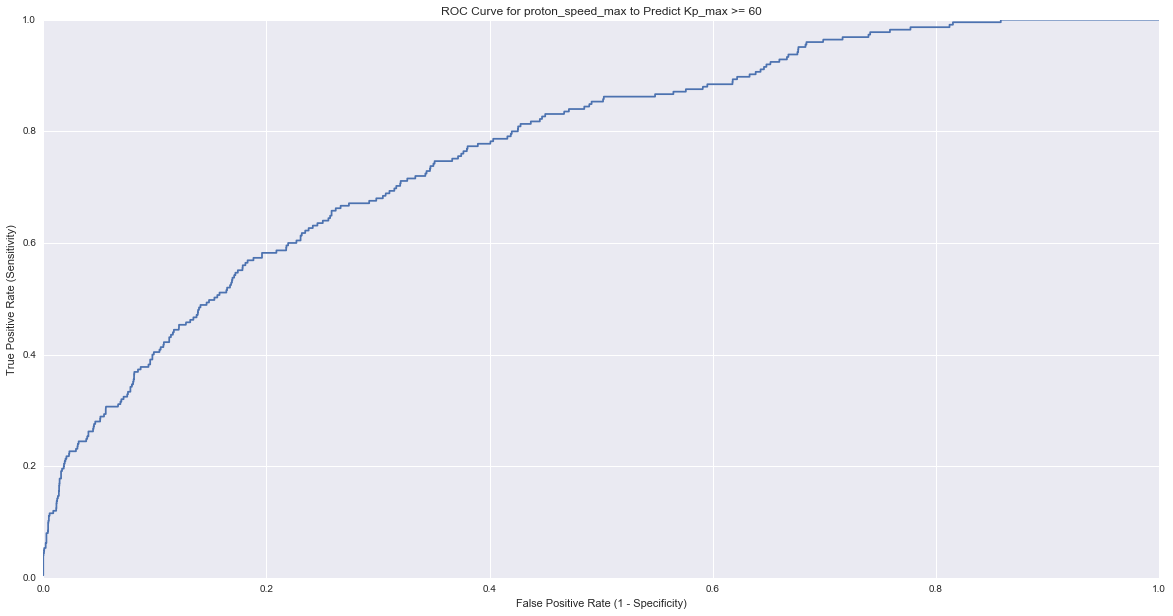
Figure 6



## Logistical Regression

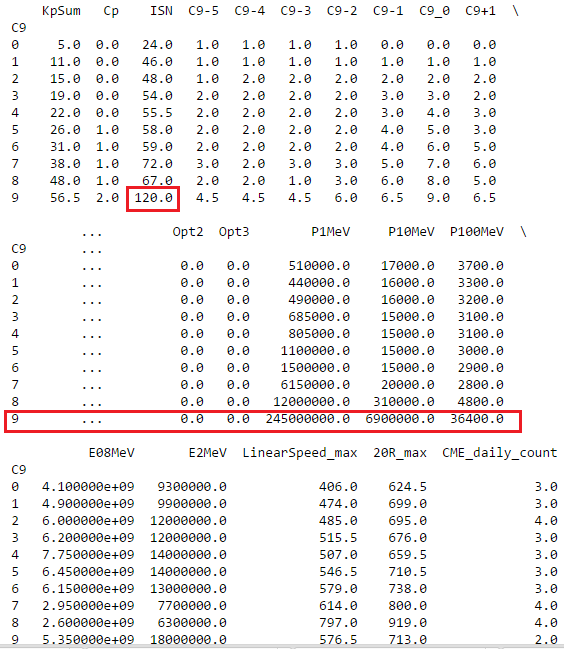
When the value of Kp > 60 there is a high probability that the geomagnetic activity will result in a substantial Aurora. Given the lack of predictive power of the models to date, I tried logistical regression in combination with several of the features to predict these events. Most of the curves followed this typical pattern as in Figure 7 below. To achieve and 80% true positive rate, the false positive rate is well over 40%. Not exactly a model you want to use when planning a trip to the Northern latitudes in order to see the Northern Lights.

Figure 7



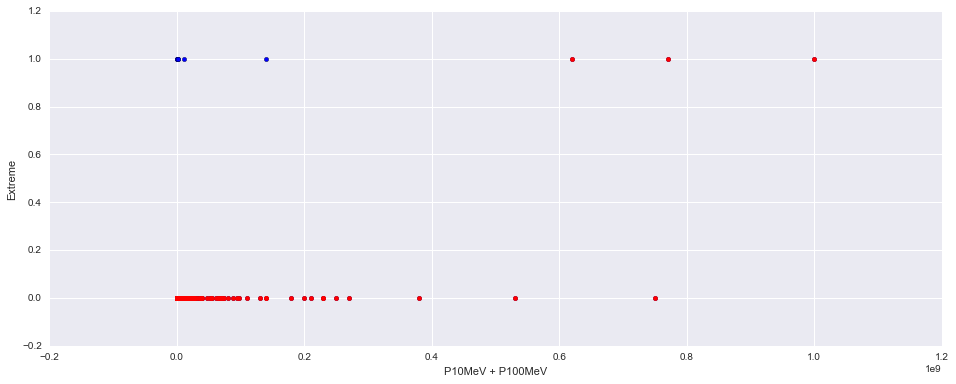
The original goal of this analysis is to predict C9=9. Based on the summary of result in Table 1 features it appears that the Proton flux counts would be good predictor of the C9=9 days

Table



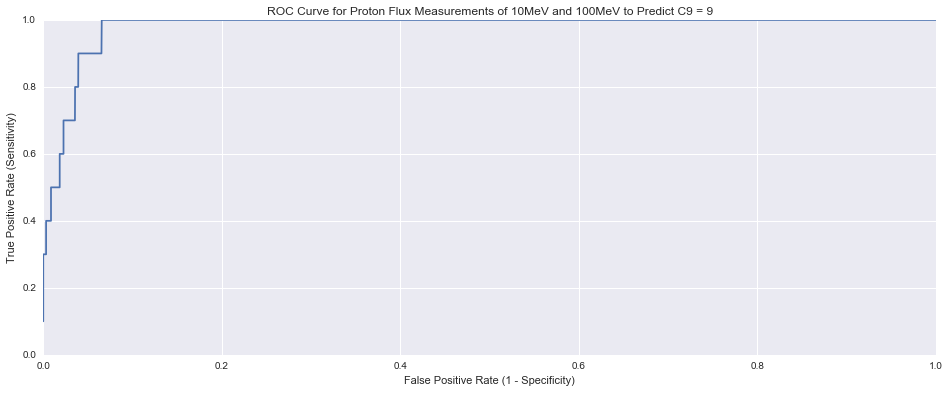
Below are the results of running a Logistic Regression model on Proton flux counts at 10MeV and 100MeV (the features) and C9=9 (response). It turns out that adding ISN as a feature diminishes the results.

Figure 6



Below is the associated ROC curve with a score of .980. Despite the high score, this is not a good predictive model for the extreme whether days due to the large number of false positives. Out of the 10 extreme C9=9 days, the model only predicted 3 of them. Most of its predictive power is in predicting majority of the events C9<9.

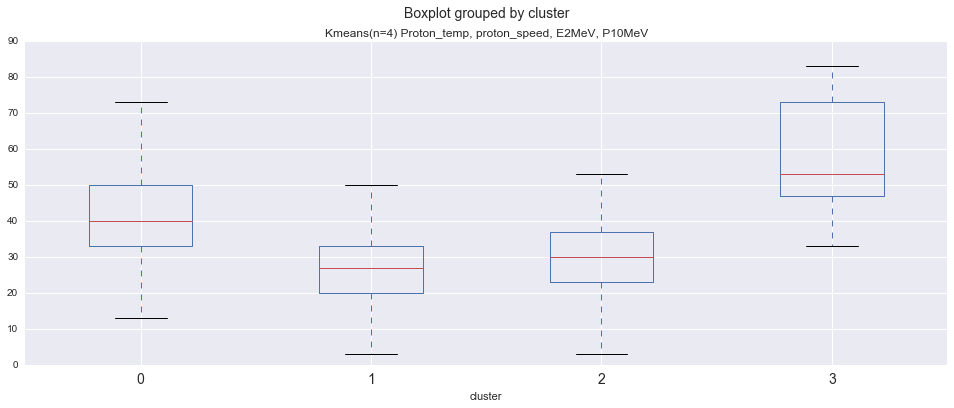
Figure 7



## Clustering Features

## I next tried Kmeans to better understand the information and to see if I could create a cluster that would be representative of the high geomagnetic activity days/events. I used the silhouette coefficient calculations as well as box chart diagrams against the Kp\_max in an attempt to isolate a set of features that would yield clusters differentiated on Kp\_max.

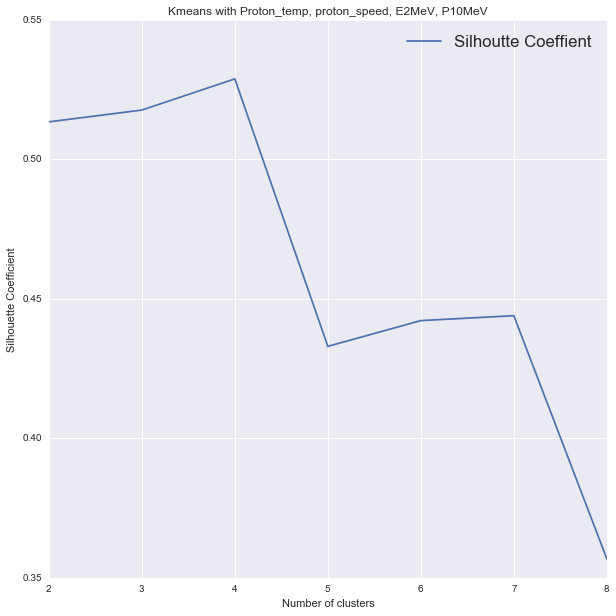
Figure 8



The features of proton temperature and speed, and electron and proton flux provided the best clustering alternative. Again the goal is to isolate the Kp\_max > 60 events. Cluster 3 accomplishes this quite well. Unfortunately, the size of the long tail of cluster 0 diminishes the effect. There are only 5 events in cluster 3 and almost 1500 is cluster 0.

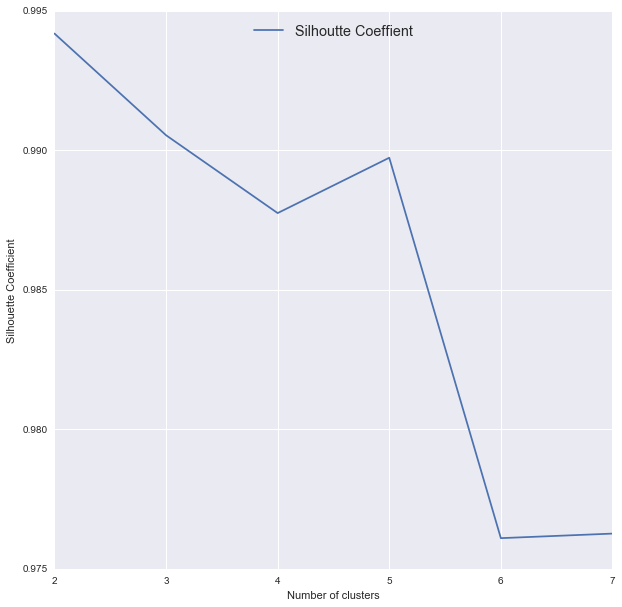
Evaluation cluster 3 as a predictive model yields only an accuracy of 1.2% for Kp\_max > 60 events.

Figure 9

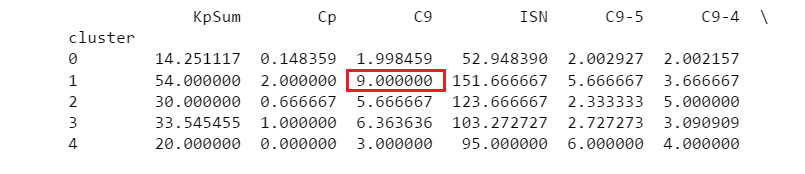


Through selectively combining features, the most promising results are using the features of the proton flux measurements (P1MeV, P10MeV, P100MeV). The graph of the silhouette coefficient is below Figure 10.

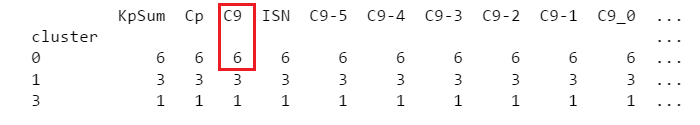
Figure 10



Using n=2 did not provide a meaningful clustering in terms of distinguishing the C9=9 events. Using n=5 yielded more promising results. One of the clusters was able to isolate these events.



However, most of the C9=9 days were clustered in with the majority of the low C9 days



## Clustering Features with Rolling Sums

For each of the features I created a new feature that is a rolling sum of the previous n days of history. I also varied the rolling sum range from 8 days to 2 days to test the cumulative effect of solar flares, cme’s, proton flux, etc.

Using Proton temperature summed over the last two days, I was able to create a predictive model for Kp\_max with an accuracy of ~25%.

# Your challenges and successes

Biggest recent challenge was finding the right catalogs of information and establishing a clear understanding and limitations of the information.

Because the response vector for the C9=9 events is so small compared to the overall dataset, I’m having a challenging time finding the right predictive model. I’ve lowered the threshold to C9 >=7 and still have not yielded good results. My next step is to continue to explore more features to add to the model. For example, there is some vector data associated with the solar wind speed that could help filter out some of the days with high proton temp but low Kp\_max values.

# Possible extensions or business applications of your project

I’ve read a couple of papers describing the use of neural networks for the prediction of sunspots but it seems like the use of these predictive techniques in this field is just emerging.

# Conclusions and key learnings

Although I’ve yet to create a predictive model, though the use of these modeling techniques I have a better understanding of the features available to build the model. There are indications in the data, for example, the linear regression correlation to proton temperature and speed that lead me to believe that exploring the other attributes associated with this information would improve the model.