# 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.

# 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.

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.

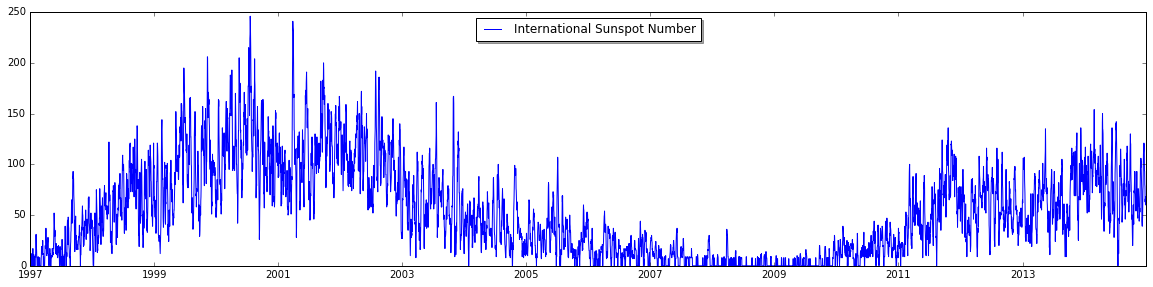
The file formats for the NGDC DSD data were consistent except for the number of header rows at the top of the files. For some years there was an extra header row. Rather than building a routine to catalog which had the extra rows, I made a local copy of the data and deleted the extra row. For years 2014 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.

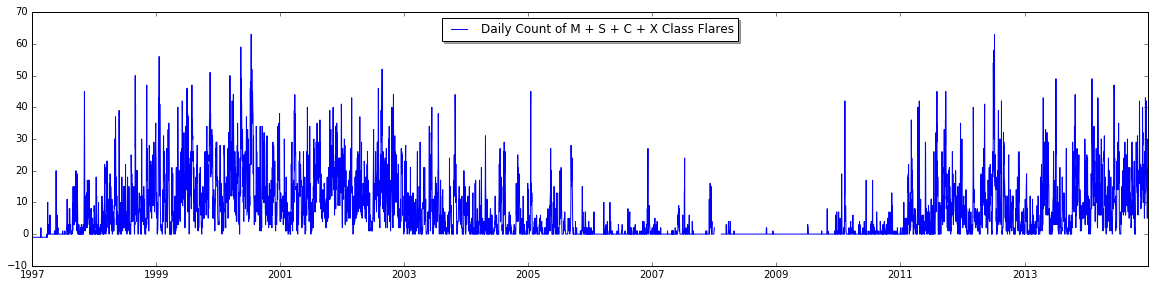
The NASA’s SOHO LASCO CME catalog has a number of data issues in terms of missing data and data flagged as unreliable. I focused my attention on the fields where the quality was not in question. One adjustment I needed to make to this data was to aggregate multiple measurements for a single day. So that 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.

# What you learned from exploring the data, including visualizations

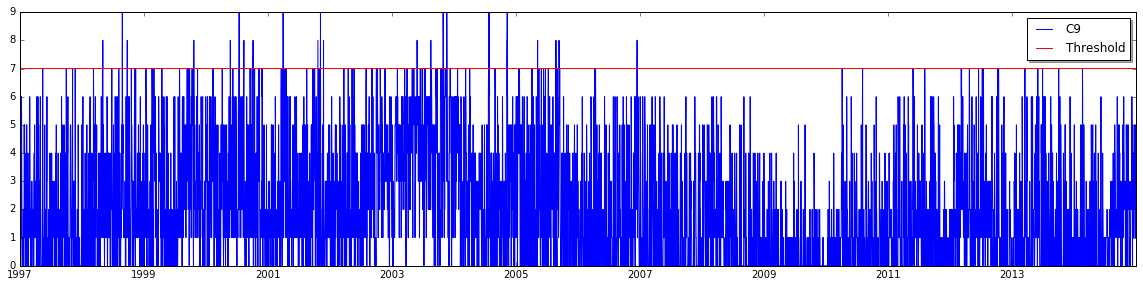
Some expected patterns are present in the data. For example, our sun is on an 11 year solar cycle where the number of sunspots displays a minimum and maximum over this period.



The occurrence of solar flares follows a similar pattern as expected

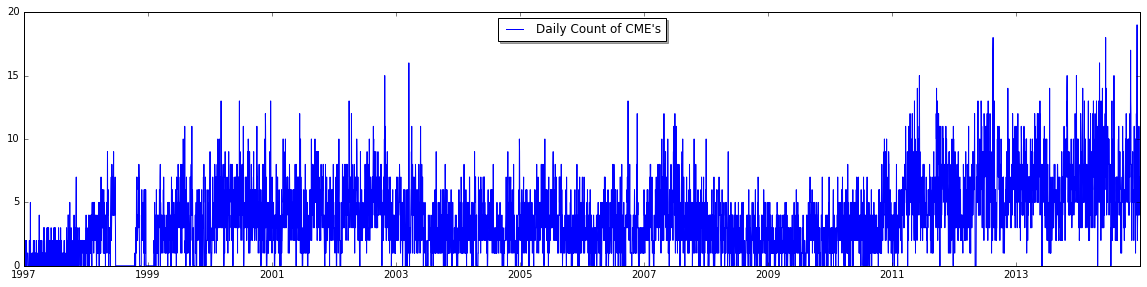


However, the occurrence of extreme geomagnetic activity on the Earth is a bit less regular



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.



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

Still a work in progress. I originally thought to focus on the CME features but am now pivoting to using counts of CME’s and solar flares.

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

Still a work in progress. Thinking that I’m going to use a clustering technique to classify the events that cause massive geomagnetic disturbances.

# Your challenges and successes

Biggest recent challenge was finding the right catalog of information on the solar flares

# 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