Dan Wald, Jason Harville, Jan Forslow
Project design document
datascience@berkeley MIDS Python Bridge Course (W18) – Project 2 - Team Assignment

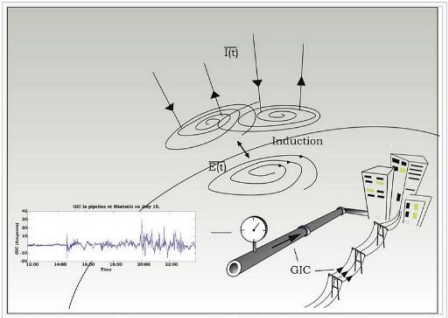
Project 2: Review of Geospatial and Land Surface Weather Impact on Energy Transmission Losses

Abstract

Power generation and transmission across the Unites States represent several significant engineering challenges: produce too little and there will be blackouts, too much and the extra energy generated is lost as there is no method to store bulk power. It is due to this fundamental trade that it is critical to understand the causes of transmission losses, derived by the formula:

$$Transmission\ Loss = Net\ Generation\ + Net\ Imports\ - Direct\ Use\ -\ Sales$$

Transmission and distribution losses in the USA were estimated at ~6.5% within the date range of 2005-2009. Two known causes of transmission loss are thermal equilibrium loss across long distances and high magnetic flux interfering with transformer inductance. Thermal losses are explained easily enough by the first law of thermodynamics: The law of conservation of energy states that the total energy of an isolated system is constant; energy can be transformed from one form to another, but cannot be created or destroyed. Therefore, significant changes in temperature can cause more or less energy to be inserted into the equation, seen in the form of power loss. Magnetic flux interactions are less obvious but equally well documented. Transformers exist across the power grid to amplify power and allow long distance transmission. This amplification is allowed by the use of magnetic induction which acts as a power amplifier by use of strong electro magnets. These transformers are disrupted by larger than normal ionosphere magnetic flux, caused by Coronal Mass Ejection interaction with the earth's Magnetosphere. This project pulls together power, weather and space weather data collected from 2005-2009 to create a unique data environment to explore the interaction and correlation between these three factors.



The basic principle for the generation of GIC: variations of the ionospheric currents (I(t)) generate an electric field (E(t)) driving GIC. Shown are also real GIC recordings from the Finnish natural gas pipeline.

Figure 1 Schematic of the formation of GIC interaction between the magnetosphere and power lines

Trade Study: Localized vs. Aggregate Data

Because of free interstate trade and lack of regulations corresponding with state or municipal borders, power data for the entire Continental United States has been used to avoid over-localization which is affected by business decisions to export or import power across state borders. By investigating the entire country, we can investigate a closed loop system to avoid misleading data.

To understand the difficulty in estimating monthly net imports, it is important to consider that the electric transmission authority in the United States is not divided by political boundaries such as states and counties. Instead, regional organizations of varying size are granted regulatory authority over regions that can be as small as a couple of counties or can span all or part of multiple states. Further complicating this irregular territory distribution is the fact that different regions of the country import/export under different circumstances. For example, California has hot summers with very high cooling demand. This drives imports at the same time as net generation increases, making California a good candidate for the method of apportioning shares of annual net imports according to net generation. However, on the other side of that coin are states like Washington, which exports large amounts of hydroelectric power to California during those same summer months while still experiencing hot summer months with highs in the 90s. We can see then, that power imports are proportional to net generation for California, but inversely proportional for Washington. It is for these complex reasons that we decided to aggregate the data across the entire US in order to remove net trade from the study of space and terrestrial weather impact to the transmission itself.

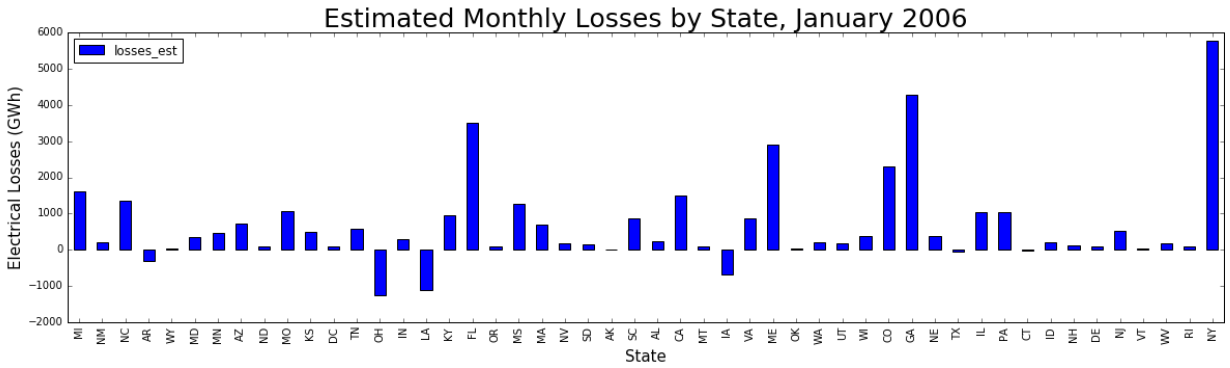


Figure 2 Estimated losses across each state line show a net trade of near zero. Negative loss is indicative of significant export of power to another state. This plot shows trade rather than environmental impacts, therefore the study was performed on aggregate US data

Hypothesis

Our team hypothesis is that a strong correlation between power loss and surface temperature will be present while only minor effects will be seen from magnetic flux

.

Analysis Methodology / Results

While our group had preconceived notions of correlations to expect, we decided instead to allow the data to guide our analysis path. After joining the data, we accomplished this by creating a new, helper DataFrame utilizing the $DataFrame.corr()$ function and sorting on absolute value in order to obtain a top ten list. Once we sorted through the values and eliminated non-value added correlations (max and max_abs or mean and median), we ended up with the top 10 correlations as seen in the table.

Table 1 Correlation Matrix

Ranking	Variable X	Variable Y	Correlation (abs)
1	Total Generation	Cooling Degree Days	0.837
2	Total Retail Sales	Cooling Degree Days	0.825
3	Total Industrial Gen.	Cooling Degree Days	0.751
4	Total Generation	Deviation from Average Temperature	0.716
5	Total Retail Sales	Deviation from Average Temperature	0.700
6	Max Temperature	Sum Magnetic Field Earthbound	0.794
7	Average Temperature	Abs Mean Magnetic Field Earthbound	0.786
8	Total Generation	Transmission Losses	0.610
9	Total Generation	Minimum Temperature	0584
10	Total Generation	Average Temperature	0.567

After that the first analysis step included to find out which variables that had the highest correlation to Transmission Losses inside the dataset.

As expected, Transmission Losses were most strongly correlated to Total Generation (0.61 correlation coefficient) and other similar variables describing the overall load on the electricity power grid. The theoretical explanation for this relationship is that for a given amount of power, a higher voltage reduces the current and thus the resistive losses in conductors (transmission lines).

We also identified a set of second-level influencers on Transmission Losses in various temperature measurements at a correlation coefficient slightly below 0.5 (specifically Cooling Degree Days). Cooling Degree Days is a measure of how much (in degrees), and for how long (in days), outside air temperature is higher than a specific indoor base temperature. It is used for calculations relating to the energy consumption required to cool buildings and showed in the analysis to have a high correlation with both Total Generation and Transmission Losses.

The resistance of a conductor increases with its temperature, which means that temperature changes in electric power lines has an effect on power losses in the line. For the Land Surface Average Temperature data at hand we discovered a seasonal effect in its relationship to Total Generation and Transmission Losses. Both colder months (September to February) as well as warmer months (March to August) had peaks for the latter two but not for the Average Temperature. The reason for this is that the impact of higher Total Generation for heating during colder months affects Transmission Losses more than impact of the temperature itself.

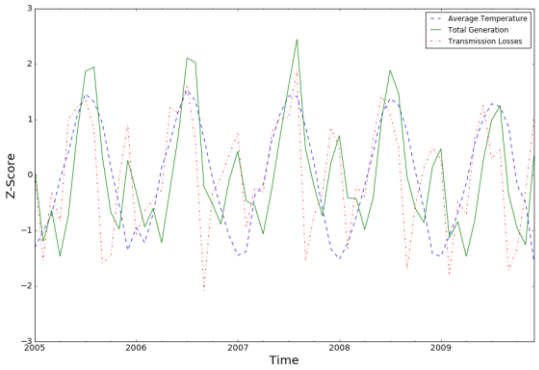


Figure 3 Normalized Z-Score for 3 Key Variables: multiplier of standard deviation delta from mean

In order to compensate for higher transmission losses for both high and low temperatures, we revisited the correlations for Transmission Losses and Average Temperature for each season. The graph below shows the warmer months (March to August) correlation (0.83) for the season much higher than the year (0.30).

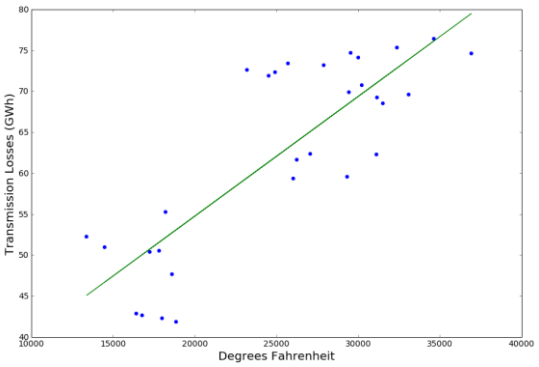


Figure 4 Monthly Transmission Losses vs. Average Temperature (September - February)

In order to obtain the impact of Total Generation on Transmission Losses by season, we set out to do a linear regression analysis on this relationship as a first-order correlation. The linear regression as shown (fig. 5) gave us an equation of Transmission Losses (y) = -32,257+0.156x, where x is Total Generation.

With the linear regression in hand, we next compared the predicted Transmission Losses vs. Actuals to obtain a residual for each observation (Table 2).

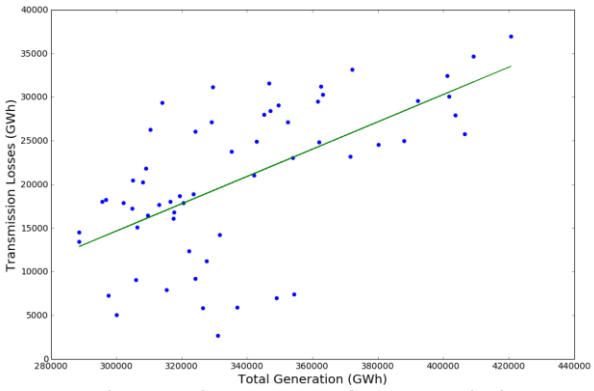


Figure 5 Linear Regression Transmission Losses vs. Total Generation

Table 2 Table 2 Correlation Matrix between Total Generation, Actual Transmission Losses and Predicted Transmission Losses and Residual based on Linear Regression.

	gen_tot	losses2	losses2_pred_gen_total	losses2_resid_gen_total
gen_tot	1.000000	0.609972	1.000000	0.031018
losses2	0.609972	1.000000	0.609972	0.810962
losses2_pred_gen_total	1.000000	0.609972	1.000000	0.031018
losses2_resid_gen_total	0.031018	0.810962	0.031018	1.000000

The residual effects were then checked for 'first-order' correlation with second-tier variables using built-in function corr(). The results of this normalized correlation were surprising in that the correlations were stronger for Total Magnetic Field (0.30) than Land Surface Temperature (-0.04)

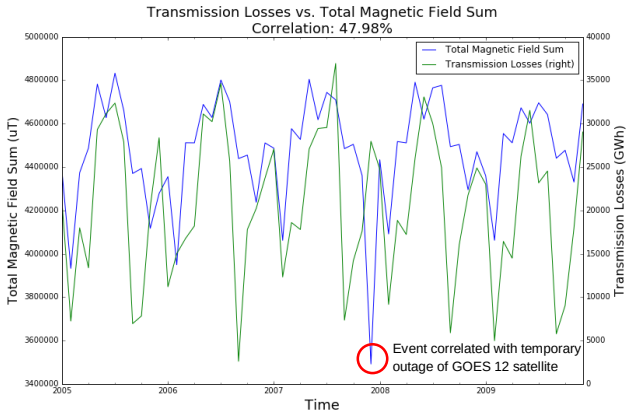


Figure 6 Point of non-correlation indicates an event took place during this month. Investigation shows that the event was a shutdown of one of the satellites. Events can be identified without a-priori knowledge of the system!

In order to explore Magnetic Flux impact further, we plotted Total Magnetic Field Sum vs. Transmission Losses. The graph shows a moderate correlation (0.48) with an exception for an event in December, 2007 when one of the two global satellites (GEOS12) was down. Continuation of analysis would use such out of family data points to infer a lack of coverage by the satellite constellation, without a-priori knowledge of the mission. In this paper, we have barely scratched the surface on what is possible in the area of event identification for government technology with weather space data: we expect more event correlations to become clear upon further effort.

As an alternate view of the impact of space weather data, we therefore also reviewed Total Magnetic Field Mean with Transmission Losses. This provided a slightly higher correlation (0.51) and with no real gaps in December 2007, which is expected as we now looked at the mean value of Magnetic Flux. As can be seen from the graph, the two curves are aligned well during summer months as was the case with Land Surface Average Temperature as well. Meanwhile, during winter months the Surface Temperature Average vs. Losses are not correlated. This ties back to the dominant effect that Total Generation has on the Transmission Losses.

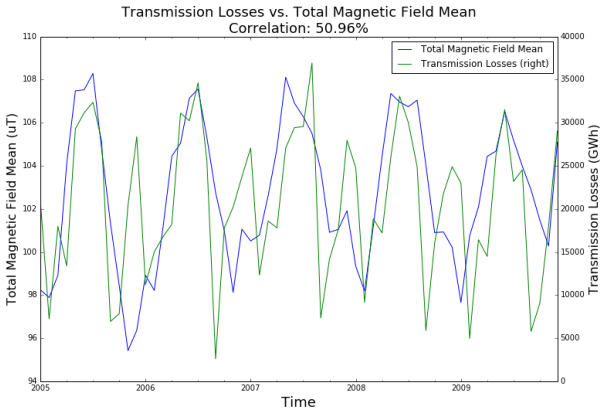


Figure 7 Transmission Losses vs. Total Magnetic Field Mean.

Conclusions

The primary conclusion for this study is the strong linear regression correlation (~0.6) between transmission loss and total generation. Therefore, all impacts that increase total generation (running A/C or heaters, high business use, etc.) will impact transmission losses the most.

After removal of generation as a variable for transmission loss, we were able to find that the linear regression correlation for Space Weather (0.30) was actually higher than Average Land Surface Temperature (-0.04) in direct opposition to our initial hypothesis.

As it turns out, temperature may be the most dominant variable only in that it drives more power generation and therefore more losses. Magnetic flux impact from Space Weather is actually more impactful than thermodynamic equilibrium loss.

Additionally, this project has taught us a lot about the challenges of finding, cleaning, and combining different datasets. We began this project thinking that the data would be readily available for analysis, and in the end we spent nearly 2/3 of our available time getting the data into workable formats. This was a limitation of our project (less time for analysis), but still a valuable lesson and an opportunity to practice many of the less-glamourous skills needed in data science.

Source Data

Magnetic Flux data for space weather (μT):

- http://satdat.ngdc.noaa.gov/sem/goes/data/new_avg
    - Sample file: ./2011/08/goes15/csv/g15_magneto_1m_20110801_20110831.csv

Land surface weather:

- Monthly temperature readings from NOAA for contiguous United States (state code 110): http://www7.ncdc.noaa.gov/CDO/CDODivisionalSelect.jsp

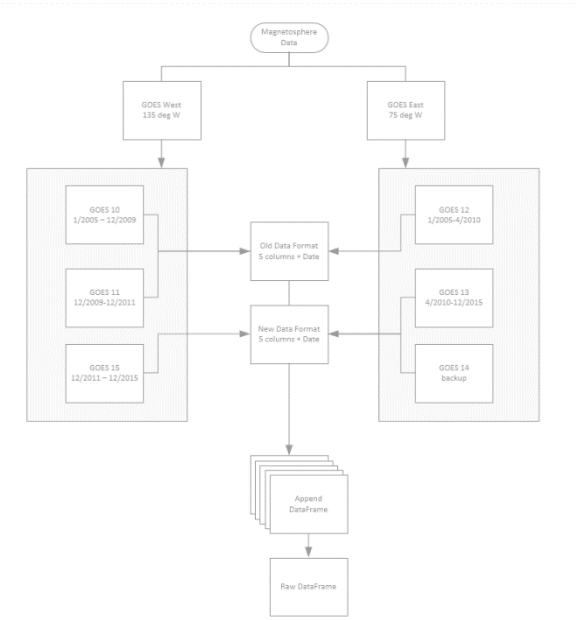
Electricity:

- Energy Information Administration:
    - API: http://www.eia.gov/opendata/qb.cfm?category=711279
- Electricity Profile report, table 10: http://www.eia.gov/electricity/state/unitedsta

Group Analysis:

- Narrative analysis, laboratory notebooks and CSV files can be found here: https://drive.google.com/folderview?id=0BxolO_eh8I-8NTUxUWVTY1Fha0E&usp=sharing

Appendix A: Data Population Methods



Data was pulled from the US Energy Information Administration (EIA) and the National Oceanic and Atmospheric Administration (NOAA), representing nearly 5 GBytes of information from > 600 individual files in 5 different formats. Our first challenge was to populate a single dataset for each system. Figure 2 illustrates the complex architecture involved with magnetosphere data population. We accomplished this by writing loops to pull all data from the date range we selected and format / save the files as a CSV on our google drive.

Step two derived new columns to allow for better data management, such as Satellite ID, Transmission Loss and custom helper columns used to assist with later manipulation.

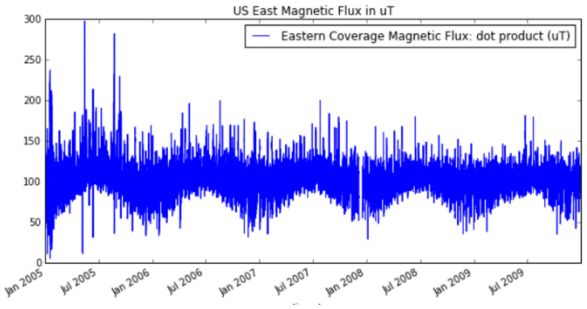


Figure 9 Initial plot of raw magnetic flux data represents ~4 million datapoints - Data reduction was required

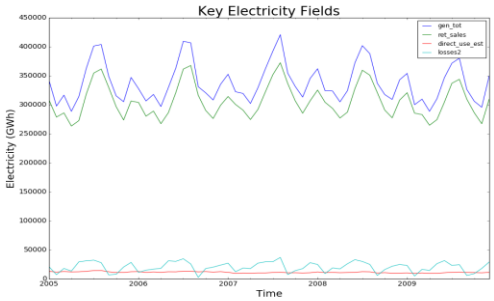


Figure 8 Initial plot of raw power data

Step three identified areas of overlap between the three DataFrames. As it turned out, the frequency of electricity data publically available was not as high as we had anticipated when entering this project. Data is only publically available on a monthly scale for power and surface temperature, whereas magnetosphere data is available minutely. Space data needed to be simplified by deriving monthly data points for each, also minimizing the memory required to manipulate and these DataFrames and reducing the data size from ~5GB to ~85KB: a **99.8% reduction of size** for the resulting DataFrame and CSV, allowing for quick loading by outside users and eliminating significant time delay.