Attributing electricity demand anomalies in the US to major events

Ian Wright, MSc Candidate, NYU CUSP

Abstract— **A novel methodology was developed to identify localized spatio-temporal events in which electricity demand from the grid decouples from its primary dependent factors, time and weather. The *decoupling events* are aligned (in space and time) with news media event data recorded in the Global Database of Events, Language, and Tone (GDELT), in an attempt to gain insight about significant events that influence power consumption in the US. Unfortunately, weather data lacked special granularity and event data lacked richness, leading to non-interpretable results. The methodology is promising, if performed with improved datasets.**

I. Introduction

The US power grid is a massively complex network of coal and gas plants, hydropower turbines, wind and solar farms, wholesale electricity buyers, retailers, transmission and distribution lines, and consumers. Countless factors impact the daily ebb and flow of demand for electricity: hot days and heavy air conditioner usage, cold days and space heater usage, and mainstream media events that draw millions of viewers to their TVs and computers, are just a few examples. But for the US grid to operate at peak efficiency, with minimal outages and power loss, it's critical that the demand for electricity be exactly matched with incoming supply, every minute of the day, on every wire in the country. A network of approximately 80 independent entities throughout the US called "balancing authorities" (BA's) are responsible for doing just that. Every day, BA's forecast exactly how much electricity will be consumed within their territory, and coordinate with local power producers to ensure that demand will be met with supply throughout the day. It's generally assumed that power demand is a function of weather and time; there are consistent daily, weekly, and seasonal trends in demand, in addition to the close relationship between demand and outdoor temperature (an unseasonably hot day sees higher demand levels).

II. Goal

The objective of this analysis, and others in the realm of exploring real-time electricity demand, is simply to gain a more nuanced understanding of the drivers of demand. In the case of a balancing authority forecasting its regional daily demand, accuracy is closely tied to efficiency. If actual demand is *lower than* forecasted, the wholesale market is contractually obligated to overpay for its power, and the excess costs are ultimately passed on to consumers as higher electricity bills. If actual demand is *higher* than forecasted, there is a power shortage in the system, and the BA is forced to turn on inefficient 'peaker plants' that emit more carbon into the atmosphere. The need for extremely accurate demand forecasts is what inspired this particular analysis. Leveraging 'big data' techniques to uncover significant trends or drivers that could be used to tune future forecasting models may eventually support a more efficient US grid. Specifically, the analysis seeks to isolate demand-effects that are independent of time and weather.

III. Hypothesis

Currently, BA demand forecasts depend primarily on time and weather. The time factor can be decomposed into a seasonal (annual) pattern, a weekly pattern that reflects a standard work week, and a daily pattern that reflects the typical waking hours for an average American. While there is an implicit weather factor that underlies the regular seasonal patterns, there is also a more volatile day-to-day weather-dependent factor. If a given summer day is significantly warmer than seasonal norms, one would expect electricity demand to also be higher than normal, to meet cooling loads.

Under the current paradigm, if the annual, weekly, and daily demand periodicity can be *modeled and removed* from demand time-series data, *and* the data is further normalized for day-to-day fluctuations in local temperature, the resultant time series should be relatively flat and featureless. That is, all dependence on time and weather is removed from the data.

If, however, this normalization process yields any significant residual structure in the demand data, the result may imply that there are non-weather- and non-time-related factors influencing electricity demand ("decoupling"). The hypothesis underlying this analysis is that significant "decoupling events" *have and do occur*, and they can be better understood by spatio-temporal matching with media news records, in an effort to study the real world events that coincided with demand decoupling events. The ideal output of the analysis would be to determine a set of *event types* that are likely to increase electricity demand, perhaps with regional dependence.

IV. Datasets

Several datasets were integrated for the analysis:

Hourly sampled temperature readings from National Oceanic and Atmospheric Administration (NOAA) weather stations throughout the country; historical (via web API)

~700 stations queried @ ~20,000 records per station → 14M records

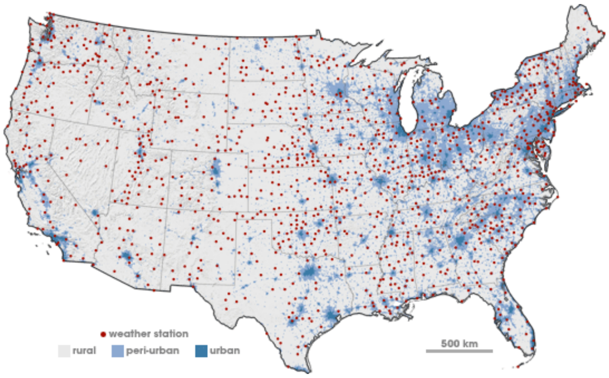


Figure 1: NOAA weather stations

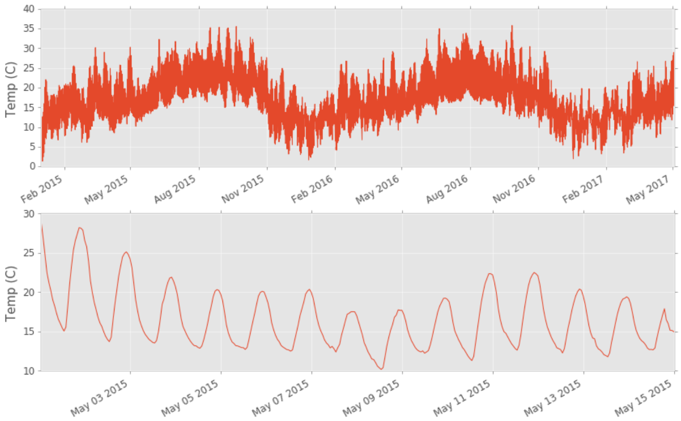


Figure 2: Weather data for California. The seasonal trend is seen clearly in the upper plot, while the weekly pattern can be seen at a closer scale in the lower plot.

Hourly electricity demand from the Energy Information Administration (EIA) for each balancing authority, and at an aggregated regional level; historical (via web API)

13 regions @ ~20,000 records per station → 260K records

EIA regional shapefiles, upon request from EIA

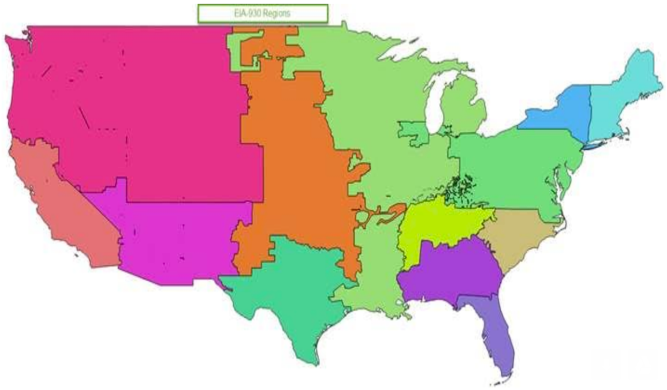


Figure 3: EIA Regions

List of all US cities with population 50K and above (US census data), and cities' associated geographic coordinates (via GoogleMaps API)

News media event data recorded in the Global Database of Events, Language, and Tone (GDELT), and categorized with respect to actors, actions, and tone (via Google BigQuery)

Unprocessed GDELT data ~25M records → 4.2 GB

V. Methodology

All relevant code for the analysis can be found in the Github repo: *https://github.com/ian-wright/demand*

At a high level, the analysis required the integration of several disparate data sources into a single pipeline, which was heavily processed in preparation for final computation on a distributed cluster (using PySpark).

Data Pipeline

Hourly electricity demand for each of 13 EIA regions in the US was pulled, via a web API. Data spans from July 2015 to present.

A list of US cities greater than 50K in population is collected from US census data. GoogleMaps API is used to associate each city with a geographical coordinate.

The NOAA API is used to locate a single weather station near each city, and the station itself is checked to ensure that complete time series data exists. The city and station are paired, and hourly weather data is pulled from the NOAA API and associated with the city.

Using a series of spatial joins, US cities are located within each EIA region, resulting in a grouping of weather time series data linked to each region.

A single *regional* weather time series is generated by computing the population-weighted average of all city-associated weather data for the region's contained cities. The underlying assumption here is that power consumption is linearly related to population, and a population-weighted regional weather average serves as a reasonable estimate for the whole region's weather patterns, despite its geographical size.

A simple method was employed to normalize demand data against day-to-day fluctuations in weather. Weather data was resampled from hourly to daily granularity, and the highest and lowest temperatures for each day were recorded. A rolling 14-day average was applied to the highs and lows, setting two important baselines: average highs for warm months, and average lows for cold months. In warm months, observed daily highs that exceed the moving average called for a proportional *downward adjustment* in the same day's electricity demand. Similarly, in cold months, observed lows that were below average lows called for a proportional *downward adjustment* in the same day's electricity demand.

Next, Facebook's *Prophet* time-series forecasting model was employed to remove trends and periodicity from the weather-normalized demand series. *Prophet* is an additive time series modelling approach that models a non-linear growth component, and a seasonal trend at multiple frequencies (if necessary) by way of Fourier transforms. The tool is often used to forecast future time series, but it can also be used to simply model the observed structures within existing observations. For each weather-normalized regional demand series, a trend term, weekly cycle term, and annual cycle term were modelled, and all three were removed from the data, leaving a relatively "flat" time series centered on zero.

To identify significant "decoupling events", in which spikes in power demand persist despite the correction for time and weather, any daily observations that exceeded a two-sigma (upwards) deviation from the series' mean were flagged. A final set of tuple anomalies was generated in the form *[(region, day), (region day), …]*

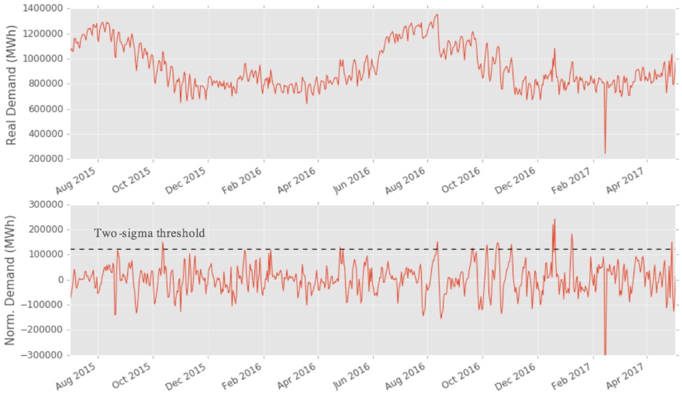


Figure 4. Raw demand data for Texas region in top plot. Post-normalization (weather and seasonality) shown in lower plot, with a 2-sigma threshold line that serves to flag anomalous structures.

Distributed computing

Google's BigQuery platform was used to collect news media event records (GDELT) corresponding to any events that happened within the US and on the days flagged as anomalies. GDELT is an ongoing project, with contributions from Yahoo and Google engineers, that attempts to store *all* news stories from across the world. The data is often used to study international geopolitical trends. BigQuery is a partitioned relational data warehouse hosted on distributed Google infrastructure. The platform is able to perform complex SQL queries on massive datasets in very little time. Because BigQuery (and all standard SQL engines) isn't capable of spatial queries, *all* event data from throughout the US on any flagged days was returned. GDELT does include geographical coordinates for each event, which were included in the output. The results were sharded into five data files, and loaded into NYU's HDFS storage (4.2GB uncompressed).

PySpark was employed for two primary operations. The first was a filtering procedure that returned only those GDELT records that occurred within a given region, on one of the "anomaly days" flagged for that region. Python's *shapely* package was useful for embedding the spatial comparison within a PySpark mapping function. The function reduced ~25M records to ~5% of original size.

Finally, another PySpark operation was designed to compute aggregate statistics for all GDELT events, grouped by region. This approach was used to uncover any regional differences in electricity demand dependence on events. Summaries for each region were written to file. All analysis code, including the script used with PySpark, can be found on the aforementioned Github repo.

As a unified PySpark script, runtime was ~18 minutes on NYU's Dumbo cluster. Processing this volume of data on a local workstation would likely take far longer.

VI. Major Challenges

Three major challenges had a significant negative impact on the outcome of the analysis:

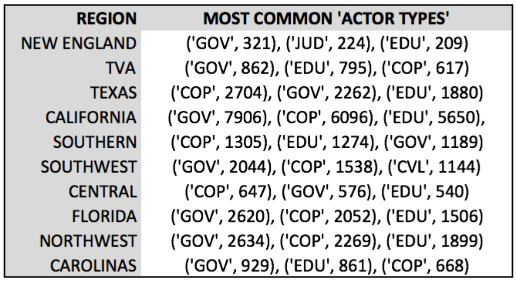
Originally, the intent was to perform the analysis for each if the US balancing authorities (~80 in total). In reality, BAs' boundaries are not fixed and concrete, by dynamic and imperfect. Power is regularly transferred across boundaries, and the boundaries themselves are not defined exactly in space. For this reason, the analysis had to be performed an aggregate level – thirteen larger regions across the country. This meant that demand anomalies couldn't be isolated precisely; for example, we may know there was an anomaly on *April 4, 2015, somewhere in the 'Midwest'*. The lack of precision makes results much less impactful.

Next, the trivial method used to normalize for weather fluctuations was not robust. A more robust method involves comparing observed actuals against *historical averages*, but in this analysis observed actuals were compared against the rolling average of themselves. The chosen method is sensitive to one-time discontinuities in temperature, but loses sensitivity in the case of long hot or cold periods.

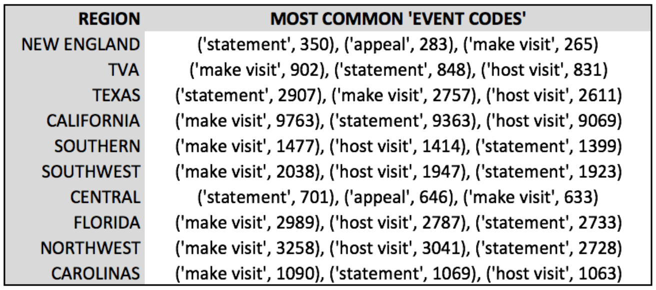
Finally, and most importantly, the richness of GDELT event data was simply not adequate for the analysis. GDELT data contains very abstracted actor codes (eg. 'gov', 'edu', 'cop', ect), and even more abstract action codes (eg. 'make visit', 'host visit', 'make appeal', ect), leaving a lot to be desired in order to understand the true nature of the news stories. It seems that this data is appropriate for studying large scale trends in the interactions of large bodies (countries or states), but comes up short for event-level analysis.

VII. Results

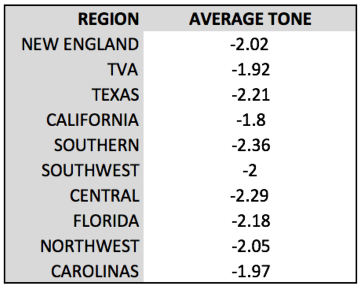
It's extremely difficult to glean any meaningful insights from the results of the analysis (primarily for the aforementioned lack of richness in GDELT data). *Table 1* shows the most common actor types for all GDELT events in anomalous place-time pairs. Government, police, judiciary, and educational entities appear in all regions. *Table 2* shows the most common event types, but the results are hard to interpret. Visits, appeals, and statements reappear frequently. Finally, the average tone from GDELT news articles (where positive numbers represent overall positive tone) are shown in each region in *Table 3*. While we don't have a baseline to compare against in this case, it's interesting that tone is *negative* for all regions. Of course, this may be a reflection of most news stories in the US.



REGION	MOST COMMON 'ACTOR TYPES'
NEW ENGLAND	('GOV', 321), ('JUD', 224), ('EDU', 209)
TVA	('GOV', 862), ('EDU', 795), ('COP', 617)
TEXAS	('COP', 2704), ('GOV', 2262), ('EDU', 1880)
CALIFORNIA	('GOV', 7906), ('COP', 6096), ('EDU', 5650),
SOUTHERN	('COP', 1305), ('EDU', 1274), ('GOV', 1189)
SOUTHWEST	('GOV', 2044), ('COP', 1538), ('CVL', 1144)
CENTRAL	('COP', 647), ('GOV', 576), ('EDU', 540)
FLORIDA	('GOV', 2620), ('COP', 2052), ('EDU', 1506)
NORTHWEST	('GOV', 2634), ('COP', 2269), ('EDU', 1899)
CAROLINAS	('GOV', 929), ('EDU', 861), ('COP', 668)



REGION	MOST COMMON 'EVENT CODES'
NEW ENGLAND	('statement', 350), ('appeal', 283), ('make visit', 265)
TVA	('make visit', 902), ('statement', 848), ('host visit', 831)
TEXAS	('statement', 2907), ('make visit', 2757), ('host visit', 2611)
CALIFORNIA	('make visit', 9763), ('statement', 9363), ('host visit', 9069)
SOUTHERN	('make visit', 1477), ('host visit', 1414), ('statement', 1399)
SOUTHWEST	('make visit', 2038), ('host visit', 1947), ('statement', 1923)
CENTRAL	('statement', 701), ('appeal', 646), ('make visit', 633)
FLORIDA	('make visit', 2989), ('host visit', 2787), ('statement', 2733)
NORTHWEST	('make visit', 3258), ('host visit', 3041), ('statement', 2728)
CAROLINAS	('make visit', 1090), ('statement', 1069), ('host visit', 1063)



REGION	AVERAGE TONE
NEW ENGLAND	-2.02
TVA	-1.92
TEXAS	-2.21
CALIFORNIA	-1.8
SOUTHERN	-2.36
SOUTHWEST	-2
CENTRAL	-2.29
FLORIDA	-2.18
NORTHWEST	-2.05
CAROLINAS	-1.97

Tables 1, 2, 3: Aggregate statistics for regional GDELT data

VIII. Conclusion

The primary achievement of the analysis was the development of a novel approach for identifying patterns in electricity demand that decouple from both weather and time. Unfortunately, weak datasets meant that the approach yielded few interesting results. With more granular electricity demand data, and richer event data, the analysis method certainly has potential to produce meaningful results.

References

Facebook Prophet:
https://facebookincubator.github.io/prophet/static/prophet_paper_20170113.pdf

Google BigQuery:
https://cloud.google.com/bigquery/docs/concepts

EIA Realtime Data:
https://www.eia.gov/beta/realtime_grid/#/summary/about?end=20170502&start=20170404