Relationships Between Weather Patterns and Power Commodity Markets

University of Michigan School of Information

May 24, 2021

Keith Bourne & Michael McManus

# Motivation

Commodity energy prices can see dramatic swings from time to time, causing significant financial and even life-threatening harm to millions of power customers while devastating power provider budgets and businesses. We saw this havoc in February of 2021 when Texas was put into a state of energy crisis due to winter storms hitting the area. These storms caused dramatic energy pricing swings and electricity power generation failure that even resulted in potable water, food, and heat shortages, leaving 4.5 million homes and business without power, and 151 people dead[[1]](#endnote-1). We look at the Michigan energy markets, seeking to gain a better understanding of the links between weather and commodity energy pricing and how those linkages can help energy markets be more efficient and better prepared for future potential crises.

HYPOTHESIS:

With day ahead locational marginal pricing (LMP) in Michigan showing much larger fluctuation in pricing than real time pricing, a better understanding of what causes these pricing differences can help market participants make better decisions about what to bid on energy in the MISO marketplace and improve the efficiencies of these marketplace. As an initial step in improving this understanding, we hypothesize that most significant changes in the day ahead locational marginal pricing (LMP) for energy in Michigan (> 20% change in a day) are due to major decreases in temperature.

# Data Sources

## Primary Dataset - MISO Locational Marginal Price

The primary data source used in this analysis was Locational Marginal Prices (LMP) in Michigan and Louisiana. In short, LMPs are the cost of electricity for the given region. The Historical LMP data was downloaded from the [MISO website](https://www.misoenergy.org/markets-and-operations/real-time--market-data/market-reports/#nt=%2FMarketReportType%3AHistorical%20LMP&t=10&p=0&s=MarketReportPublished&sd=desc) in CSV file format. The variables included for analysis in this dataset were date/time, real-time LMP, day ahead LMP for Michigan Hub and Louisiana Hub for the years 2013-2020. During our analysis, we decided to Load Forecast and Actual Load data to better understand the possible effect of weather on electric prices. In short, Load Forecast and Actual Load is the measure of demand on the MISO grid. The load data was also downloaded from the same [MISO website](https://www.misoenergy.org/markets-and-operations/real-time--market-data/market-reports/#nt=%2FMarketReportType%3ASummary%2FMarketReportName%3AHistorical%20Daily%20Forecast%20and%20Actual%20Load%20by%20Local%20Resource%20Zone%20(xls)&t=10&p=0&s=MarketReportPublished&sd=desc) data source as our primary dataset, in CSV format hourly for the years 2013-2020.

Once all aggregation and merging of datasets was complete, our final dataset included 65,038 records and a total of 62,851 kilobytes. To facilitate reproducibility and further analysis, both our raw and aggregated datasets are available on [GitHub](https://github.com/kbourne/MADSmilestone1data).

## Secondary Dataset - Copernicus Weather API

We used [Copernicus Climate Data Store](https://confluence.ecmwf.int/display/CKB/Climate+Data+Store+%28CDS%29+documentation) to gather historical weather data for the MISO footprint, specifically Shreveport, Lake Charles, and New Orleans in Louisiana and Detroit, Grand Rapids, and Lansing in Michigan.  The files returned by the Copernicus Weather API were of NC file format.  The variables we gathered from the API for use in this analysis were date/time, wind speed, surface pressure, precipitation, and temperature.  We gathered hourly data from the three cities in each of the two states for the years 2013-2020, resulting in around 420,000 records total.

To access the weather data, you are required to create a free user account at Copernicus, but once that is complete, you can access the API with the key you will be provided with your free account.  They provide a tool to help you build an API request for the specific data you are seeking here: <https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=form>

Once you have you API request formulated, we suggest using a python script to pull the data. Please note, due to the API data limits, we had to make 24 separate requests that took 1-2 hours to complete. You also have to store the credentials file provided to access the API in the same directory as the python script. The credentials are in the form of a system file (.cdsapirc), so you will need server permissions to store that, which did not work with University of Michigan or Colab environments.

You need to install two additional packages in your local environment: cdsapi (for accessing the Copernicus data store) and netCDF4 (provides utilities for extracting data from the NC files). Here is a simple example of a function to pull weather data:

REFERENCE get\_weather\_data() FUNCTION & A TYPICAL CALL TO THE FUNCTION

## Full Dataset

Once all aggregation and merging of datasets was complete, our final dataset included 65,038 records and a total of 62,851 kilobytes. To facilitate reproducibility and further analysis, both our raw and aggregated datasets are available on [GitHub](https://github.com/kbourne/MADSmilestone1data).

# Data Manipulation Methods

## Copernicus Weather Data

Upon retrieval of the weather data, scripts were used to convert the NC file format to CSV file format.

REFERENCE read\_netcdf\_file()

In addition to this transformation, several of the weather attributes in the dataset required manipulation and/or unit conversion. For example, wind speed was formulaically derived from the u and v components of wind provided by the API.

Additionally, weather measurement units were converted to “local” weather related scales, such as converting the units provided in Kelvin to Fahrenheit and meters per second measurements to miles per hour. Finally, as the weather API had call limits and only two years of data could be pulled at a time, we created scripts to aggregate the weather data for the years 2013-2020 running constantly over the course of several days.

## Locational Marginal Price

As LMP data could only be pulled in time increments, the downloaded CSV files were concatenated using a script which pulled each file in, converted it to a Pandas data frame, and concatenated it to an existing data frame iteratively.

REFERENCE SCRIPT FOR concat of pd

One issue we encountered while concatenating the LMP data was a change in format of column headers. In older data the hour ending columns were formatted as “HE 01” and newer data column headers had a naming convention similar to “HE01”. To alleviate foreseeable issues with concatenation on dissimilar column names, the script changed all column names to a common format prior to concatenation. Once all LMP data was concatenated over the years 2013-2020, the data frame was manipulated from wide format to long format using the below function.

REFERENCE transfor\_lmp\_csv function

In short, the function takes in a file path removes unneeded columns. The index in the LMP data set was a date format while the hours of each date were listed column-wise. To create a data frame which could be merged with the weather data, the data frame was melted into long format. ‘HE’ was removed from the hourly column names and converted to time delta data types to allow them to be merged with the date column to provide a datetime format as 2013-10-01 01:00:00. The function then shifts the ‘time’ column to the left side of the data frame to allow for merging with the weather data set. This function was then used in a script to pull all LMP data from a local folder and concatenated into one large data set.

## Missing Data

We originally sought to show data back to the origination time of the MISO data, which is in the mid-2000’s. However, upon further analysis, we found that each state started at different times, with Louisiana starting at the most recent time period (late 2013). Due to this limitation, we reduced all of our datasets to align with that same time frame.

## Conversion Process

We utilized several common data conversion and processing steps during the analysis. These included:

* Converting data/time columns – the two datasets had different date/time formats that had to be converted. We used the Pandas to\_datetime method to set the weather time column to the same format as the LMP pricing time column.
* Creation of categorical data, temperature categories, season categories – To do the final analysis and gain further insight into when pricing and weather showed the most insightful interactions, we used categorical temperature data (i.e. below freezing, room temperature, warmer than room temperature) and season categories (i.e. Winter, Spring, Summer, Fall) to observe the SPLOM visuals in distinct categories. An example of what we used to create this additional field with categorical data can be found:
  + REFERENCE TEMPERATURE CATEGORY CODE

## Steps to Join Data Sets

We followed these steps to transform the data from raw (NC and CSV files) to the final format for analysis:

1. Pull weather data down in 2 year increments, for 6 different locations across 2 different states, in .nc file format
2. Pull power pricing data down for 2 different states in csv file format
3. Write/execute script to extract data from the nc weather files, including significant manipulation per field with list comprehensions
4. Write/execute scripts to combine all the weather data for each state into Michigan and Louisiana weather csv files
5. Convert date/time columns for weather to match power files
6. Combine weather and power data joining by date/hour - Once the weather and LMP data were in their proper formats, all data was merged into one data set on the ’time’ column and written out to a CSV file. Having all data in one location allowed for simpler analysis.

REFERENCE FINAL FILE CREATION SCRIPT

## Data Analysis Workflow

Our data analysis started with pulling down the data from the APIs described above. We then performed the steps described above to consolidate the data into one csv that enabled a more efficient analysis during the analysis stage. At this point, we performed a couple round of exploratory data analysis (EDA), looking for missing data, exploring structure of data, and looking for general aspects of data that are more obvious with some basic EDA functions (such as .describe and .isna().sum). This step led to some insights, such as needing additional wind speed data and the addition of the load data to the energy data set.

Once the new data was collected, an additional round of EDA was performed, focused on building off the basic EDA, starting to run more visualizations that showed correlations among the data points. The goal at this stage was to explore what we may need to use for the final analysis and finalize the specific ways we are going to analyze the data. For example, this stage of the analysis showed that the large energy pricing swings were more reflected in the day ahead energy data, rather than the real-time energy data, and this helped guide the final analysis.

Finally, an analysis was performed to focus on answering the key questions we proposed, what causes energy pricing differences that can help market participants make better decisions about what to bid on energy in the MISO marketplace.

## Challenges and Solving Them

We encountered many challenges during the analysis. For example, the weather data source had a data cap for downloading, and their server bandwidth was significantly throttled, causing us to have to download 2 years of data per location at a time. With there being 6 different locations, this was 12 downloads taking that took several days to download, and then this data had to be recombined to represent the final data set. After the initial download, we found that there were actual multiple wind variables that needed to be downloaded to get the final wind speed data, which caused us to have to repeat the entire data download process a second time.

Due to time and project size constraints, we also had to cut down the scope of our analysis from looking at both Louisiana and Michigan (to draw comparisons in different regions of the country for what we were analyzing) to looking at just Michigan.

## Bringing it all together

Once the weather and LMP data were in their proper formats, all data was merged into one data set on the ’time’ column and written out to a CSV file. Having all data in one location allowed for simpler analysis.

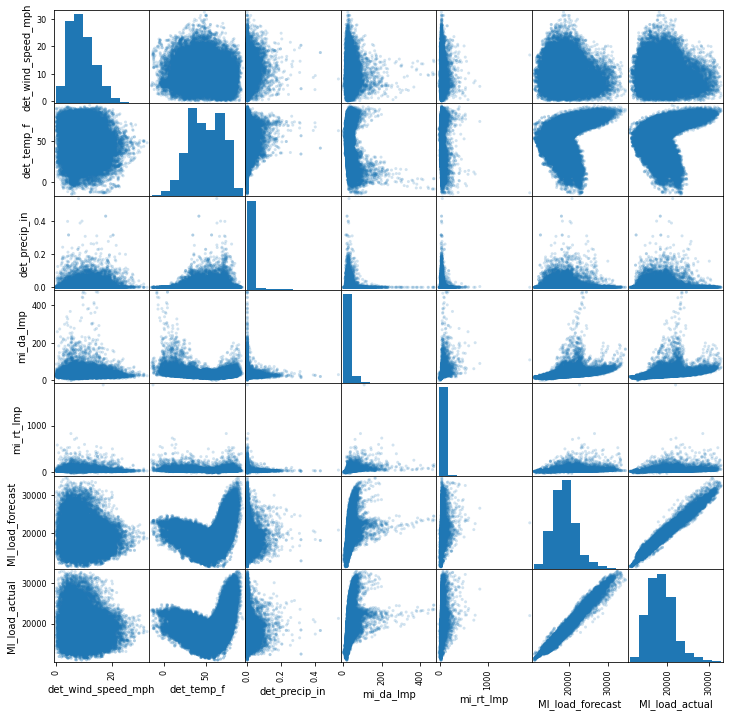
REFERENCE FINAL FILE CREATION SCRIPT

# Analysis and Visualization

## Analysis Steps Performed on Datasets

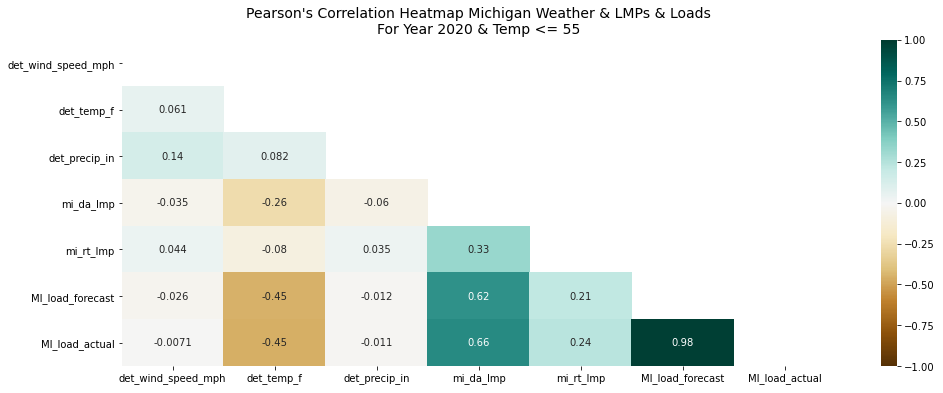
Correlation between various weather data points and energy pricing and loads was a key focus of the analysis. Initially we looked at the actual text-based correlations combined with a heatmap to quickly visualize where the correlations occur and how relatively strong those relationships are. We also utilized multiple scatterplot matrices (SPLOMs) to help us visualize these data interactions, as they are widely considered an effective tool for this specific type of application. We chose Seaborn as our visualization tool to display the SPLOMs, as it is known as a useful tool for this particular type of visualization.

We began our analysis by examining the correlation between temperature and LMP prices. Intuitively, and through some prior domain knowledge, we believed there would be a strong correlation between temperature and LMP prices due to seasonal heating and cooling load on the grid. The correlation we measured between temperature and LMPs in the real time market was -0.141365. This coefficient seemed strikingly low to us which prompted us to consider other confounding. It was at this point we decided to incorporate load data to our analysis. In looking at a SPLOM of our dataset, a pattern emerged between the temperature and load data.

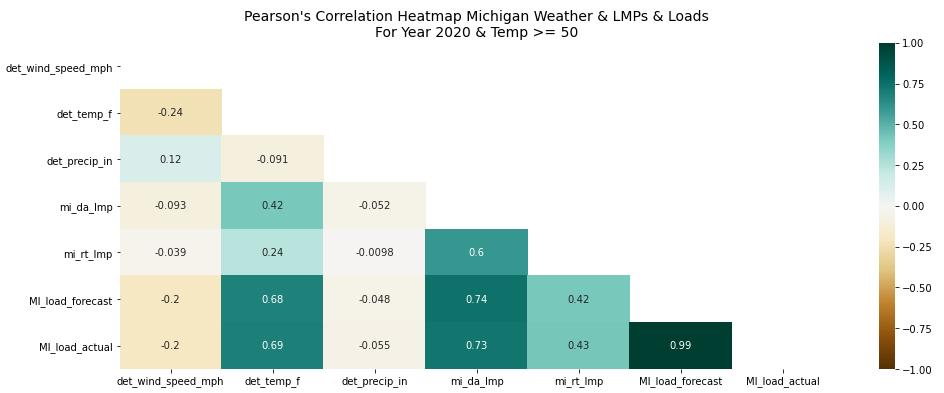


Through this visualization technique, we quickly recognized the relationship between temperature and load. However, this chart was still not detailed enough to give deeper insights into what aspects of temperature had the most influence on major pricing fluctuations. Was there a temperature threshold where pricing changed significantly? What happened to the other variables during the same time that temperatures were having an impact on the pricing? Did the expectation for energy load have an impact on the pricing and was there a strong correlation between the expected temperature and the expected load demand?

Looking further into this correlation, we found that when temperatures are at or below 50 degrees Fahrenheit, there is a -0.45-correlation coefficient between temperature and actual load.



Similarly, when we looked at temperatures at or above 50 degrees Fahrenheit, we found a correlation of 0.69 correlation coefficient between temperature and actual load.



With this newfound insight we began exploring seasonal patterns in the data by encoding temperature categories with color on our SPLOM visualization. Utilizing the script shown below, we categorized temperature data into their corresponding bins.

## Summary of Results

Usiing the insights conducted so far, we chose to chart the correlations between the weather data points and the energy pricing, but with categories that indicate specific ranges of temperature that seemed to have the most impact on pricing.

***Temperature categories:***

REFERENCE PAIRPLOT SCRIPT

Temperatures below 0 degrees Fahrenheit were encoded red, temperatures between 34-65 degrees were encoded green, and temperatures above 65 degrees were encoded blue.

Scatter chart

Description automatically generated

For day ahead pricing, there is clearly a break visually that indicates the impact lower temperatures have on the pricing fluctuations that did not stand out when you only looked at temperatures overall. You can also see how the high surface pressure and price seem to be correlated. For example, all the high pricing occurred when the temperature was below freezing, and the precipitation was very low. Wind speed combined with low temp seems to push the price up as well, but that is more spread out than the other factors. And then much of the high prices seem to show up when the surface pressure was in a specific range.

Seasonal Categories

Chart, scatter chart

Description automatically generated

In addition to the temperature range based categories, we look at categories from a seasonal perspective. As might be expected, the colder temperatures that occur in the winter are highly consistent with the temperature categories charted above. Similar conclusions can be drawn, but with a seasonal perspective, such as the low temperature, low precipitation, and higher surface pressures of the winter months are where you see the larger price fluctuation.

Given these results, we see a clear relationship between times when temperatures are below zero, during the dryer, higher surface pressure winter months and higher energy pricing. \_\_\_show the correlation data for these relations??\_\_\_

The analysis showed that day ahead locational marginal pricing (LMP) in Michigan shows a much larger fluctuation in pricing than real time pricing. We showed that temperature is a key data point related to these fluctuations, showing high correlation, and particularly within the below freezing temperature range. This conclusion can help market participants to use temperature forecasts to help them make better decisions about what to bid on energy in the MISO. Our hypothesis that the most significant changes in the day ahead locational marginal pricing (LMP) for energy in Michigan (> 20% change in a day) are due to major decreases in temperature was proven to be positive.

## Suggestions for Further Analysis

While these results show strong correlations between certain temperature ranges and energy pricing, it does not necessary indicate if this is the main variable causing this, or if there are other confounding variables that may be influencing the energy pricing. Further analysis of new data sources, such when there are pipeline breakages or congestion. There are a number of other aspects that may or may not have an impact on energy pricing that we suggest to explore.

# Statement of Work

Keith Bourne and Michael McManus each collected one of the datasets initially. Once it became apparent the wind data needed to be expanded, they split re-collection of the weather data that second time. Both team members performed their own exploratory data analysis (EDA) and then held discussions regularly to determine best next steps based on their findings. Both then worked together to consolidate their findings into a final report format that provided the most relevant of their efforts to the final project report.

1. “2021 Texas Power Crisis.” Wikipedia. Wikimedia Foundation, May 13, 2021. <https://en.wikipedia.org/wiki/2021_Texas_power_crisis>. [↑](#endnote-ref-1)