Relationships Between Weather Patterns and Power Commodity Markets

University of Michigan School of Information

May 24, 2021

Keith Bourne & Michael McManus

1 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[i]. 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 the market. As an initial step toward 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 below freezing.

2 DATA SOURCES

2.1 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 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 discovered it was necessary to include 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 energy demand on the MISO grid. The load data was also downloaded from the same MISO website 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.

2.2 SECONDARY DATASET - COPERNICUS WEATHER API

As our secondary dataset, we used Copernicus Climate Data Store 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 total records.

To access the weather data, you are required to create a free user account at Copernicus. Once a user account is created, the user can access the API with a key provided with the free account. The Copernicus team provides a tool on their website to help users build an API request for the specific data they are seeking.

Once an API request is 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. Users may also have to store their credentials file which is provided to access the API in the same directory as the python script. These credentials are provided in the form of a system file (.cdsapirc), which require server permissions to store, and which will not work with University of Michigan or Colab environments.

Users will need to install two additional packages on their local environment: cdsapi (for accessing the Copernicus data store) and python netCDF4 library (provides utilities for extracting data from the NC files). Here you can find a simple example of a function to pull weather data.

2.3 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.

3 DATA MANIPULATION METHODS

3.1 COPERNICUS WEATHER DATA

Upon retrieval of the weather data, scripts were used to convert the NC file format to CSV file format. 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 continuously over the course of several days.

3.2 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.

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 a function. In short, this function takes in a file path, converts the CSV file to a pandas dataframe, and removes unneeded columns. The index in the downloaded LMP dataset was in date format similar to `2013-10-01`, 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 dataset.

3.3 MISSING DATA

Our initial intent at the onset of this analysis was to include data ranging from the induction date of MISO in 1998 to present. Upon further research we discovered MISO market participants joined the ISO at different stages of its lifespan. Louisiana, being one of the states we chose to examine in our analysis, joined the ISO in December of 2013. Due to this limitation, we reduced all of our datasets to align with that same time frame.

Another instance of missing data was in our historical LMP dataset. The LMP record for October 14, 2020 was missing from the MISO historical dataset. We used different approaches to alleviate this issue. When looking more granularly at the data, we simply used interpolation using the mean LMP of the month. In other instances, resampling the data by month or week on the mean also alleviated negative effects of this record.

3.4 Conversion Process

We utilized several common data conversion and processing steps during our analysis. In order to merge our data successfully, we converted date/time columns to datetime objects using Pandas `to_datetime` function. As the LMP datasets utilized wide format for dates and times, these were changed to long format concatenating hours on the date and then converted to datetime objects.

To gain further insight into the relation between weather and electric pricing, we categorized temperature in two ways. Firstly, we categorized temperatures as 'below freezing', 'between 34 & 70', and 'above 70'. We used the Numpy select method to apply a list of conditions and their respective value names to the temperature, effectively turning temperature into a categorical variable that can utilize color as a secondary visual way to display the differences in the correlations between specific categories of temperature ranges and energy pricing. These categories were encoded with color on our SPLOM representation of the dataset, allowing the SPLOM to display additional information about what ranges of the temperature the correlations were most prominent, rather than just displaying the correlations in general.

We applied this same categorization technique and analysis using seasons as well (winter, spring, summer, and fall). In this case though, we wrote an additional function that turned date stamps into a season category. Once all temperatures were categorized, we applied the same secondary visual of color by season to the SPLOM charts for added visualization benefit.

3.5 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.

3.6 Data Analysis Workflow

Our data analysis started with pulling down 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 loading during the analysis stage. At this point, we performed a couple rounds 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 reflected more in the day ahead energy data rather than the real-time energy data which 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.

3.7 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. Given that we pulled data for 6 locations over 8 years in two-year increments, data was downloaded in 16 chunks which took several days to complete. The data then had to be recombined to represent the final data set. After the initial download, we found that meteorological weather data provides wind speed in two components, u and v, which can be used to derive wind speed and direction. Without this prior domain knowledge, we excluded one component of the wind data and had to repeat the process of downloading the weather data from the API a second time.

Due to time and project size constraints, we also had to narrow the scope of our analysis from looking at both Louisiana and Michigan to draw comparisons in different regions of the country within the MISO footprint, to analyzing data in Michigan only.

3.8 Bringing it all together

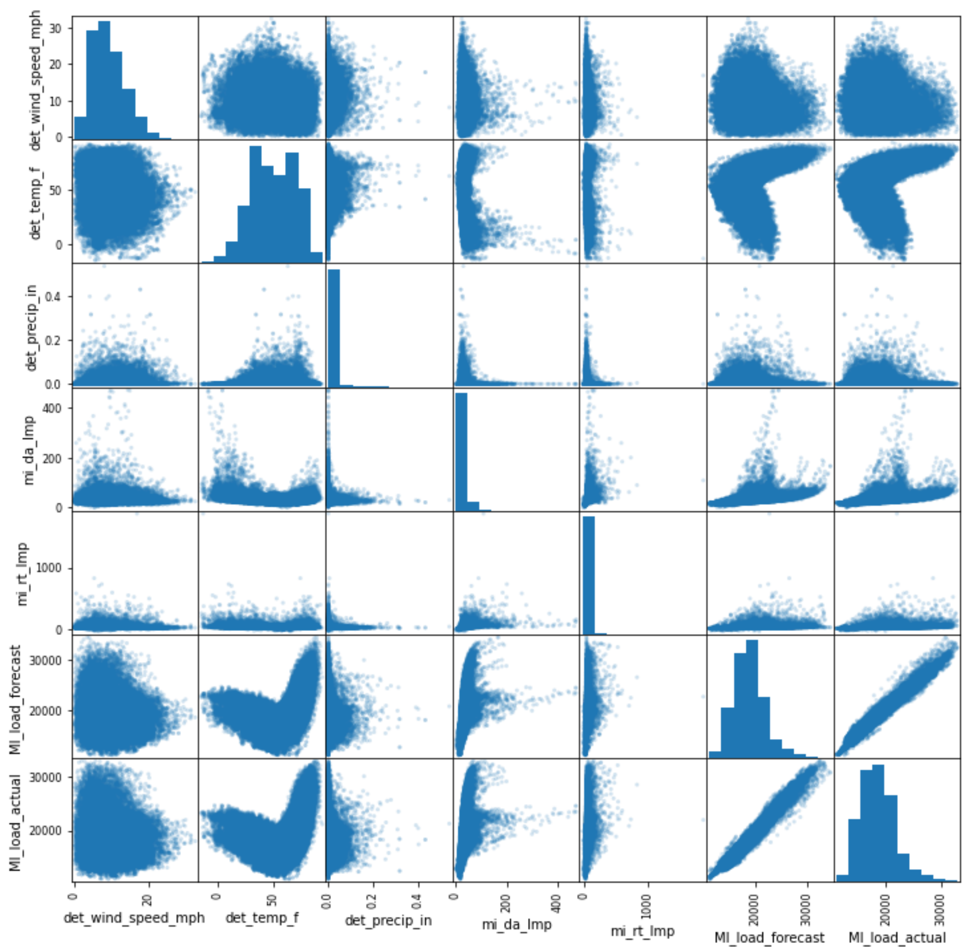
Once the weather and LMP data were in their proper formats, all data was merged into one data set on the `'time'` column and written out to a CSV file. Having all data in one location allowed for simpler analysis.

4 Analysis and Visualization

4.1 Analysis Steps Performed on Datasets

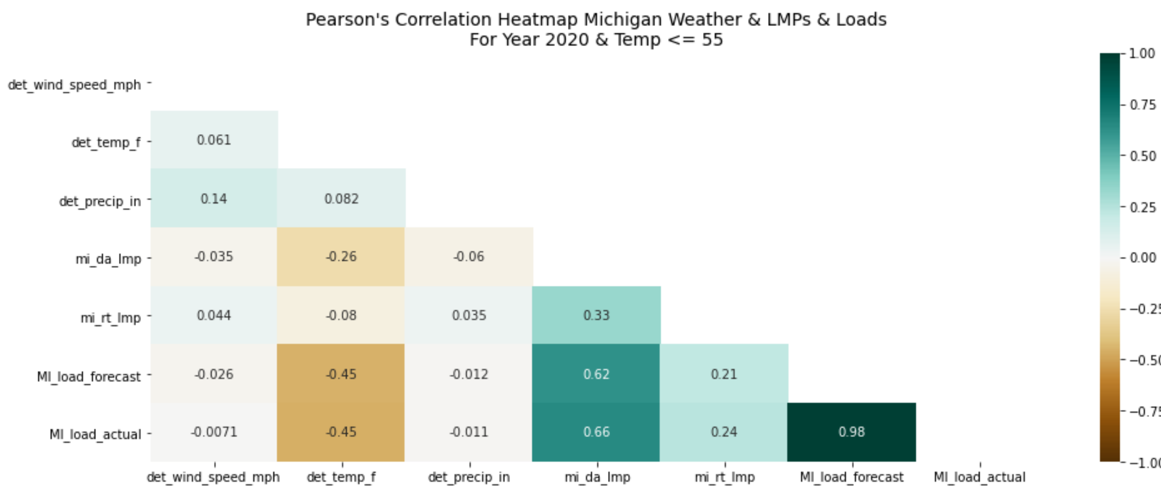
Correlation between various weather data points and energy pricing and loads was a key focus of the analysis. Initially we looked at the actual text-based correlations, using a correlation matrix combined with a heatmap to quickly visualize where the correlations occur and how relatively strong those 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 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 variables. It was at this point we decided to incorporate load data into our analysis. In looking at a SPLOM of our dataset, a pattern emerged between the temperature and load data.

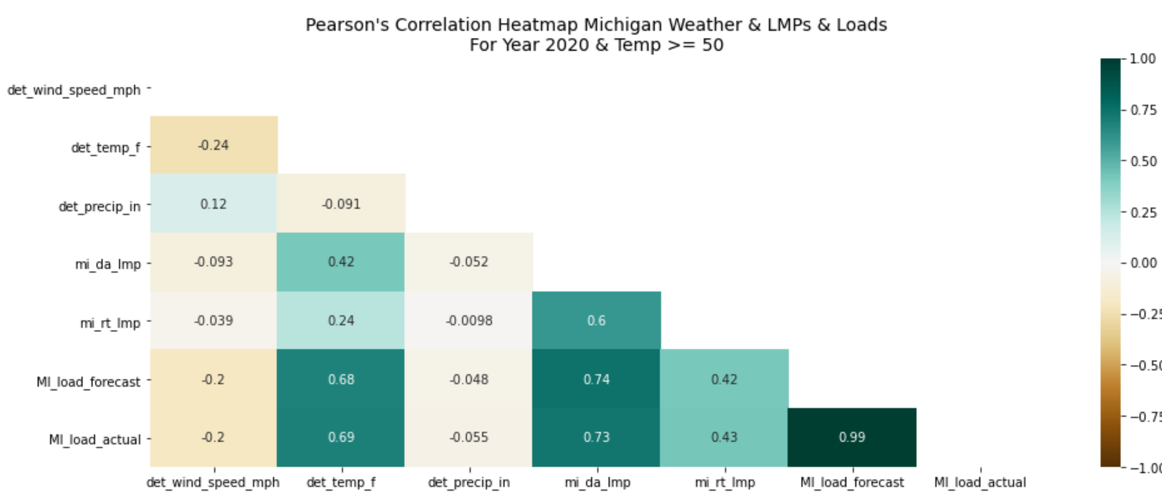


Through this visualization technique, we quickly recognized the relationship between temperature and load. However, this chart was still not detailed enough to give deeper insights into what aspects of temperature had the most influence on major pricing fluctuations. Was there a temperature threshold 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? There were many directions that we could take the analysis and areas of interest to help address the original hypothesis.

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.



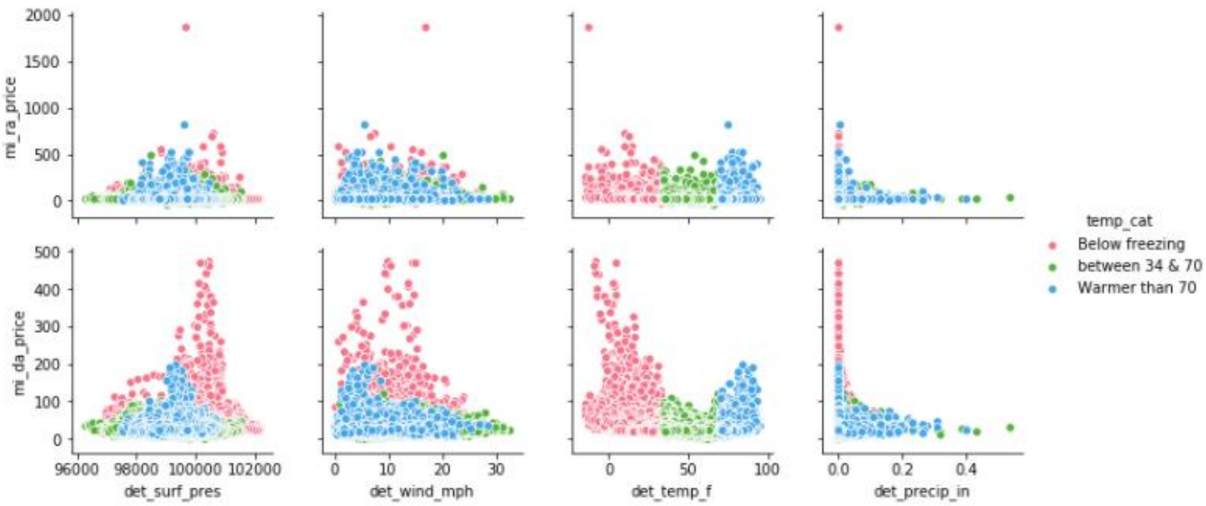
With this newfound insight we began exploring seasonal patterns in the data by encoding temperature categories with color on our SPLOM visualization. Utilizing the script described below, we categorized temperature data into their corresponding bins.

4.2 Summary of Results

Using 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. This helped us to gain a better understanding of the nuances of the interactions between temperature and energy pricing, particularly where among the data the correlations were strongest (or weakest).

Temperature categories:

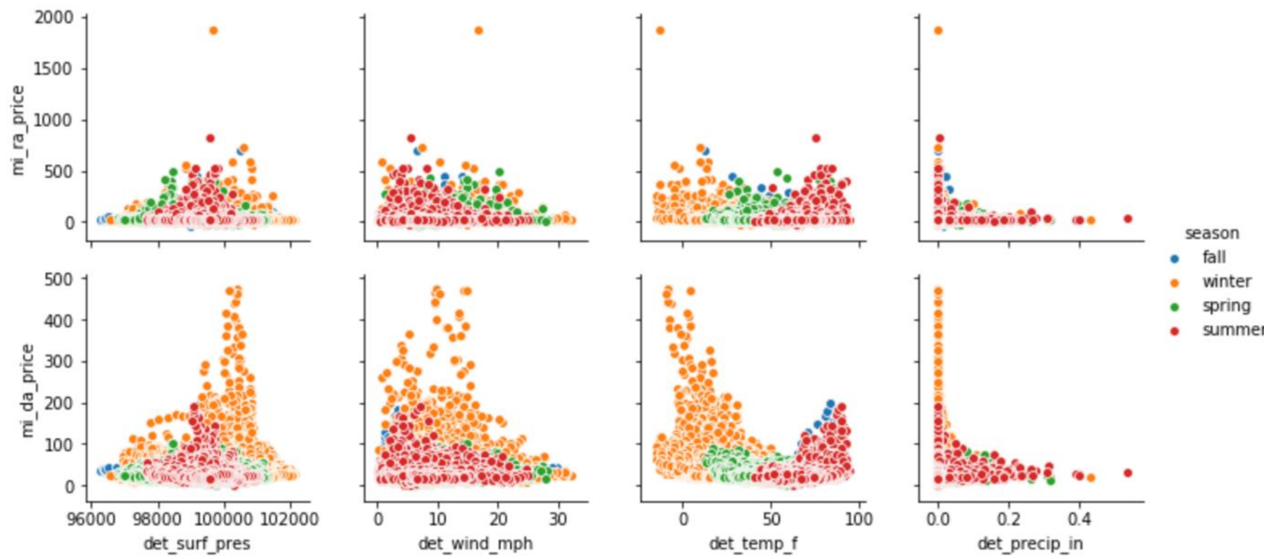
Temperatures below 0 degrees Fahrenheit were encoded red, temperatures between 34-70 degrees were encoded green, and temperatures above 70 degrees were encoded blue.



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. A central range of 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. These were all interesting insights that helped to paint the overall picture of how temperature interacts not just with the energy pricing, but also with the energy pricing combined with other weather factors.

Seasonal Categories:

Given the insightful results from the categorical temperature analysis, we decided to look further at how categories could be used to further understand weather and energy pricing interactions. We chose to explore the seasonal impact of weather, since in general, seasons represent "groups" of weather effects within each category, such as winter representing cold, snow, and winds. This was more of an exploration, since the seasons are not as clear-cut. For example, winter-type weather in Michigan does not stop right in December and end in February. But it did give us a general sense of patterns exhibited in specific groups of months.



As might be expected, the colder temperatures, among other winter weather characteristics, that occur in the winter are highly consistent with the temperature categories charted above, and further supports the conclusions of the temperature categorization analysis. This lends an additional aspect to the analysis, since similar conclusions can be drawn, but with a seasonal perspective, such as the impact of low temperature, low precipitation, and higher surface pressures of the winter months. This helps to explain what this wintry combination of weather elements does to energy pricing fluctuations. Given these results, we see a clear relationship between times when temperatures are below 34 degrees (freezing) Fahrenheit, during the dryer, higher surface pressure winter months and higher energy pricing.

The analysis showed that temperature is a key data point related to the day ahead locational marginal pricing (LMP) fluctuations in Michigan, 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. This resulted in a positive 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 below freezing.

4.3 Suggestions for Further Analysis

A secondary conclusion of the analysis, beyond showing the positive hypothesis about the interactions between temperature and energy pricing below freezing, was that it was not just the temperature that played a factor. 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 the data sources indicated in the analysis, such as wind speed, surface pressure, and precipitation, would give a better understanding of how all of these weather behaviors conspire together, and when they do not, to create the environment by which energy prices fluctuate widely. An additional area that we began to look at, but could be the focus of further analysis, was the energy load amounts, and how these interact with all of the other data points, both weather and energy pricing based. And third, there are numerous additional data sources that are not weather related that could be added to the analysis, such when there are pipeline breakages or congestion. Ideally, all of these data points can be taken into account to provide a full understanding of how energy pricing is impacted over time.

5 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.

---

[i] "2021 Texas Power Crisis." Wikipedia. Wikimedia Foundation, May 13, 2021. https://en.wikipedia.org/wiki/2021_Texas_power_crisis.