Data Analytics Capstone Topic Approval Form

Student Name: Aaron Balke

Student ID: 011005116

Capstone Project Name: Comparing Regional Energy Grid Consumption using Paired T-Testing and TBATS Time Series Analysis

Project Topic: Comparing regional max energy consumption of the PJM energy grid will enable effective resource allocation, maintenance, and repair estimation, and generalized executive decision-making. Using PJM's energy load dataset, t-testing and TBATS time series analysis can be applied to evaluate max energy consumption.

☒ **This project does not involve human subjects research and is exempt from WGU IRB review.**

Research Question: Is the difference between PJM regional max energy consumption means statistically significant? And if so, can PJM Max Energy Load per Region be forecasted using TBATS time series modeling?

Hypothesis: **Null hypothesis**- The difference between PJM regional max energy consumption means is not statistically significant. **Alternate Hypothesis**- The difference between PJM regional max energy consumption means is statistically significant.

Context: The contribution of this research to the MSDA Program, and the field of Data Analytics is to create a paired t-test which can evaluate if the mean energy consumption between geographical regions is statistically significant. PJM Interconnection LLC is a regional transmission organization in the United States. It is part of the Eastern Interconnection grid operating an electric transmission system serving all or parts of Delaware, Illinois, Indiana, Kentucky, Maryland, Michigan, New Jersey, North Carolina, Ohio, Pennsylvania, Tennessee, Virginia, West Virginia, and the District of Columbia (Kaggle, 2018). PJM directly supplies their hourly energy load data on their website (PJM, 2024), which has been compiled onto Kaggle under a Public Domain License. Comparing maximum energy consumption between regions will enable effective resource allocation, maintenance, and repair estimation, and generalized executive decision-making. This study will provide a comparison of maximum energy consumption between PJM regions, using t-testing, and a forecast of energy consumption using TBATS time series analysis. The use of T-testing for conducting energy consumption analysis has already been applied by the Behavioral Insights Team on behalf of the Department for Business, Energy, and Industrial Strategy. In their work, energy consumption was measured to evaluate the effectiveness of residential energy-saving methods (Schein, 2020). While this is not identical to this research, they share insights on methods that will have to be deployed. For example, they required a "difference in differences" method to complete the t-testing, which is analogous to verifying stationarity in the time series data for this research, before t-testing. Beyond T-Testing, TBATS time series analysis will be performed to create a forecast of the complex seasonal components of the individual regional max energy consumption data.

Data: The dataset is published on Kaggle under a Public Domain License. Hourly data is provided for all regions between 1998-2003, and after 2003 data is split by region until 2018. Just the 2003-2018 data for PJMW and PJME will be used, which stands for PJM West, and PJM East respectively. PJM West consists of Ohio, West Virginia, and portions of Illinois, Pennsylvania, Kentucky and Virginia, and PJM East consists of New Jersey, Delaware and Maryland. These regional files total, 288,572 records, with the West region missing the first four months of the year. The first four months of the East data will be removed, which will change our total to 286,412 records, setting the time period of the analysis to be equal between regions.

Feature	Data Type
Date Time	Qualitative
Energy Consumption (MW)	Quantitative
Region*	Qualitative

*Separate files are provided for each region, with identical features. However, the datasets will immediately be merged to simplify analysis.

The first limitation of this dataset is the hourly nature of it. This, while providing a ton of data, does open the model up to having randomness effects. The smaller slices of time are more affected by random moments of extremes. To remove the effect randomness would have, the dataset will be aggregated into daily max values. Daily values should limit the effect random changes have on the data, making it easier to forecast (Grogan, 2022). Important context regarding the dataset includes that in 1996-1999 US electrical utilities were

restructured and regulations were put in place to aid smaller players in the space, not PJM. However, many regulations were removed in 2002 causing expansion (Tomain, 2004). This should not affect the dataset since it falls in after these events, however, it would have a major effect on previous datasets. The data is originally provided directly from PJM's website, where they provide metrics for the public to access. The data has been compiled and uploaded to Kaggle at https://www.kaggle.com/datasets/robikscube/hourly-energy-consumption under a CC0: Public Domain License (Kaggle, 2018) by Rob Mulla, a Senior Principal Data Scientist at H20.ai, an open-source AI/ML platform provider (Mulla, n.d.).

Data Gathering: The dataset is available on Kaggle as individual CSV files or an Apache Paruqet file. Files are provided for each region from 2003 through 2018 and as total values for 1998-2002. All data is hourly energy consumption values. Only the regional data after 2002 will be used. Only PJMW January through April 2002 is missing, and all data is consistent. Our analysis will chop off the equal timeframe from the PJME data, that way data is similar in shape. After the aggregation mentioned in the Data portion, values will become quite large. A conversion from Megawatts to Gigawatts will make the data more readable. For T-testing, the regional data will need to be stationary (Schein, 2020). This is to make sure differences in one region do not confound the identification of another region. Additionally, the differences between data must have outliers removed and be normally distributed. This ensures an inaccurate rejection of the null hypothesis does not take place in t-testing. Trend and Seasonal Components will not be removed prior to t-testing since they will need to be taken into consideration for comparing the means of each region's max energy consumption. For example, if climate differences affect the mean values, this needs to be kept in our analysis.

Data Analytics Tools and Techniques: Paired T-testing has been chosen as the prominent technique for this analysis since it allows the comparison of the means of different region's max energy consumption.
This will allow PJM to allocate resources effectively between regions and make plans for future maintenance and repairs when combined with TBATS time series forecasting. Geographic regions will be studied to identify significant differences in energy consumption. These regions are PJM West, consisting of Ohio, West Virginia, and portions of Illinois, Pennsylvania, Kentucky and Virginia, and PJM East, consisting of New Jersey, Delaware and Maryland. Differences between the regions will undergo the following data preparation: First outliers are removed, and then data is checked for normality using a Shapiro-Wilk test. If the data is not normally distributed, data transformations will be applied to adjust the skew. The null hypothesis will be rejected if a p-value < 0.05 is achieved since this would imply a 95% confidence in our alternative hypothesis. If the regions have statistically significant differences in mean max energy consumption values, then a time series analysis will be completed to forecast future max energy consumption. Exploratory Data Analysis will be performed on the dataset analyzing the daily max loads. This will include using Dickey Fuller testing to check the data is stationary, decomposition of the data to check each component, and finally autocorrelation to check data seasonality information. For the presentation, Google Presentation will be used, after exporting the images/plots from the Python/Jupyter Notebook environment. This has been chosen since it provides a simple method of creating a professional presentation for executive decision-making. Tableau and Power BI were decided against since interactivity would not be necessary for the limited features.

> **Justification of Tools/Techniques:**
> Python will be used for this analysis. Python has been chosen because of its English like syntax when compared to R. Additionally, R would require more to set up since an environment with checkpoint functionality such as Jupyter Notebook, which comes with Python by default (Luna, 2022). SAS would be a good option for direct manipulation of an external database; however, this is not a problem in our research (Johnson, 2023).

Project Outcomes: The expected project outcome is a support for the alternative t-testing hypothesis, showing the PJM region max energy consumption means are statistically different. Support for the alternative hypothesis is seen in The Behavioural Insights Team's work for the Department for Business, Energy, and Industrial Strategy (Shein, 2020), which shows t-testing is an effective method for comparing mean energy consumption between groups.

Projected Project End Date: February 1st, 2024

Sources:

BATS and TBATS time series forecasting. (2022, December 22). GitHub. https://github.com/intive-DataScience/tbats

De Livera, A.M., Hyndman, R.J., & Snyder, R. D. (2011), Forecasting time series with complex seasonal patterns using exponential smoothing, Journal of the American Statistical Association, 106(496), 1513-1527.

Grogan, M. SARIMAX model slopes downwards instead of upwards. (2021, August 2) Cross Validated. Retrieved January 8th, 2024, from https://stats.stackexchange.com/questions/538823/sarimax-model-slopes-downwards-instead-of-upwards

Hourly Energy Consumption. (2018, August 30). Www.kaggle.com. Retrieved January 8th, 2024, from https://www.kaggle.com/datasets/robikscube/hourly-energy-consumption

Johnson, D. (2023, November 21). SAS vs R: What's the Difference? Www.guru99.com. Retrieved January 8th, 2024, from https://www.guru99.com/sas-versus-r.html

Luna, J. (2022, December 28). Python vs R for Data Science: Which Should You Learn? Www.datacamp.com. Retrieved January 8th, 2024, from https://www.datacamp.com/blog/python-vs-r-for-data-science-whats-the-difference

Mulla, R. (n.d.). https://www.linkedin.com/in/rob-mulla/ [Review of https://www.linkedin.com/in/rob-mulla/]. Linkedin; Linkedin. Retrieved January 8th, 2024, from https://www.linkedin.com/in/rob-mulla/

Nadeem. (2021, November 30). Time Series Forecasting using TBATS Model. Analytics Vidhya. Retrieved January 8th, 2024, from https://medium.com/analytics-vidhya/time-series-forecasting-using-tbats-model-ce8c429442a9

PJM - Home. (2024). Pjm.com. Retrieved January 8th, 2024, from https://pjm.com/

Schein, A., Bogiatzis-Gibbons, D., & Hardy, T. (2020). Guidance on conducting energy consumption analysis Created by the Behavioural Insights Team on behalf of the Department for Business, Energy and Industrial Strategy. In bi.team. The Behavioral Insights Team. Retrieved January 8th, 2024, from https://www.bi.team/wp-content/uploads/2020/12/Guidance-on-conducting-energy-consumption-analysis.pdf

Schendzielorz, T. (2020, January 15). A guide to Data Transformation. Medium. Retrieved January 8th, 2024, from https://medium.com/analytics-vidhya/a-guide-to-data-transformation-9e5fa9ae1ca3

Tomain, Joseph and Cudahy, Richard (2004). Energy Law in a Nutshell. Thomson - West Group. ISBN 9780314150585.

Course Instructor Signature/Date:

☒ The research is exempt from an IRB Review.

☐ An IRB approval is in place (provide proof in appendix B).

Course Instructor's Approval Status: Approved

Date: 1/10/2024

Reviewed by: 

Comments: Click here to enter text.