**A Time Series Analysis: ERCOT Power Load Forecasting**

Sarah Alqaysi

Bikram Gill

Jack McCullers

Travis Lloyd

University of San Diego

ADS-506 Applied Time Series Analysis

6 December 2021

# TABLE OF CONTENTS

[TABLE OF CONTENTS 2](#_Toc89713613)

[Problem Statement 2](#_Toc89713614)

[Data Description 3](#_Toc89713615)

[Literature Review 3](#_Toc89713616)

[Project Overview 5](#_Toc89713617)

[Data Exploration 6](#_Toc89713618)

[Data Wrangling 6](#_Toc89713619)

[Pre-processing 6](#_Toc89713620)

[Model Strategies 7](#_Toc89713621)

[Results and Final Model Selection 10](#_Toc89713622)

[Conclusion 11](#_Toc89713623)

[Next Steps and Suggestions 11](#_Toc89713624)

[References 12](#_Toc89713625)

# Problem Statement

Load forecasting is an important aspect of any power grid. Due to the lack of large-scale storage options, system operators must forecast the daily load (demand) ahead of time and ensure adequate power generation is secured. Load forecasts are used to determine which power-generating resources will need to be turned on, which will need to be on standby in case of increased demand and overall affect the economics of the power grid. The time-series data in this study shows historical load profiles over time and the goal of this paper will be to determine if there are any models that can explain the load profile historically (and perhaps be useful in predicting the near-term future load profiles).

# Data Description

The data shows hourly historical power load by region in Texas, as well as the overall number. The study will take data from 2017-2020, it will be in four separate excel sheets. The total records for the cumulative dataset are 35060 hours, each record to include the hourly power load for eight different regions in Texas.

The regions are broken down into COAST, EAST, FWEST, NORTH, NCENT, SOUTH, SCENT, AND WEST. There is a tenth column labelled ERCOT which is the load’s total sum of all 8 regions. This last column will be used to perform univariate analysis The first column in the dataset is the hourly timestamp which is in string format. Checking for missing values, the result was that there were no missing values, expediting the pre-processing step.

# Literature Review

The first source found was a paper completed for the 2018 International Conference on Machine Learning. In this paper, the team of researchers were focused on establishing different machine learning models for load forecasting in China. This study was conducted by J. Yang and Q Wang. Their paper supports the use of a Support Vector Machine model that outperforms their other explored options. This was important for the team to review because it outlined a baseline approach to fitting traditional and non-traditional models to a power load time-series dataset. The researchers from this conference had focused on a national scale within the country of China to accurately model and forecast power load requirements for the nation. This related well back to the focus on one of the largest states in the United States. Texas could be used as an example for how to accurately model for the United States as a whole. This is important as new legislation pushes to explore renewable energy sources as well as pushing for an expansion in the electric vehicles marketplace. All of these new ideas will require power grids to support the increase in power loads.

The second source found was another paper completed for the 2018 International Conference on Machine Learning. This study was completed by Juri, Micu and Muesan on providing a valuable overview of electrical energy forecasting methods and models. In this paper, the team of researchers provided an overview of the actual forecasting methods and models used in renewable energy resources. This provides alternate and available forecasting tools to show which is the most efficient. The team chose to explore this research as it pertains to selecting the most advantageous models based off univariate and bivariate datasets. The team also connected back the findings found in this paper to the new legislation push for finding reliable renewable energy sources within the United States. The team saw applicability with the focus on power load forecasting on the current electrical grid as well as other potential sources such as solar and wind power.

The third source found was an article from IEEE Transactions on Power Systems November 2013 issue completed by Paparoditis and Sapatinas. The paper introduced a novel functional time-series methodology for short term load forecasting. This was applied to data of historical daily loads in Cyprus. This study can pull future sources from this same IEEE Transaction on Power Systems journal. This is because they have many publications on load forecasting. The team saw the importance in their study as it connected to a focus on short term load forecasting. The team drew connections to how they could apply model fitting over shorter-term periods as a future focus area because the team had already decided to focus specifically on the years 2017-2020. This short-term load model fitting could be accomplished off the team’s dataset since it contained daily power load amounts with time stamps associated with them. This again could be a future area of exploration.

This study identified sources are important in understanding what load forecasting is and the role machine learning plays with the creation of models to support the objective. The study plans on using these sources to help support the models created within this project. This study also focuses on how to determine if there are any models that can explain the load profile historically.

# Project Overview

The team decided to approach this project with the objective of fitting multiple traditional time-series models and one neural network model to the Texas power load dataset. The plan the team created was to first perform exploratory data analysis, data wrangling, and then move into modeling the data. The team had established a good baseline understanding of the applicability that traditional and non-traditional modeling can have for forecasting time-series data. The team had decided to focus on fitting the traditional models and providing commentary on the finding. The team would then attempt to fit and forecast using the neural network model. The team had seen from the sources identified in the literature review how other researchers had attempted support vector machines, and other modeling techniques to fit and forecast power loads both on short- and long-term scales. As stated previously, the team decided to focus specifically on the years 2017-2020 for the analysis that will be outlined below. The team planned to provide final conclusions as well as identifying next steps to take the time-series modeling further.

# Data Exploration

In the Exploration step the data was plotted using ggplot2 to create subplots of all the columns, less the data column. When the plot was created, it was visually clear that there was seasonality to the data. The data was then plotted in individually to look for trends, checking for stationarity. The data was found to overall have a mild trend with increasing power demand. This will then be addressed in the pre-processing step.

# Data Wrangling

## Pre-processing

The load dataset for 2017 had a space in the Timestamp column, preventing the binding of the four years. Using bind\_rows from tidyverse, the data from 2017-2020 was merged. Following that, using strptime, the time variable data type was set. The dataset also had several NA's which arise due to issues with daylight savings (1 record in each year), hence, they were removed. Finally, the dataset was converted to a time series format for the project analysis.

Addressing the mild trend to the time series, there were different approaches taken to check for the most optimal way to detrend the data. The measures taken were Standardization, Normalization, Box-Cox and Differencing. Eventually differencing was chosen as the preferred method as it retained most of the seasonality. All four of the datasets had the same variable names except 2017’s HourEnding variable as it had a space in-between hour and ending. This was fixed to match 2018-2020’s HourEnding without a space. The timeseries was then formatted from a data.frame to a timeseries using the forecast package.

# Model Strategies

To begin with, P/ACF of the raw series and subsequent transformed series were evaluated. As can be seen in Figure 1; the raw data shows high cyclicality, and the series is certainly not stationary. This is to be expected, power demand has daily cycles (power demand being higher at certain hours of the day than others) and seasonal cycles (Winter has higher power demand than warmer months due to air-conditioning usage).

In order to overcome this, the first and second-order differenced data was observed, however the P/ACF plots still showed cyclicality. In particular, there was a continuous correlation of lag 24 (and multiples of it). Once again, this made sense in the context of the data, as the demand at any given time would be highly correlated to the demand 24, 48, 72 etc. hours before it.

As such, the data was differenced on its lag 24 (seasonal difference), which finally showed a P/ACF plot that could be modelled. Based on the plot shown, an ARIMA (24, 24, 0) model appeared to be most appropriate as the ACF tailed off and the PACF cut off for this final plot. Another model that was tested which is the GARCH (1,1) with ARIMA (24,24,0). The results included Akaike (AIC), Bayes (BIC), Shibata and Hannan-Quinn criteria for the model estimation. Basically, the lower these values, the better the model is in terms of fitting. And it was shown that results were in the (-4) range, which is good.

A second evaluation metric is the Ljung-Box test for testing the serial correlation of the error terms. The null hypothesis states there is no serial correlation of the error terms. Basically, if the p-value is lower than 5%, the null hypothesis is rejected. The model results show that the p-value is higher than 5% on the on "Standardized Squared Residuals", meaning that there is no serial correlation of the error term. Lastly, another interesting one to check (Figure 2), is *Adjusted Pearson Goodness-of-Fit* which measures the goodness of fit of the error. The null hypothesis says that the conditional error term follows a normal distribution. If the p-value is lower than 5%, the null hypothesis is rejected. The model results show three of the p-value higher than 5% and one lower. Figure 3 shows key plots of the model parameters, including graphs of ACF, Sample and Theoretical quantiles and the model Conditional SD plot.

**Figure 1**

*P/ACF Plots of ERCOT Time Series*

Chart

Description automatically generated with medium confidence A picture containing text, antenna

Description automatically generated

Chart, box and whisker chart

Description automatically generated with medium confidenceA picture containing chart

Description automatically generated

**Figure 2**

*GARCH Model: Adjusted Pearson Goodness-of-Fit Test*

Text

Description automatically generated

**Figure 3**

*GARCH Model: Adjusted Pearson Goodness-of-Fit Test*

Graphical user interface

Description automatically generated

# Results and Final Model Selection

Based on information gathered in the P/ACF exercise, the ARIMA (24, 24, 0) model was fit and the results can be seen in Figure 4. Overall, this model fit quite well with a low MAPE (0.07) and low AIC (-216,565.9). What this shows is not only was the ARIMA model a good fit, it suggests that power demand is highly predictable according to the time of day. In the broader context of this assignment, this shows that forecasting power load is achievable.

**Figure 4**

*ARIMA (24, 24, 0) Model Summary and Forecast*

Text

Description automatically generated

# 

# Conclusion

## Next Steps and Suggestions

In terms of next steps, having a model that can fit and predict power demand is useful for power grid planning and commercial exercises.

For planning purposes, Independent System Operators can use the load forecast to determine their power generation needs, which need to be confirmed in advance of actual power usage.

For commercial enterprises, particularly trading firms, having an accurate load forecast is one component of predicting power prices which could then be traded against. Load forecasts represent the demand side of power usage, which could then be extended to predict which generation units would need to be turned on to meet this demand. These two forces create the market price of electricity every hour of the day and could be speculated against.

# References

Du, J., Cheng, Y., Zhou, Q., Zhang, J., Zhang, X., & Li, G. (2020). Power load forecasting using BILSTM-attention. IOP Conference Series: Earth and Environmental Science, 440(3), 032115. https://doi.org/10.1088/1755-1315/440/3/032115 https://iopscience.iop.org/article/10.1088/1755-1315/440/3/032115/pdf

Jurj, D. I., Micu, D. D., & Muresan, A. (2018). Overview of Electrical Energy Forecasting Methods and Models in Renewable Energy. 2018 International Conference and Exposition on Electrical And Power Engineering (EPE), Electrical And Power Engineering (EPE), 2018 International Conference and Exposition On, 0087–0090. https://doi-org.sandiego.idm.oclc.org/10.1109/ICEPE.2018.8559807

Madrid E. A., Antonio, N. (2021). Short-Term Electricity Load Forecasting with Machine Learning. Information, 12(2), 50. https://doi.org/10.3390/info12020050.

Paparoditis, E., & Sapatinas, T. (2013). Short-Term Load Forecasting: The Similar Shape Functional Time-Series Predictor. IEEE Transactions on Power Systems, 28(4), 3818–3825. 10.1109/TPWRS.2013.2272326

Tsafack, I. (2021). *GARCH models with R programming : a practical example with TESLA stock*. idrisstsafack.com. <https://www.idrisstsafack.com/post/garch-models-with-r-programming-a-practical-example-with-tesla-stock>

Yang, J., & Wang, Q. (2018). A Deep Learning Load Forcasting Method Based on Load Type Recognition. 2018 International Conference on Machine Learning and Cybernetics (ICMLC), Machine Learning and Cybernetics (ICMLC), 2018 International Conference On, 1, 173–177. https://doi-org.sandiego.idm.oclc.org/10.1109/ICMLC.2018.8527022