Forecasting Energy Demand in Ontario

An in-depth analysis of the demand of energy in Ontario

By:
Suhani Didwania
Jonah Chua
Kaarthick Rameshkumar

Table of Contents

Introduction

The Independent Electricity System Operator (IESO) ensures Ontario's electricity grid operates reliably by balancing supply and demand. Accurate demand forecasting is critical for optimizing energy generation, reducing costs, and preventing outages or waste.

Ontario's electricity system is influenced by factors like weather, seasonality, and consumer behavior. With growing reliance on renewable energy and conservation policies, precise forecasts are essential for efficient power generation and distribution.

The data for this project was retrieved from Kaggle. This dataset provides hourly information on electricity demand and average hourly prices spanning two decades (2002–2023). Each observation includes:

- **Hour**: The hour of the day (Eastern Standard Time), from midnight to midnight.
- **Hourly Demand**: The total hourly electricity consumption in kilowatt-hours (kWh).
- **Average Hourly Price**: The Hourly Ontario Energy Price (HOEP) in Canadian cents/kWh, representing the average of twelve five-minute market clearing prices weighted by electricity usage.

By analyzing this dataset, we aim to build a forecasting model to predict energy demand in Ontario. Such a model will enable the IESO to:

- Proactively manage electricity supply to meet varying demand levels.
- Enhance the integration of renewable energy by predicting periods of high or low demand.
- Support planning for future electricity infrastructure development and grid expansion.
- Improve cost efficiency and reliability of the electricity grid.

Goal: The project will focus on leveraging advanced models taught in STAT 443 to forecast future energy demand in Ontario. By incorporating trends, seasonality, and external factors into the model, we aim to provide actionable insights that can help the IESO optimize Ontario's energy management system for the benefit of residents, businesses, and the environment.

1. Data Preprocessing

The data was already fairly clean (with no null values), however, based on our modelling requirements, some preprocessing was required.

Key Preprocessing Steps:

- **Aggregation to Monthly Average Demand**:

The original data contained hourly electricity demand, which exhibits multiple seasonal patterns (e.g., daily, weekly and monthly seasonality). Since our analysis focuses on monthly average demand, we aggregated the hourly data into monthly averages. This approach simplifies the seasonal components and is consistent with the scope of our coursework, as handling multiple seasons was not covered.

- **Handling Incomplete Data for April 2023**:

The dataset for April 2023 contained only 4 days of data, which was insufficient to calculate a reliable monthly average. As a result, April 2023 was excluded from the dataset.

- **Resulting Dataset**:

The final dataset contains monthly average electricity demand from May 2002 to March 2023. This time series data is suitable for modeling and forecasting long-term trends and seasonal patterns.

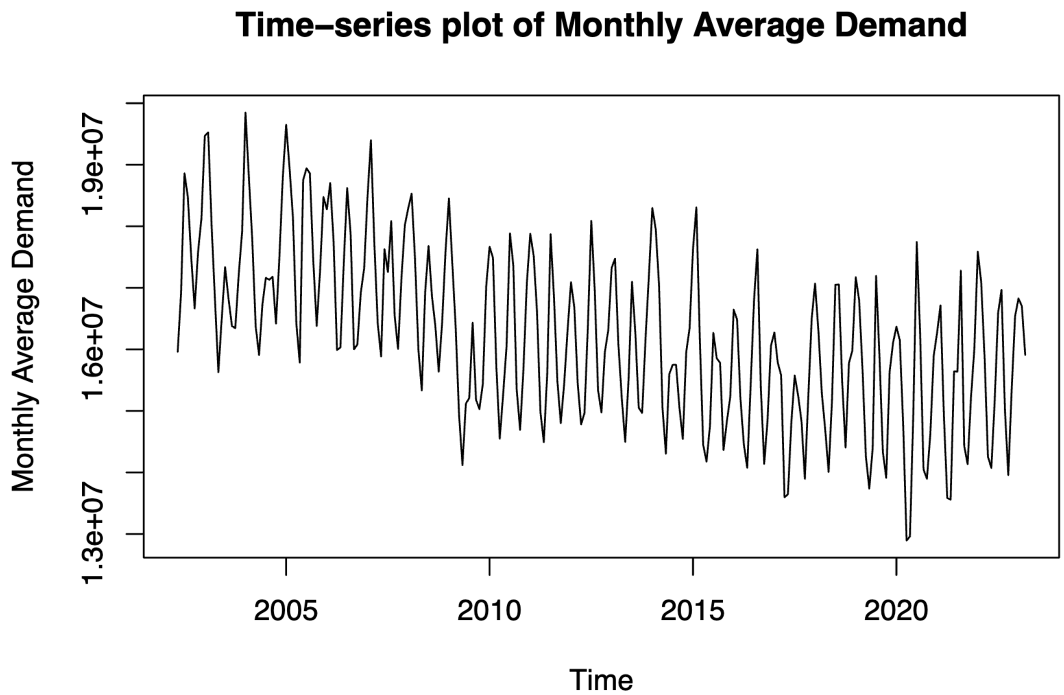
This preprocessing ensures that the data is appropriately structured for analysis while addressing potential limitations of the original dataset.

- **Train-Test Split:**

The dataset is split into training and test sets:

Training Data: Up to March 2022.
Test Data: From April 2022 onwards i.e., last year of the dataset

2. Exploratory Analysis



We do not see an obvious trend (although it could possibly have a slow downward one), but there is a clear seasonality. The variance seems to be constant which is further confirmed by the small value p-value for the Fligner-Killeen test of homogeneity of variances.
We conclude that the data is not stationary due to the presence of seasonality.

3. Regression Modelling

Monthly Average Demand was modeled using combinations of the following predictors:

- Time: Represented as a fractional year with polynomial terms (up to degree 20).
- Month: Binary indicator for each month.
- Monthly Average Price: Electricity price in cents/kWh.

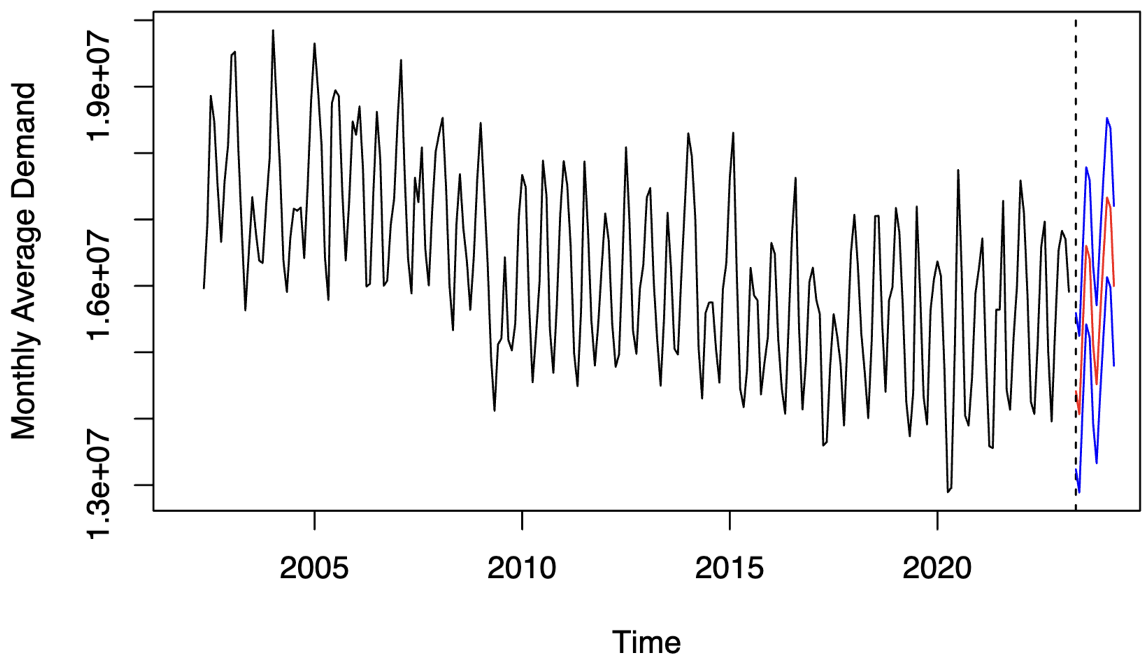
The following models were tested using these covariates:

- Regular Linear Regression: Minimizing square loss.
- Regularized Models: Ridge, Lasso, and Elastic Net with cross-validation to optimize lambda.

Best Model

- Regular regression with **time and month (degree 3, with intercept)** achieved the lowest APSE.
- Residuals showed no significant correlations and were normal and stationary.

This model refitted on the entire dataset, and its forecast for the next 12 months is generated below:



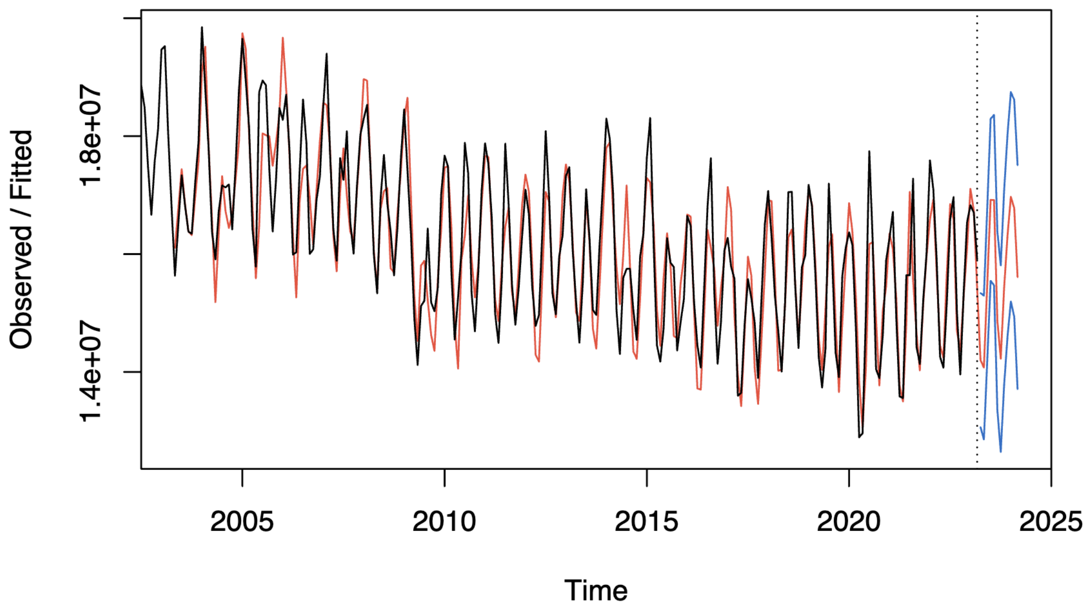
We also notice that the residuals from the fitted and the full observed data seem to be normal with no time correlation and the ACF seems stationary. This supports the validity of our model and its prediction intervals.

4. Smoothing Methods

Various smoothing models like Exponential Smoothing, Double Exponential Smoothing, Additive Holt-Winters and Multiplicative Holt-Winters are fitted to the training data. Predictions are made on the test data, and their squared errors are averaged to compute the APSE for each model.

Model Comparison:

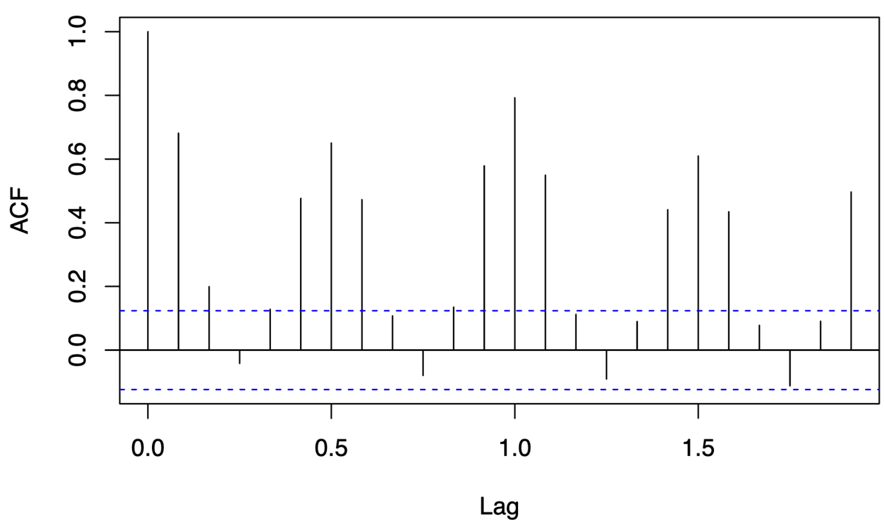
The **Additive Holt-Winters model** emerged as the best performer based on APSE. This model was refitted on the entire dataset, and forecasts for the next 12 months are generated below:



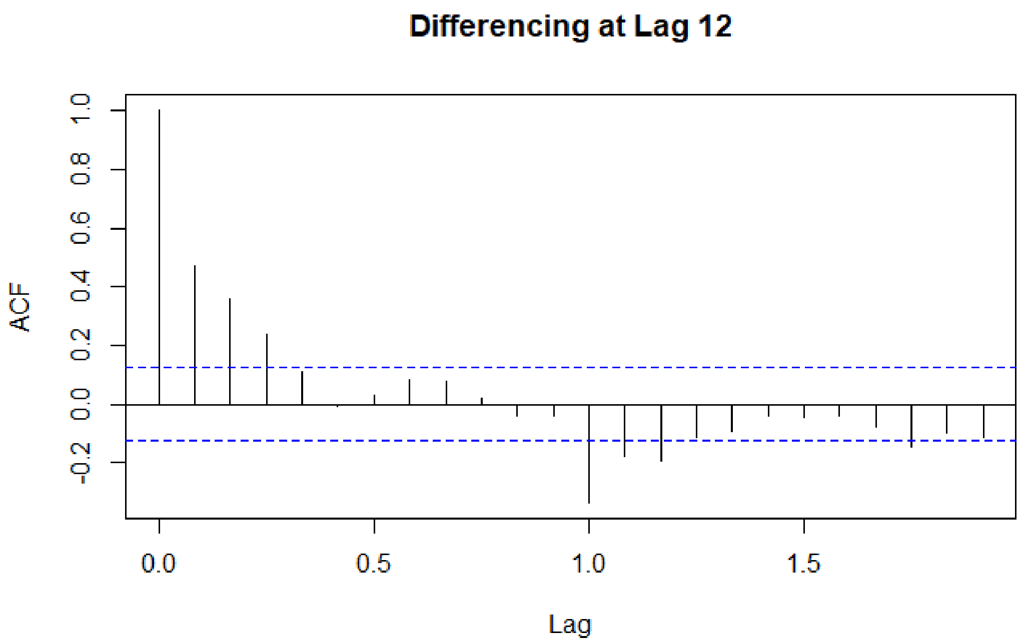
The residuals of the **Additive Holt-Winters model** represent white noise, which means the model fits well and captures the underlying patterns of the data.

5. ARIMA/SARIMA Models

The ACF plot of the data is as follows:



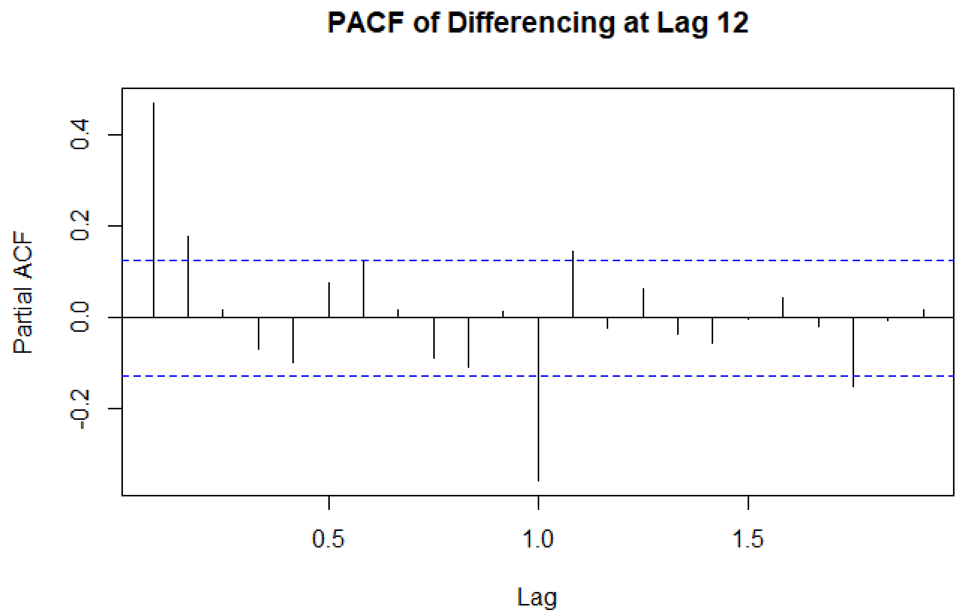
From the ACF of the data, we see that there is certainly evidence of seasonality with period 6 or 12, as discussed in the exploratory data analysis. It is unclear if there is a slow linear decay or not because the spikes near the lag of 6 months are lower than the spikes at lag of 12 months. However, after 12 months we do indeed see evidence of decay. Therefore, it could be beneficial to difference this data at lag 6, 12, or to do one time regular differencing. These three methods were each performed, but the best differenced data was the data which was differenced at lag 12. The ACF of this data is as follows:



From this ACF plot, we see that although there is still seasonal correlation, there is neither a slow decay nor periodicity. We can conclude that stationary has been achieved after performing 1 time differencing at lag 12.

Proposing ARIMA/SARIMA models:

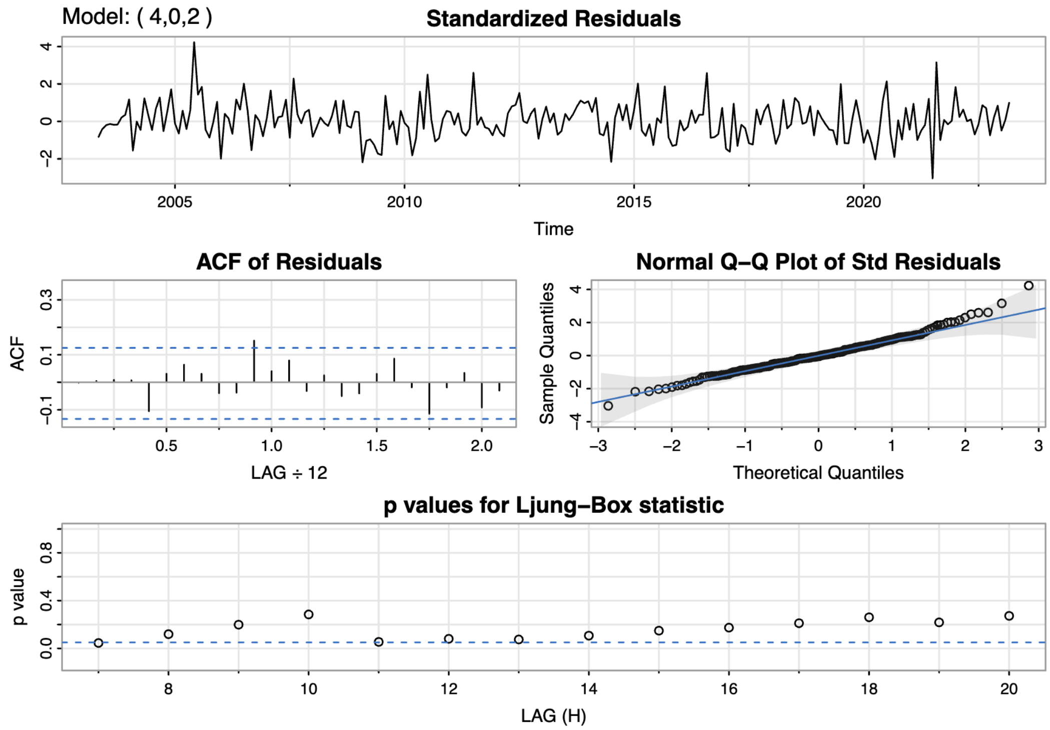
An appropriate model for the data may be a SARIMA with seasonal differencing performed once with season set to 12 months. We also see that there is exponential decay in the ACF and it cuts off at either a lag of 3 months or 14 months. For the purpose of fitting a SARIMA model, we will assume that the ACF cuts off at lag 12 and assume that the remaining lags (which are not at lag 12) are not significant. There is clearly seasonal correlation present at lag 12. Hence, we may propose a SARIMA model where D = 2, P = 0,1, and p = 0,1,2,3. The PACF of the data which has been differenced once at lag 12 months is as follows:



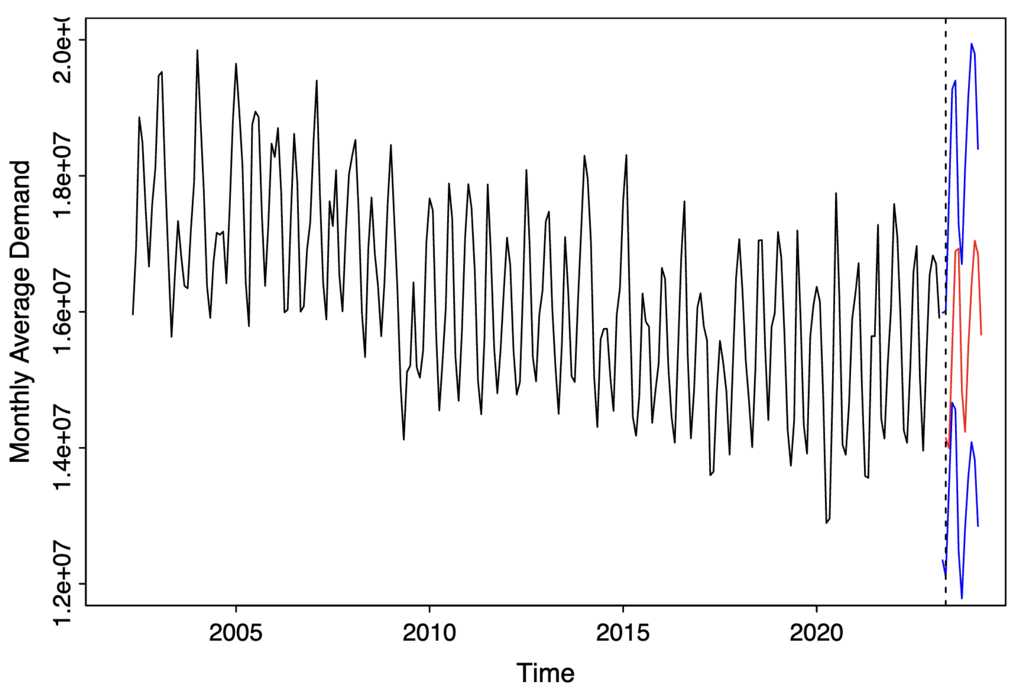
In this PACF, we see exponential decay and a significant seasonal correlation at a lag of 12 months. We can also conclude that the PACF cuts off at lag 1. Hence, we may propose SARIMA models with Q = 0,1 and q = 0,1. All these SARIMA models were fit and their APSE values were calculated.

In addition to these models, SARIMA models were also fit to the residuals of the exponentially smoothed data, the elastic net regression models, and also the regular linear models. The ACF and PACF plots for these data will not be presented in the main section of this report, but they can be found in the appendices. The best of these models, which was also the best overall SARIMA model based on APSE, was a SARIMA(4,0,2)(1,1,1) model fit with a season of 12 months. Its diagnostic plot is as follows:

SARIMA(4, 0, 2)x(1,1,1)_12 on Holt-Winters Multiplicative residuals:



From this plot, we see that the model generally satisfies all its assumptions. This model's fitted values are as follows:



Final Model Selection:
After training and testing various models like regression, regularization, smoothing and SARIMA models, we identified **regular regression with time and month (degree 3, with intercept)** as the most effective model for forecasting Monthly Average Demand. This model demonstrated the lowest APSE (which can be found in Appendix A) on test data, indicating strong predictive performance. Residual diagnostics (which can be found in Appendix C) confirmed that the model's assumptions were met, with no significant autocorrelations and stationary, normally distributed residuals.

Conclusion

The goal of this project was to accurately forecast monthly electricity demand in Ontario using various time series models. After evaluating multiple models, we identified a simple regression model using time and month as factors as the best-performing model based on its low Average Prediction Squared Error (APSE). This model performed the best in capturing seasonal patterns and provided accurate predictions for future demand.

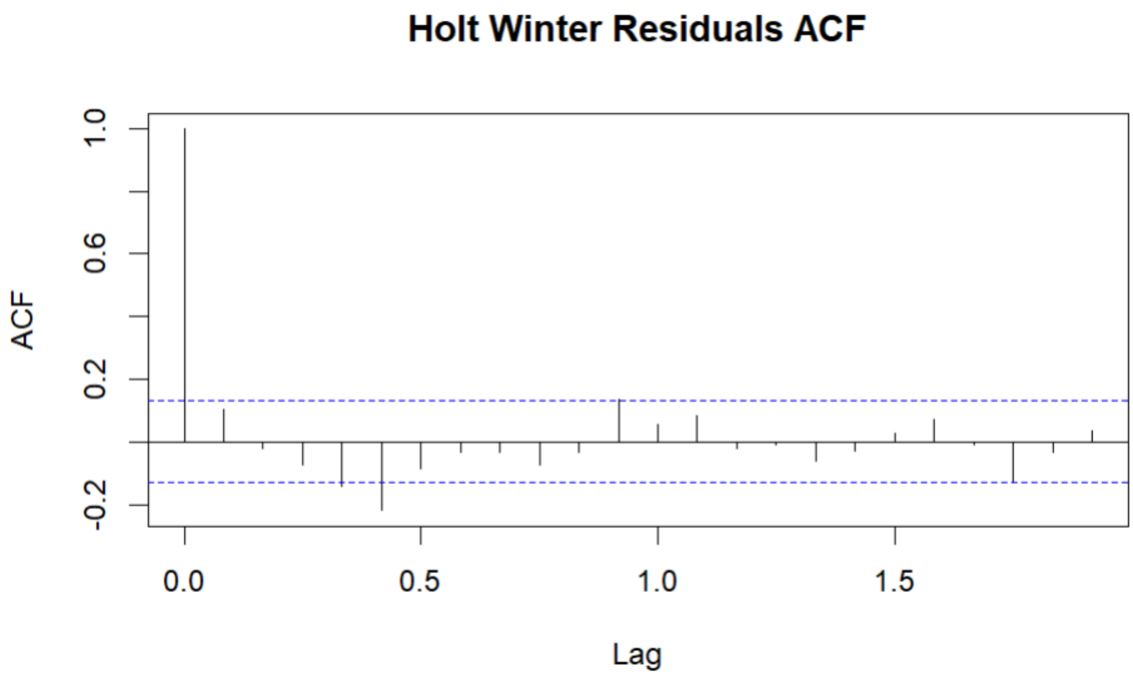
Using this model, we successfully predicted electricity demand for the next 12 months, which can help the IESO plan energy supply more efficiently, avoid outages, and reduce costs. This approach supports better energy management and contributes to the integration of renewable energy sources, ensuring a more reliable and sustainable electricity system for Ontario.

Appendix A: Best APSE Values

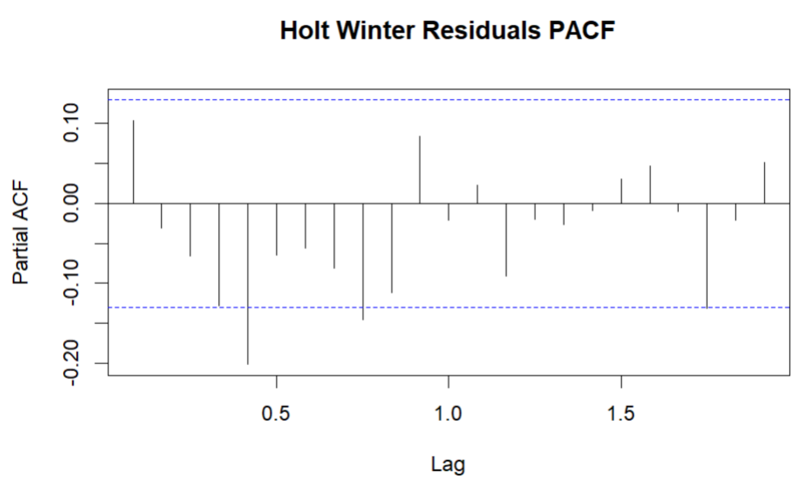
Model Class	**Best APSE Value**
Regular SARIMA Models	161023649802
SARIMA Models on Residuals of Linear Model with Price as Covariate	1.604847 x 10^12
SARIMA Models on Residuals of Holt-Winters Multiplicative Model	115129287334
Exponential Smoothing Methods	139667294919
Regular Linear Regression Model (Covariates were Subset of Time (Polynomial up to Degree 20), Time, Price, and Month)	97227915187
Elastic Net Models (Same Covariates as Linear Regression)	155009840923.835
SARIMA Models fit to Residuals of Elastic Net Models (Always fit with Intercept)	161023649802

Appendix B: Analysis of Holt Winters Residuals

In order to fit SARIMA models to the residuals of the Holt-Winters smoothing methods, their ACF and PACF plots were analyzed. Note that models were only fit to the residuals of the multiplicative Holt-Winters residuals, since this model had the lowest in-sample APSE. Its ACF is as follows:



From this plot, we could propose a SARIMA model with p = 0,1,2,3,4, or 5 though it is possible that the correlation we see at lag 4 and 5 are not actually significant. We also clearly see there is exponential decay. The PACF is as follows:

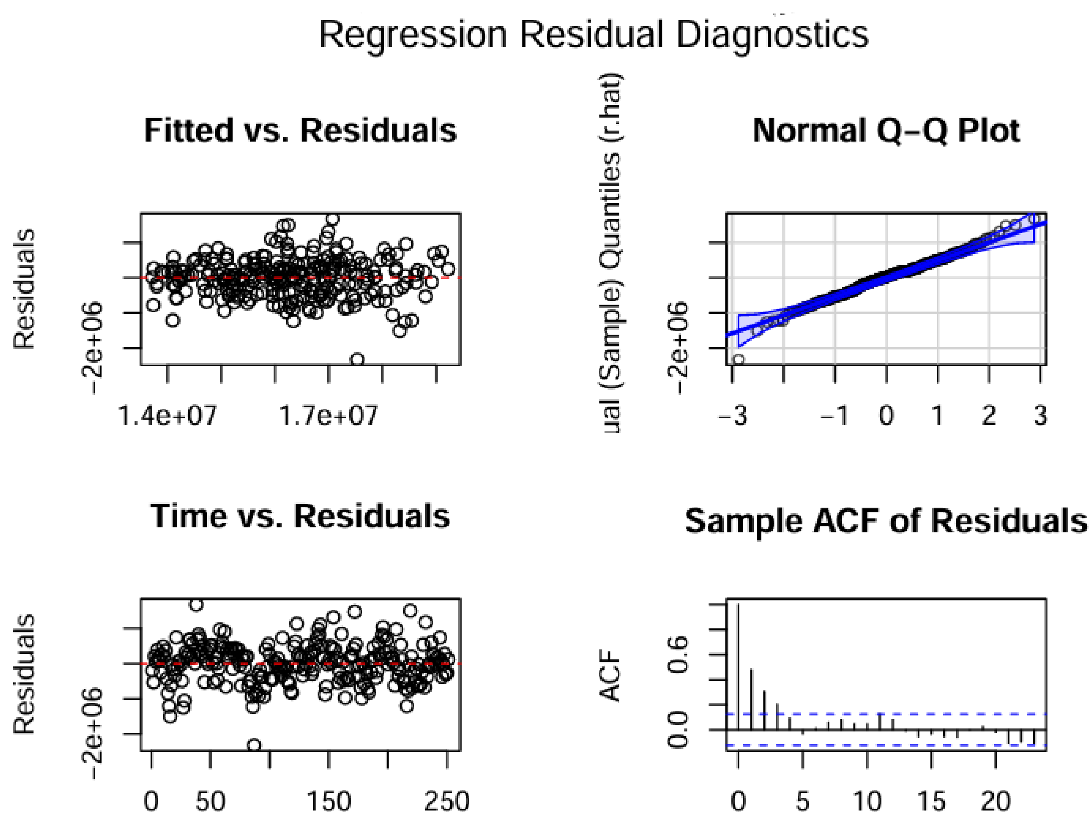


From this PACF, we could propose a SARIMA model with $q = 0,1,2,4$, or 5. There is no evidence of any seasonal correlation or need for differencing, so we set $P = Q = D = d = 0$. All these models were fit to the data.

ACF plot analysis was also done for the residuals of the linear models and elastic net models. However, there are far too many plots and also the models didn't end up being very good in the end. Hence, an analysis of these plots will not be performed. Those who are curious may check our code for these plots to understand the models which were fit.

Appendix C: Residuals Analysis of Best Regression Model

We verified the validity of the predictions from the regression model by checking the ACF, QQ-plot, Residuals vs Time, and Fitted vs. Residuals as seen in the plots below.



From the plots, there does not seem to be a significant correlation between time and fitted values. This is further supported by the sample ACF. This indicates that the data is stationary with exponential decay in the ACF. Furthermore, the data fits well with the bands of the QQ plot. Given these, we feel confident that the predictions and their intervals are valid since the normality and lack of correlation assumptions are satisfied.