Modeling energy consumption in Tetotuan, Morocco

Team 5
Akane Fujimoto, Victor Lai, Xingjian Wang

Summary

The project aims to identify patterns and forecast energy consumption in in Tetotuan, Morocco using data collected in 2017. Energy consumption modeling is important as it allows improved grid management and load balance in energy plans, avoiding potential blackouts and waste.

The analysis shows that energy consumption exhibits significant hour-of-the-day and monthly seasonality. We utilized univariate and multivariate models, including ARIMA, ARIMAX, VAR, and VARX, to obtain forecasts. The results show that ARIMA with exogenous models can capture the seasonality of the data and produce acceptable forecasts. Additionally, adding temperature, humidity, and wind speed as exogenous variables improved model accuracy.

Table of Contents

Project Description

The project focuses on the analysis and prediction of power consumption in Tetotuan, Morocco in 2017. Tetotuan is located in northern Morocco, between the Mediterranean coast and the Rif mountains. It is the 11^{th} most populous city in Morocco with 380,787 inhabitants. Morocco's energy is heavily dependent on fossil fuels and on the private sector, as fossil fuel imports accounts for 90% of the total primary energy supply [1]. The country has a plan to source more than half of its energy from renewable sources by 2030, with a focus on solar and wind power plants.

Energy plants are interested in understanding energy consumption to achieve more optimal planning. Insights on energy consumption in the short term (days or weeks ahead) allow improved grid management and load balancing. Using consumption forecasts, operators are able to balance the supply and demand on the electrical grid, thus potentially preventing overloading the system during peak demand times and reducing waste.

We aim to identify patterns in energy consumption as well as evaluate various forecast models to predict energy consumption using data from Tetotuan.

Data

The dataset is found in the UC Irvine Online Machine Learning Repository [2]. It provides time series data on power consumption (kWh) of Tetouan, Morroco, from January 1, 2017, to December 31, 2017, collected daily every 10 minutes. The power consumption is measured in 3 distribution networks that power 3 zone stations: Quads (zone 1), Smir (zone 2), and Boussafou (zone 3). The data was originally collected by the Supervisory Control and Data Acquisition System (SCADA).

Environmental factors include temperature (C), humidity (%), wind speed (m/s), diffuse flows (representing a specific type of diffuse flow in the city), and general diffuse flows (general diffuse flows in the city, which may include air or water flows). These external variables were collected every 10 minutes, matching energy consumption data. The environmental data was collected by the research team using sensors (cite). All data fields are complete.

Objectives

The project has the following objectives:

1. What is an appropriate level of aggregation for data analysis? The data is reported every 10 minutes.

We hypothesize that aggregating in 1-hour intervals is appropriate as energy consumption is not expected to significantly vary within the same hour. Additionally, this interval might allow us to utilize hour-of-the-day seasonality.

2. What are the trends and seasonal patterns in the energy consumption data, if any?

We hypothesize that trends are not significant as we do not expect to consumption increase throughout the year. We hypothesize that the hour of the day, day of the week, and monthly seasonality are significant. Energy consumption is expected to be higher during nighttime and during the summer months. Day of the week seasonality could be present depending on the behavior of the residents of the city.

3. Are environmental factors (temperature, humidity, wind speed) associated with energy consumption? What are these relationships?

We hypothesize that environmental factors are associated with energy consumption, especially temperature. We expect that energy consumption increases as temperature increases.

4. Can we predict future energy consumption using historical data and environmental factors?

We hypothesize that energy consumption forecast can be done using time series models, as the data is believed to depend on previous data as well as exhibiting significant seasonal patterns.

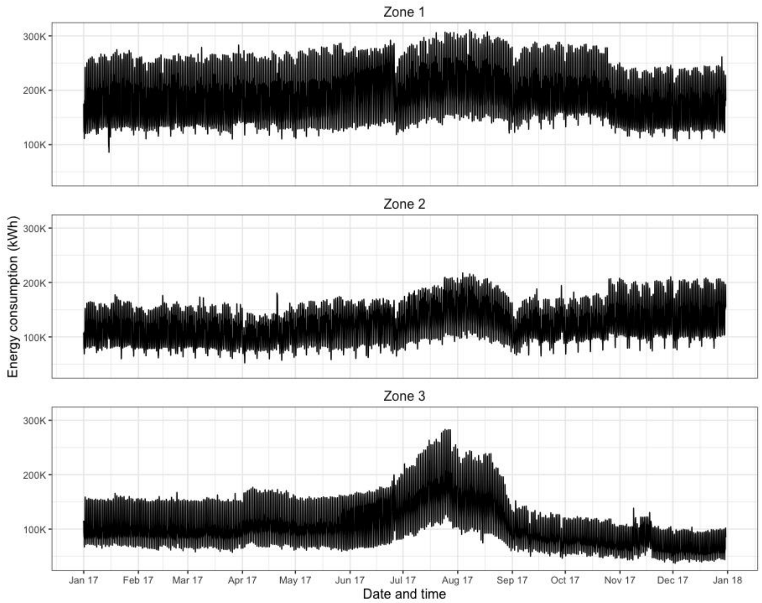
Data exploration

In this section, we explore the data and provide descriptive summaries to better understand the data and identify patterns in the data that can aid model configurations.

The energy consumption time series is collected every 10 minutes, which is very granular for prediction, especially when using time series models that use past data. Under this data structure, there are 144 data points per day, which creates complex models with high orders to capture relationships with lagged data. Additionally, it creates more layers of seasonality in the data.

Figure 1 shows the energy consumption aggregated in 1-hour intervals. We can see seasonal patterns through each day, as the consumption goes up and down daily, as well as monthly seasonality with higher consumption starting in July through September, which corresponds to summer months. We don’t observe a clear trend in the data.

Figure 1: Energy consumption aggregated in 1-hour intervals from January 1, 2017 to December 31, 2017,



We aggregated the yearly data by time of the year (figure 2A), day of the week (figure 2B), and day of the month (figure 2C) to understand seasonality better. We observe that there is a strong hour of the day seasonality. The energy consumption steadily decreases from midnight to 7 am, increases from 7 am to noon, remains constant until 4 pm, increases in the early evening peaking at 8 pm, and then decreases until midnight. This hour of the day seasonality shows that aggregation to 1-hour blocks seems appropriate as it can be used as a predictor. The data does not exhibit a strong day of the week pattern as all days have the same median energy consumption, with a decrease in consumption on Sundays. The monthly plot shows that there is month of the year seasonality, with an increase in energy consumption in the summer months (June, July and August). This analysis shows that hour of the day and month of the year are appropriate seasonal factors to consider in the modeling.

As observed in the presence of monthly seasonality, energy consumption increases in the summer months which indicates that energy consumption might also be associated with environmental factors such as temperature, humidity, and wind speed. Figure A in the Appendix shows the time series for the environmental factors described in the data section. Figure 3 shows the contemporaneous correlation between the variables. As expected, energy consumption is correlated with temperature, humidity, wind speed, and general diffuse flows in different levels. In the modeling steps to follow, we evaluate the cross-correlation among the variables.

Figure 2: Aggregation of energy consumption yearly data into (A) hour of the day, (B) day of the week, and (C) month of the year to evaluate seasonal patterns.

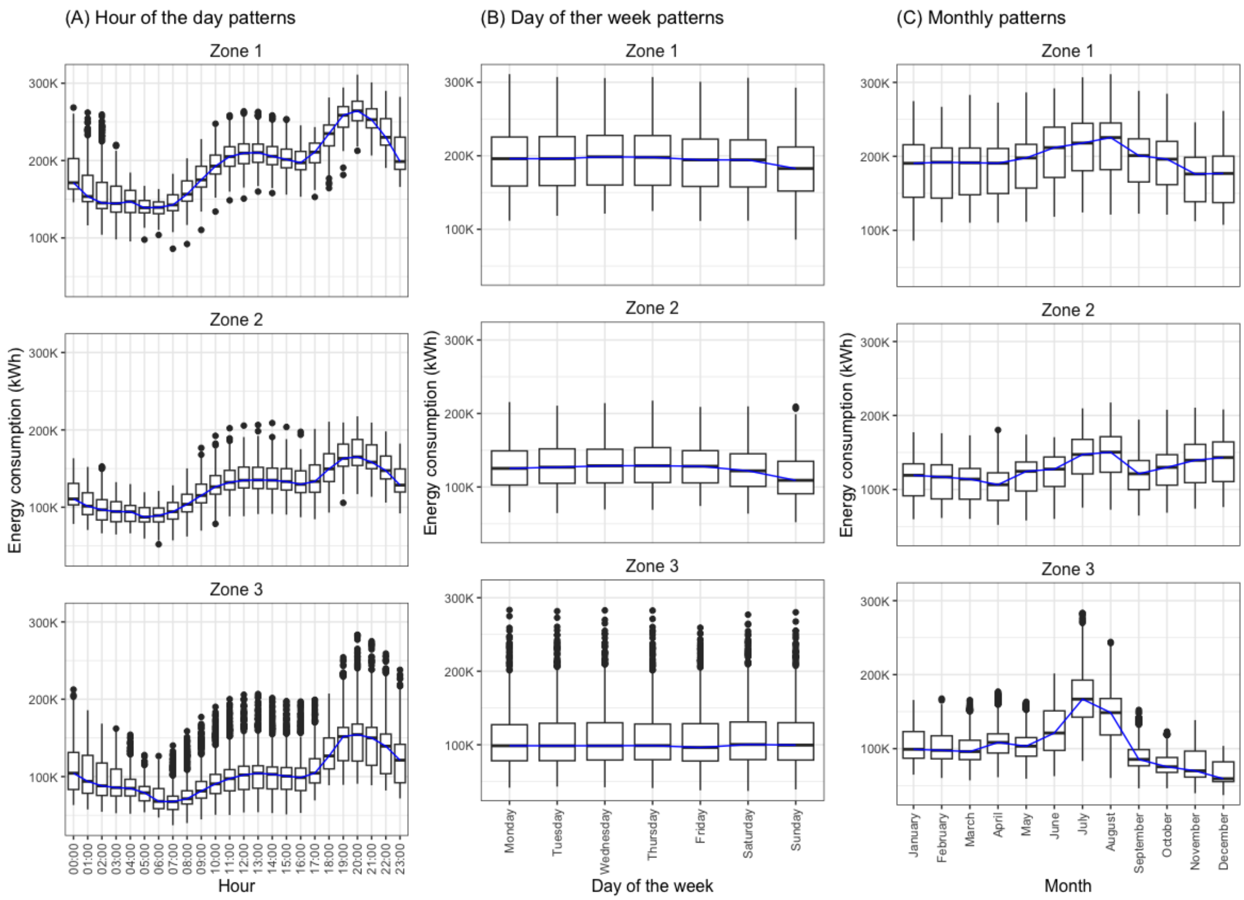
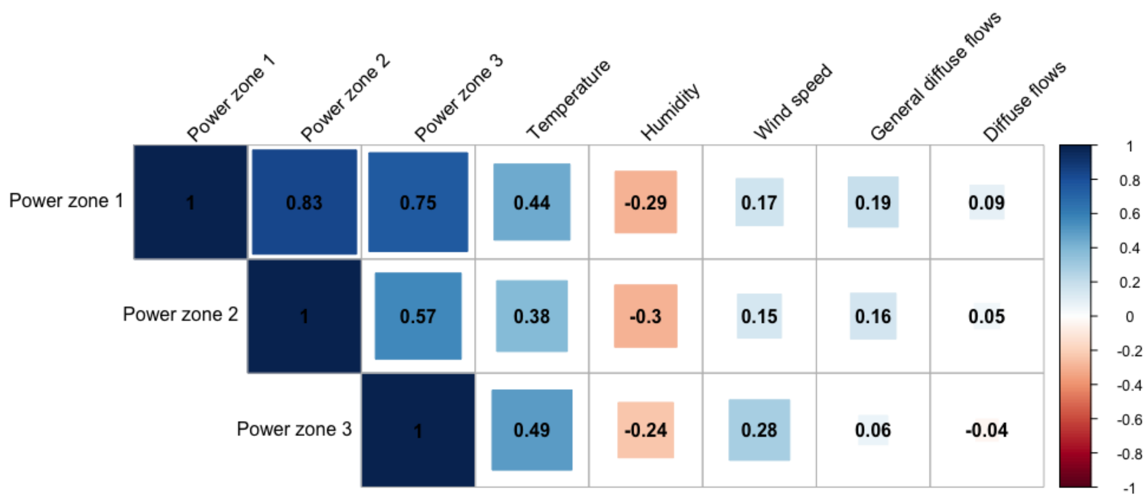


Figure 3: Pearson correlation between the energy consumption and the environmental variables.



Prediction Model

We split our hourly dataset into training and testing subsets for our analysis. The testing set comprises the final two weeks' data, totaling 336 observations (14 days, 24 hours per day), while the remainder forms the training set.

In the next section, we begin by applying a basic model to capture trend and seasonality. Subsequently, we fit an ARIMA model to the residuals of this initial model. We then advance to more complex multivariate models, focusing specifically on ARIMAX and VARX models.

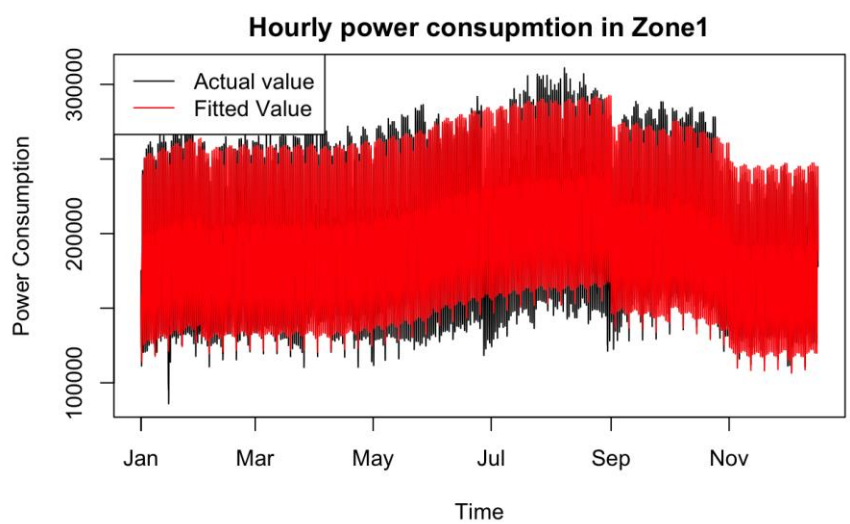
Univariate models

Trend-Seasonality models

We initially focus on modeling the power consumption trends in Zone 1, utilizing four distinct methods: Moving Average (MAV), Quadratic Polynomial (PARA), Local Polynomial (LOC), and Generative Additive Model (GAM) employing the Spline method.

As depicted in Figure B (see Appendix), the GAM model demonstrates superior efficacy in capturing the trend. Consequently, we have selected GAM to construct the trend-seasonality model, incorporating hourly, weekly, and monthly seasonal patterns. This decision is based on multiple experiments confirming the significance of all three seasonality. Figure 4 presents both the fitted data and the original series for comparison

Figure 4: Prediction from trend-seasonality model using a generative additive model (GAM)



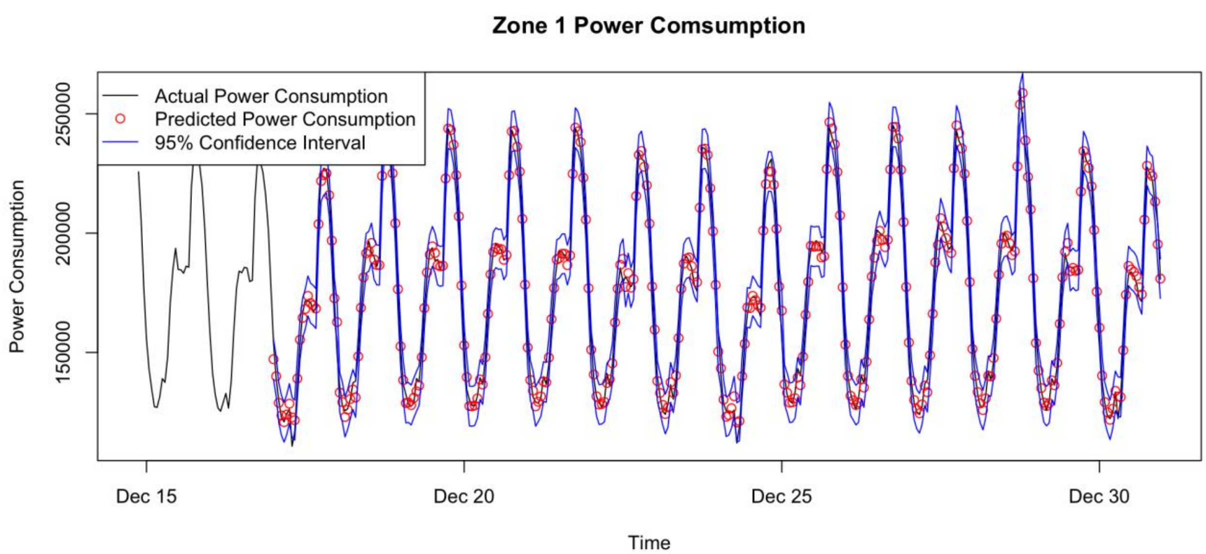
From the Figure C in Appendix, we can see that the residual of the trend-seasonality still shows a strong seasonality, so we consider applying the seasonal ARMA model to the residual process.

ARIMA models

We applied an ARIMA model to the residuals from the trend-seasonality model. The ACF plot of these residuals (Figure C in Appendix) revealed persistent strong seasonality, prompting the consideration of a seasonal ARMA model with parameters autoregressive (AR) and moving average (MA) parameters equal to 1 and seasonal AR and MA parameters equal to 1 (i.e., (1,0,1) * (1,0,1)) and a seasonality of 24. The residuals' ACF plot, along with that of the squared residuals, are depicted in Figure D in the Appendix. From the ACF plot, it is evident that weekly seasonality remains, although other lags appear adequately addressed. The squared residuals suggest the presence of some heteroskedasticity. However, due to the absence of a joint SARIMA and GARCH implementation in R, which limits predictive improvement, we did not pursue a GARCH model.

Subsequently, we employed rolling prediction with a step size of 1 to evaluate our model's accuracy. Figure 5 showcases the predicted power consumption along with the confidence intervals, compared against the actual data.

Figure 5: Prediction from SARMA and trend-seasonality model

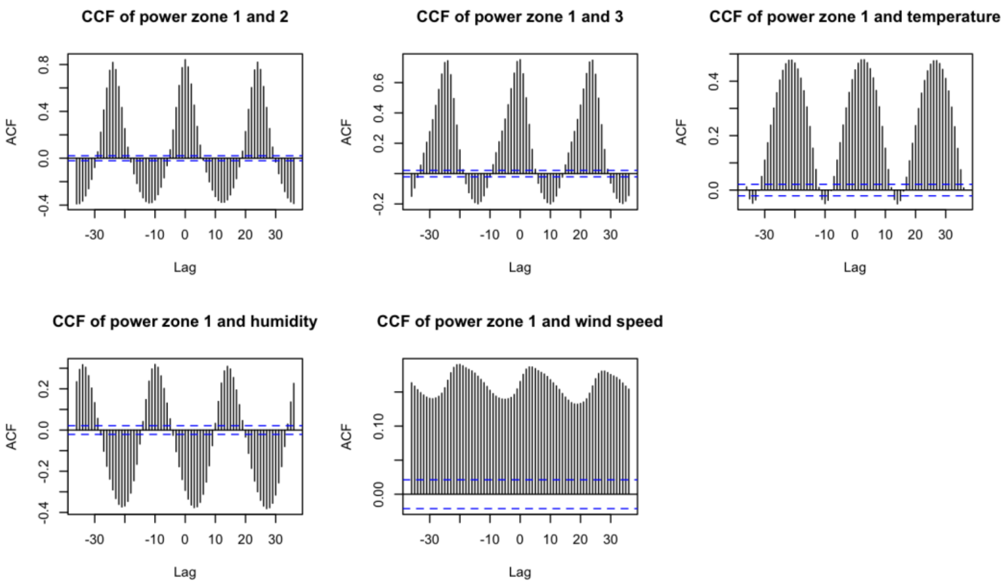


Multivariate models

ARIMAX models

In developing the ARIMAX model, we incorporated a range of exogenous predictors available in our dataset. These include power consumption in zones 2 and 3, temperature, humidity, and wind speed. To determine the most effective lag for each of these regressors, we conducted a cross-correlation function (CCF) analysis with the power consumption in zone 1. The results of this analysis are presented in Figure 6, as shown below.

Figure 6: Cross-correlation plot between exogenous regressor and power consumption in zone 1



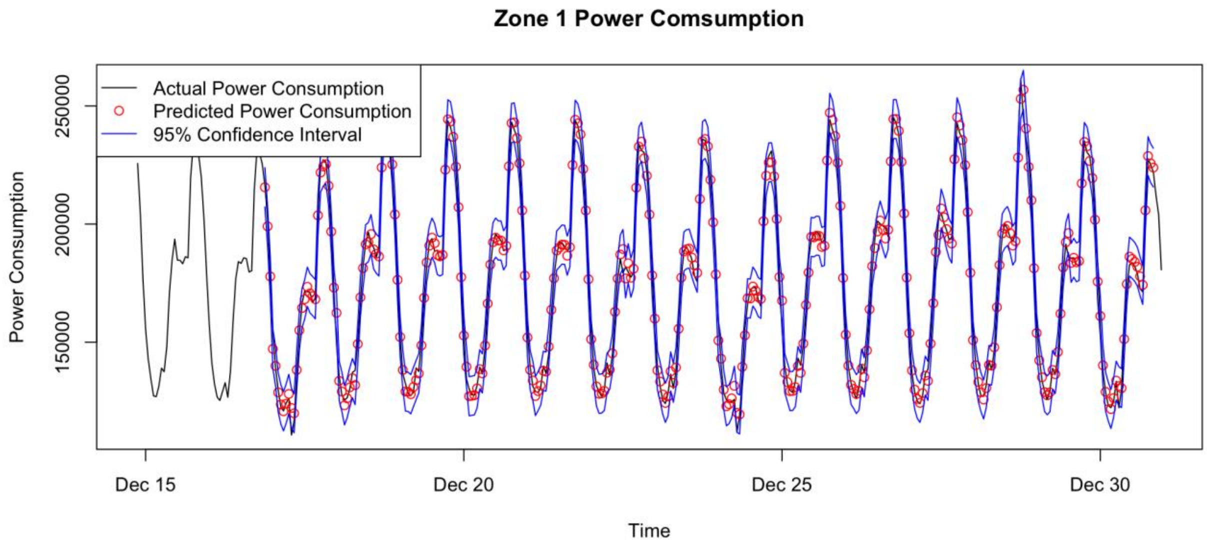
Analysis of the plot reveals that for zones 2 and 3, the correlation peaks at lag 0, indicating synchronous movement among these zones. However, for predictive purposes, using lag 0 data is not feasible. Therefore, we opt for lag 1 as the predictor. Regarding the three weather indicators, we observe that the correlation peaks at lag 3, suggesting that power consumption precedes changes in weather conditions. This correlation, however, should not be misconstrued as causality, as weather patterns are not influenced by power usage. A more logical interpretation is that power consumption is adjusted in anticipation of forecasted weather changes.

To integrate predictors into our model, we employed a forward selection method. Starting with a model devoid of external information, we sequentially introduced one external regressor at a time, choosing the one that yielded the lowest AIC in each iteration. Ultimately, the selected model includes one-step lagged data from zone 2 and three-step-ahead temperature data. The ACF plots for the residuals and squared residuals are presented in Figure E in the Appendix. The results of the rolling prediction, compared with actual values, are displayed in Figure 7.

VAR models

We built four Vector Autoregressive (VAR) models, VAR model with a maximum lag of 24, as incorporating more lags create models with a large number of predictors. The selected number of lags using the AIC indicator was 24, which results in a large model. Due to large number of predictors in the VAR model, we decide to restrict the VAR model to reduce the number of predictor variables. The restricted VAR model results in lower AIC indicating that it is a better model. The VAR model shows the factors of temperature, humidity, and general diffuse flows are all statistically significant in predicting energy consumption of all power zones.

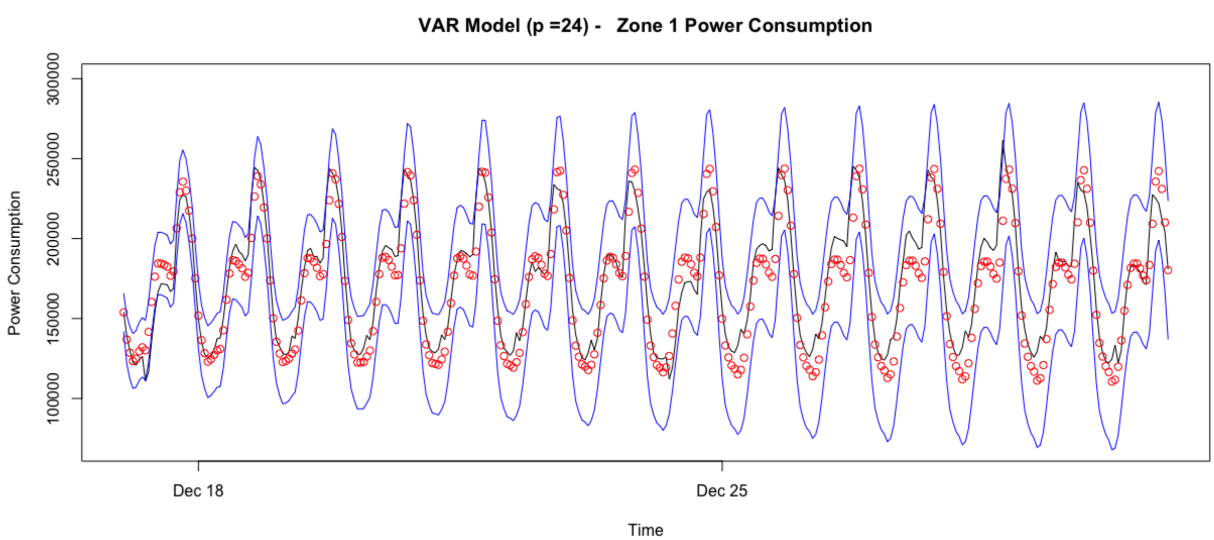
Figure 7: Prediction from SARMAX and trend-seasonality model



In the prediction plot of power zone 1 (Figure 8), the model accurately predicts the seasonality, capturing the vast majority of fluctuations. This accuracy with fluctuations is repeated in the power zone 2 (Figure F in the Appendix) and power zone 3 (Figure G F in the Appendix) models. The actual data stays within the 95% confidence interval for all power zones but cuts it close with power zone 3. Overall, the model appears to be most accurate predicting power zone 1.

While the predictions seem to be good, the VAR model fails multiple goodness of fit tests. Analysis shows the VAR model failing the constant variance test, normal distribution test, and serial correlation test.

Figure 8. Prediction of Power Zone 1 from the Restricted VAR model with a lag of 24



VARX models

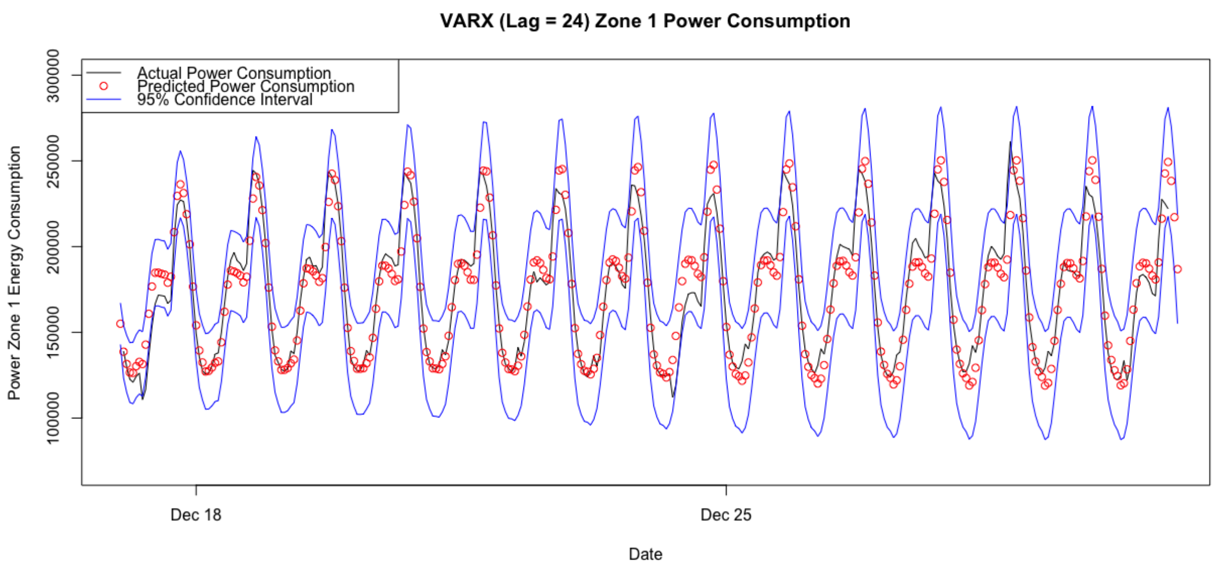
We fit Vector Autoregressive with Exogeneous Factors (VARX) models to predict energy consumption in zone 1 using the same exogenous variables as the ARIMAX model.

The VARX model also shows the factors of temperature, humidity, and general diffuse flows are all statistically significant in predicting energy consumption of all power zones.

Upon analysis of the results, we see similar results to the predictions using the VAR models. Despite the improvements in prediction accuracy, the VARX model also fails the constant variance test, the normal distribution test, and the serial correlation test. Therefore, there is likely to be some trend or seasonality or pattern unaccounted for and more analysis must be done.

Overall, the VARX perform slightly better than the VAR models. However, we believe the VARX model is a better model due to simplicity.

Figure 8: Prediction of Power Zone 1 from the VARX model with a lag of 24



Model performance

We calculate the MAE and the root mean squared error (RMSE) for each model for direct comparison, as seen in Table 1. The ARIMAX model outperforms all the other models, hence is selected as our final predictor. Such result is determined by the nature of ARIMAX model and the behavior of power user. Compared with VAR type of models, ARIMAX has the advantage of utilizing the shock information, which leads to better performance. For example, if there is a power outage in zone 1, only the ARIMAX model can quickly capture such shock information and update the prediction, which is due to its MA part. However, for VAR model, it will take several hours for the shock to be recorded by all lagged data.

Table 1: Mean absolute error (MAE) and root mean squared error (RMSE) for each model.

Model	ARIMA	ARIMAX	VAR (lag=24)	Restricted VAR (lag=24)	VARX (lag=24)
MAE	2,158.2	2,094.6	9,617.8	9,884.8	8,605.9
RMSE	2,999.9	2,905.2	12,727.5	13,036.4	11,427.7

Findings

The study of Tetouan's energy consumption provides results and predictions that can be used to improve energy management and environmental impact. Having accurate predictions of energy fluctuations enable proper and efficient allocation of resources such as manpower, equipment, and maintenance. The predictions can allow the city to save on costs of energy, as well as better plan for contingencies such as power losses (how much energy should be stored in reserve if possible) and power surges (how much energy the power grid needs to be able to handle). In addition to saving costs, better energy management improves efficiency and reduces energy waste. For power grids that rely on non-renewable resources such as coal, such improvements will reduce negative environmental impact.

The analysis showed the presence seasonal patterns in the data that aid in the forecast of future energy consumption, including hour-of-the-day and monthly seasonality. Additionally, including exogenous variables such as temperature, wind speed, and humidity improved model performance. Decision makers in this space can take advantage of the use of weather forecast to improve energy consumption predictions.

Based on the analysis, the ARIMAX model is the superior model which as it has the advantage of utilizing the shock information, which leads to better performance. Though the ACF plot for the residual and squared residual still shows some nonstationary behavior and hetero heteroskedasticity, we did not apply GARCH model. This decision was influenced by the GARCH model's computational complexity and its limitations in providing tangible, real-world interpretations.

In future investigations, more focused research on the sources of Tetouan's power consumption should be done. Doing so will further improve energy management and environmental impact, but most importantly provide a foundation for better government policies. Government policy can be better adapted to support energy efficiency through more targeted legislature. For example, Tetouan may consider placing renewable energy mandates, tax incentives, or new energy standards on industries that consume high levels of energy. In addition, such research would help in the development of new technologies or microgrids. New technologies such as cheaper renewable energy implementation can be used and microgrids for localized power consumption will be more efficient and can be positioned better.

References

1 International Energy Agency (2019) ‘Energy policies beyond IEA countries – Morocco’. Available at: https://www.iea.org/reports/energy-policies-beyond-iea-countries-morocco-2019

2 Power consumption of Tetouan City (no date) UCI Machine Learning Repository. Available at: https://archive.ics.uci.edu/dataset/849/power+consumption+of+tetouan+city (Accessed: 08 December 2023).

Appendix

Figure A: Time series of environmental factors aggregated in 1-hour intervals from January 1, 2017 to December 31, 2017.

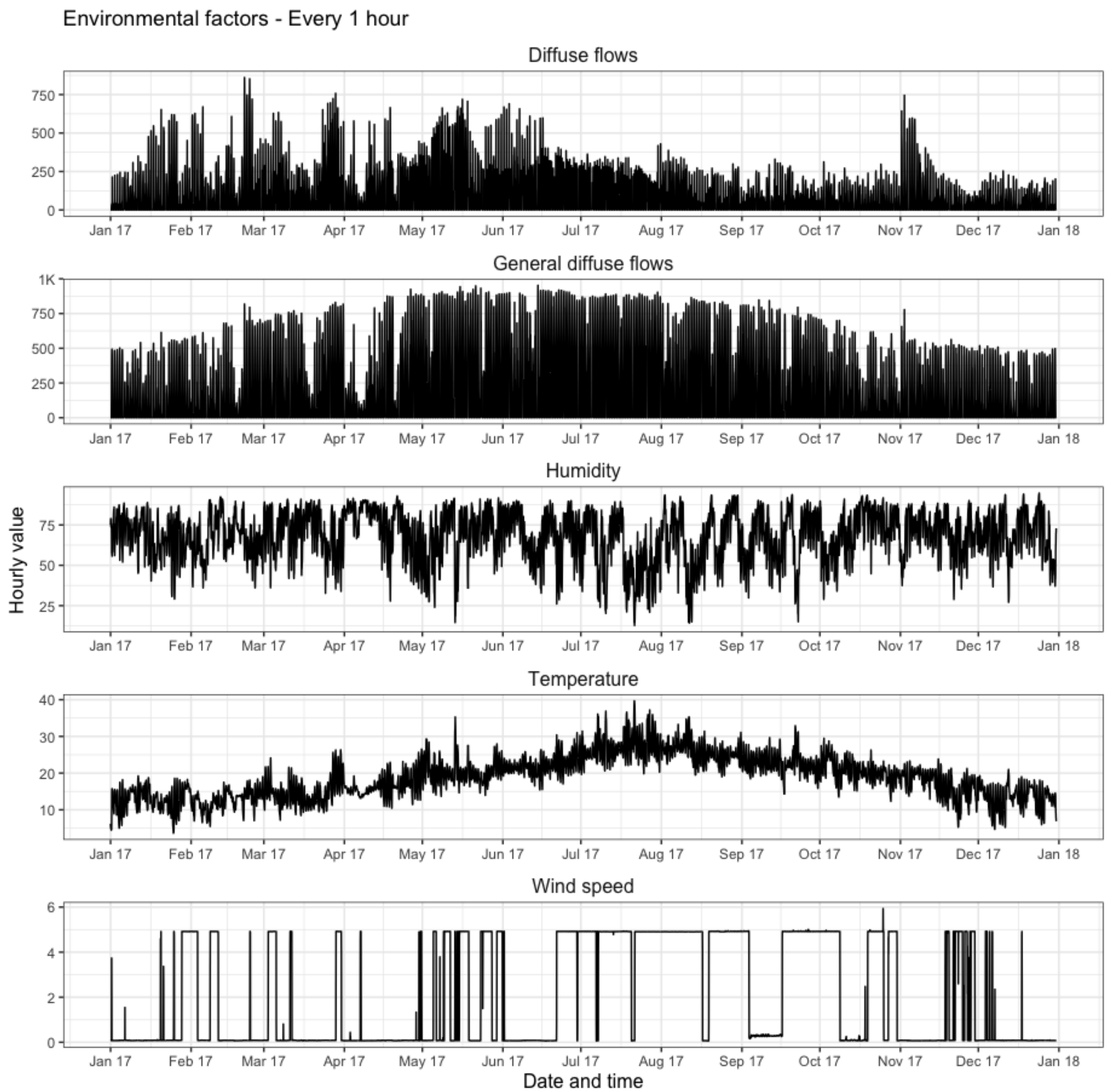


Figure B: Trend Estimation Using Four Methods and Actual Data

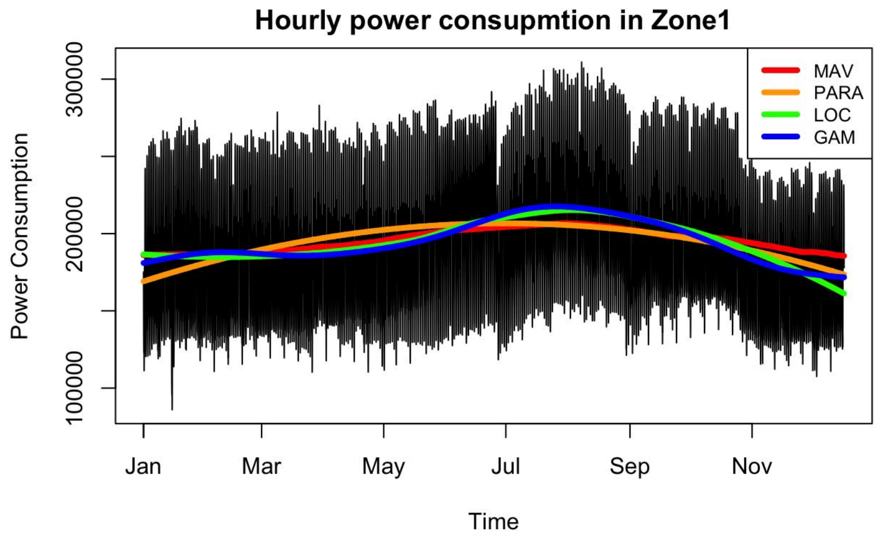


Figure C: ACF plot of the residuals of Trend-Seasonality model

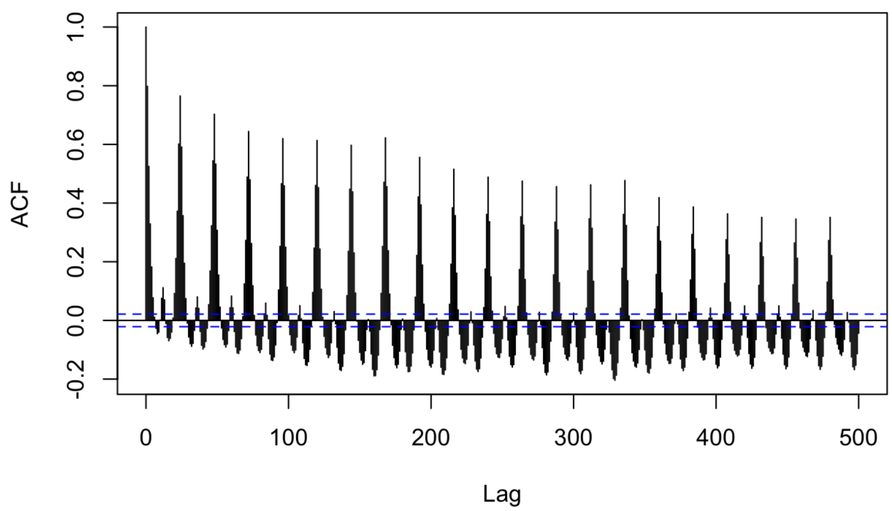


Figure D: ACF plot of the residuals and squared residuals of ARIMA model

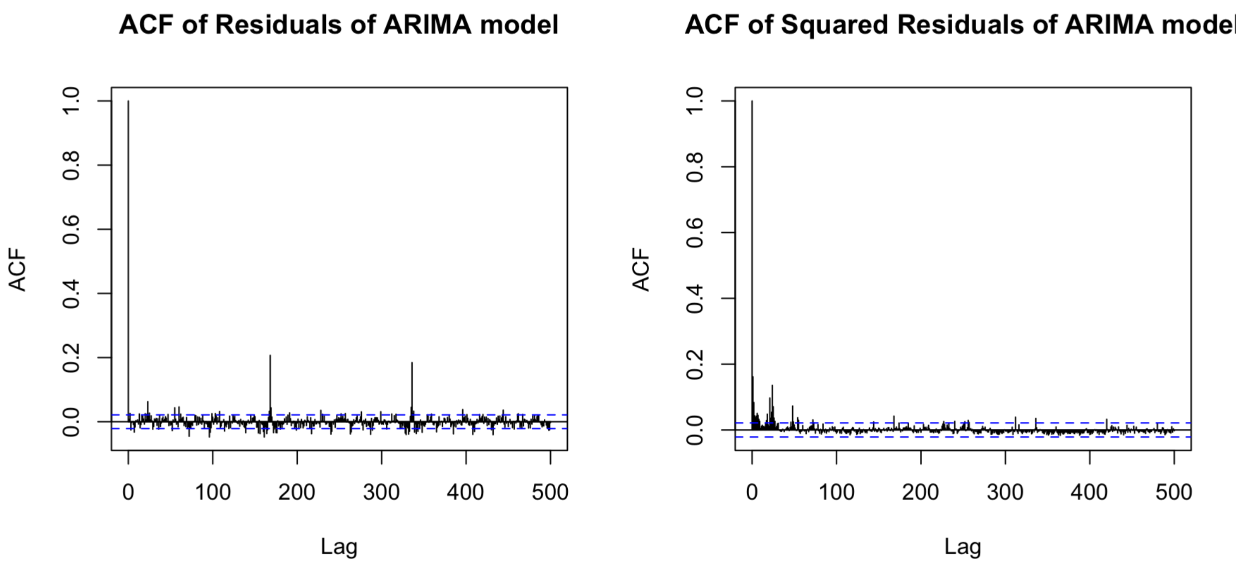


Figure E: ACF plot of the residuals and squared residuals of ARIMAX model

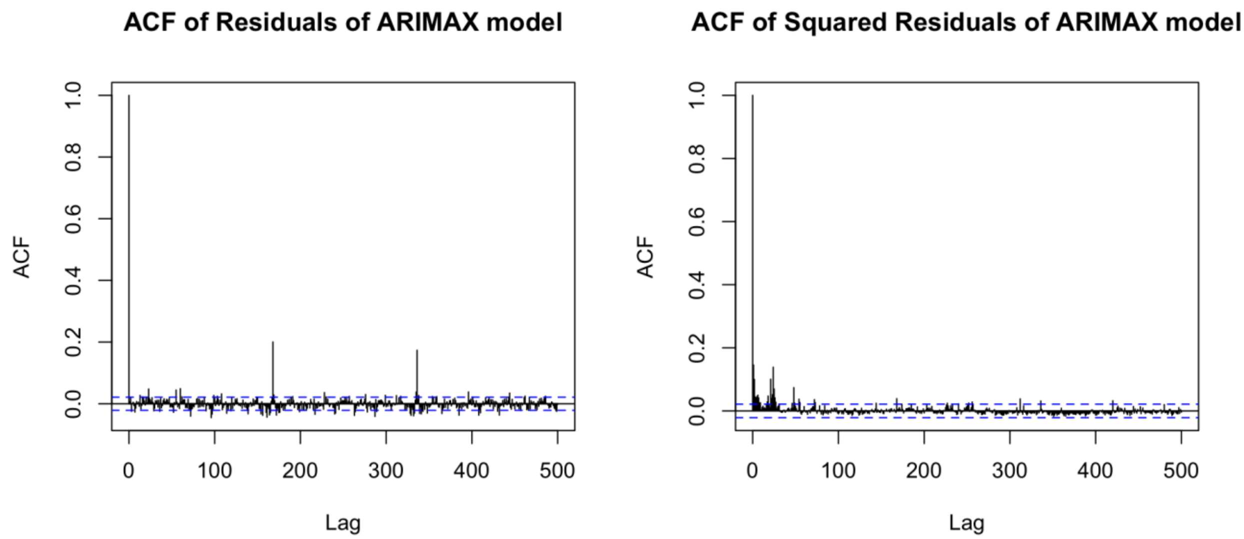


Figure F. Prediction of Power Zone 2 from the Restricted VAR model with a lag of 24

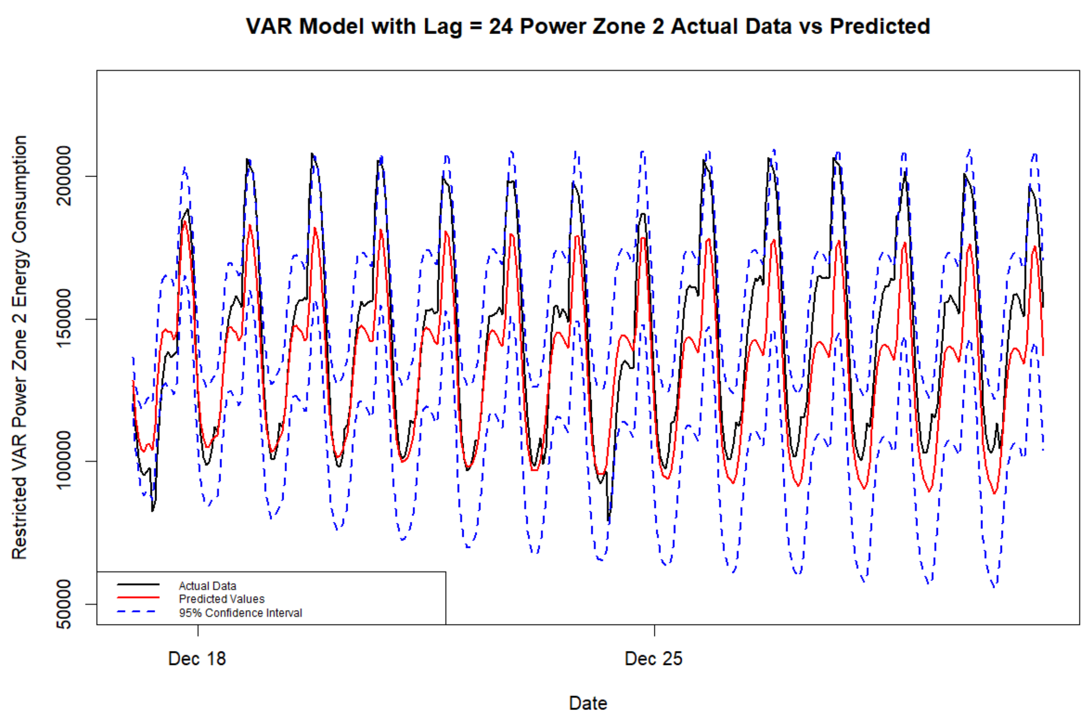


Figure G. Prediction of Power Zone 3 from the Restricted VAR model with a lag of 24

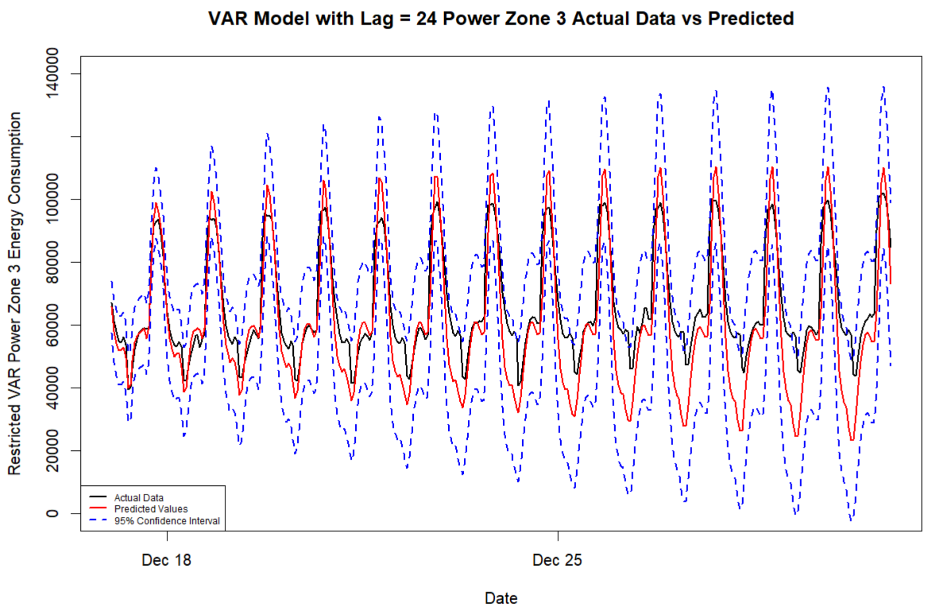


Figure H. Prediction of Power Zone 2 from the VARX model with a lag of 24

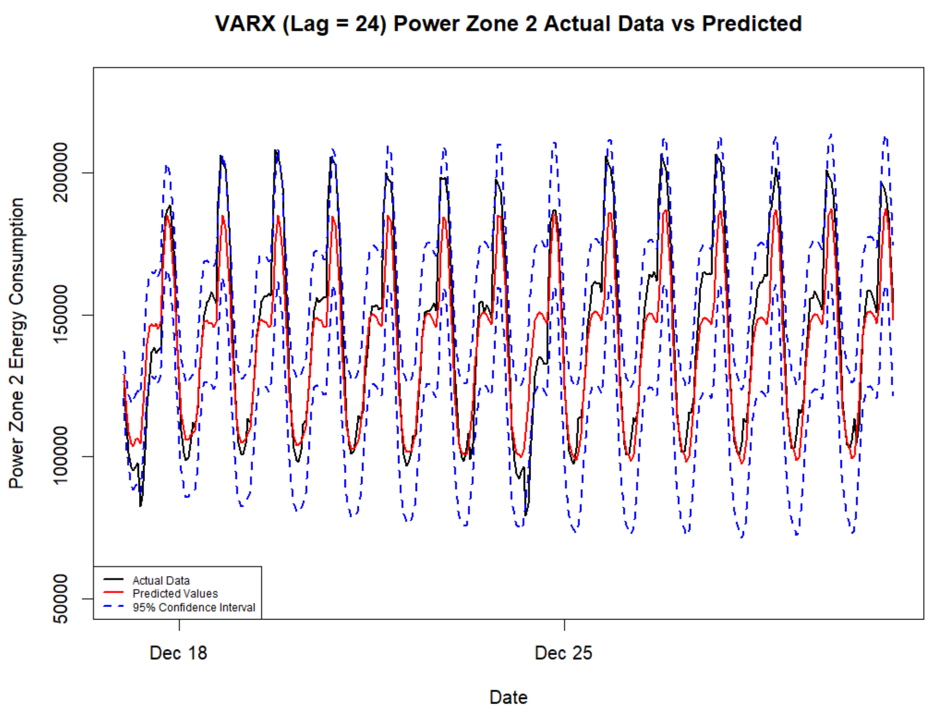
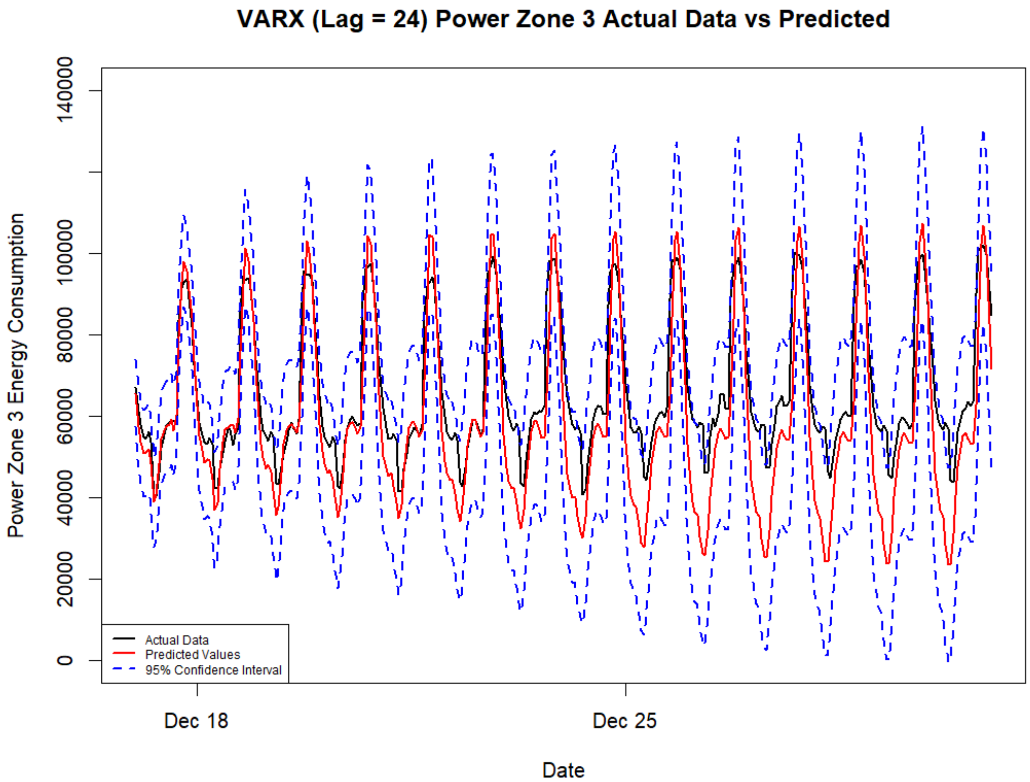


Figure H. Prediction of Power Zone 3 from the VARX model with a lag of 24



VARX (Lag = 24) Power Zone 3 Actual Data vs Predicted
Power Zone 3 Energy Consumption
0
20000
40000
60000
80000
100000
140000
Actual Data
Predicted Values
95% Confidence Interval
Dec 18
Dec 25
Date