Electricity Demand Forecast

GROUP E

Zijue Fang, 101716879
Thinh Tran, 903259
Linh Duong, 894025
Ha Do, 894601
Bùi Anh

Contributions of each group member:
• Bùi Anh did not contribute (0%).
• Other group members contributed equally.

Contents

1. Introduction

1.1 Problem description

Electricity demand-along with prices-has been a regular discussion point in the media. Concerns about cost and efficiency call for accurate forecasting of electricity demand in order to reduce cost and avoid unneeded greenhouse gasses emission. Electricity consumption is affected by numerous factors (i.e. weather, holidays, day of the week. etc.), thus making it hard to forecast correctly. With the help of current technological development, accurately predicting demand is made easier than before. For the assignment, the group is expected to create a demand forecast for the total electricity consumption in Finland.

1.2 Objectives and expected results

The goal of this project is to produce electricity demand forecasts for 3 separate days: 14.3, 16.3 and 20.3, using historical data of electricity consumption in Finland from the start of 2020 onwards. To increase the accuracy of the forecast, the analysis should be based on the weather data for Helsinki, Tampere and Rovaniemi from the start of 2020, date information such as weekends and holidays and possibly, Covid data as well. The result of the forecast should be the mean of the natural logarithm of the following day's consumption, and the standard deviation of the forecasted number.

2. Data cleaning and preparation

To commence the analysis, the total daily electricity consumption in Finland data, the weather data for Helsinki, Tampere, Rovaniemi, and the day-ahead weather forecasts for those three cities were first employed. All of the data tables were initially checked to see if there were any missing values. The data types were also taken into account to ensure proper handling and manipulation of the information throughout the preprocessing process.

The first data table included 55,348 entries and three columns: Start Time UTC, End Time UTC, and Electricity Consumption in Finland. The data spanned from 31st of December, 2019 to 29th of February 2024, covering multiple hourly intervals. However, since 2024, the intervals have been changed into either hourly or every fifteen minutes. In order to transform the raw data into a structured format suitable for analysis, we first converted the date time data from character strings to datetime objects. The dataset was then grouped by date, and the total daily consumption was calculated by multiplying the mean consumption of each day by 24 hours. The total consumption column was then log_transformed and stored in a new column called log_value. The purpose of this transformation is to stabilize variance and make it more suitable for data modeling. Finally, the processed data was converted into a tsibble.

For the weather observations datasets of cities around Finland, there are three separate tables which contain the mean temperature in every one hour of Helsinki, Tampere, and Rovaniemi respectively. The total entries were 36 452, 35,057, and 36 370 respectively. The date range is from the beginning of 2020 to 29th of February 2024. For each table, the date column was generated by concatenating the Year, Month, and Day columns and converting them into date format. The daily mean temperature for each city was calculated and the data for Helsinki, Rovaniemi, and Tampere were merged into a single dataset using the date column. The merged weather data was also converted into a tsibble and then merged with the previously preprocessed electricity consumption data using date column. In order to fill the missing values caused by the gaps in the time series data, the fill_gaps function was applied to the combined dataset.

The day-ahead weather forecast table contains the temperature in every six hours of the three cities Helsinki, Tampere, and Rovaniemi. There are 375 observations in total. After the forecast date was converted into date format, the forecasted temperature for Helsinki, Tampere, and Rovaniemi were grouped by date and the daily average temperature were calculated. Three separate aggregation operations were performed for each city and the values were stored in three different variables which are ‘daily temperature Helsinki’, ‘daily temperature Rovaniemi’, and ‘daily temperature Tampere’. Later on, the aggregated daily temperature of three cities were merged into a complete table based on the date column. Finally, the table was converted into a tsibble.

3. Initial forecasting approach

3.1 Visual analysis

We initiated our analysis by visualizing the log-transformed electricity consumption data against time. The time series plot (Figure 1) indicated potential seasonality, given the recurring patterns at regular intervals. This aligns with our expectations for electricity demand which can be influenced by factors like seasonal weather changes and holidays. To encapsulate potential trends and seasonality in the data, we initially explored a Time Series Linear Model (TSLM).

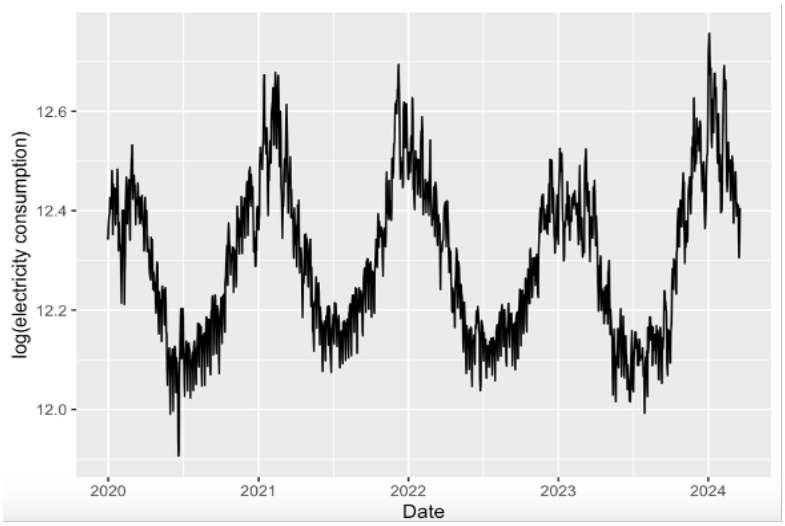

Figure 1: log-transformed electricity consumption

3.2 Evaluation of the initial choice

Our TSLM incorporated both trend and seasonality components. The model's fit was less than ideal, as indicated by several seasonal coefficients not being statistically significant and a low adjusted R-squared value, suggesting a poor model fit (Table 1).

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	1.23E+01	1.33E-02	923.841	< 2e-16 ***
trend()	1.46E-05	1.27E-05	1.156	0.24805
season()week2	2.24E-03	1.58E-02	0.142	0.88727
season()week3	-4.75E-02	1.58E-02	-3.01	0.00267 **
season()week4	-7.15E-02	1.58E-02	-4.532	6.45e-06 ***
season()week5	-3.02E-03	1.58E-02	-0.191	0.84827
season()week6	3.42E-04	1.58E-02	0.022	0.98273
season()week7	1.57E-03	1.58E-02	0.1	0.92053
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
Residual standard error: 0.1433 on 1147 degrees of freedom				
Multiple R-squared: 0.03751, Adjusted R-squared: 0.03164				
F-statistic: 6.386 on 7 and 1147 DF, p-value: 2.1862e-07				

Table 1: summary statistics of TSLM model

Further inspection of the residuals from the TSLM (Figure 2) revealed a lack of randomness, with clear patterns suggesting unaccounted seasonality. Additionally, the autocorrelation function (ACF) plot shows that most autocorrelation bars extend beyond the blue dashed significance boundaries, signaling the presence of autocorrelation within the residuals. Moreover, the distribution of the residuals, as depicted in the histogram, deviates from normality.

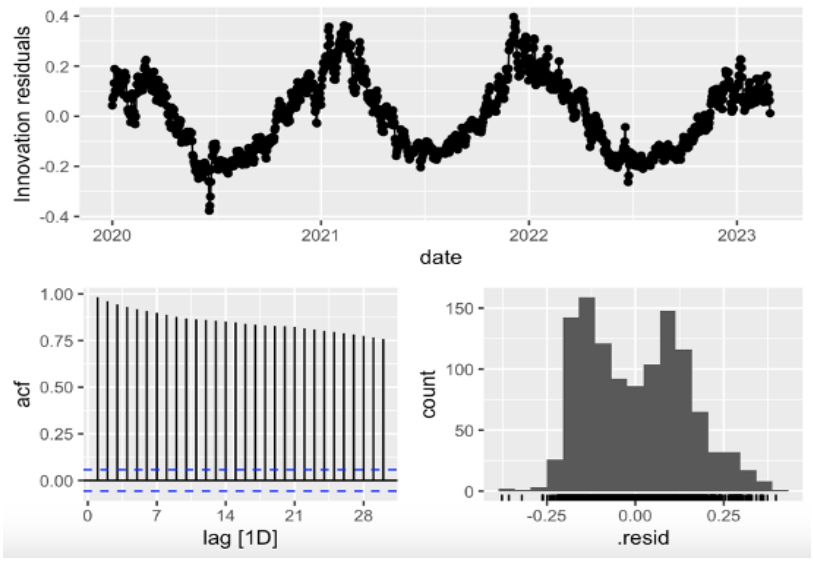

Figure 2: residuals plots of TSLM model

The one-year-ahead forecast plot generated from TSLM (Figure 3) indicated that the model's prediction intervals were wide and expanded over time, reflecting high uncertainty in the forecasts.

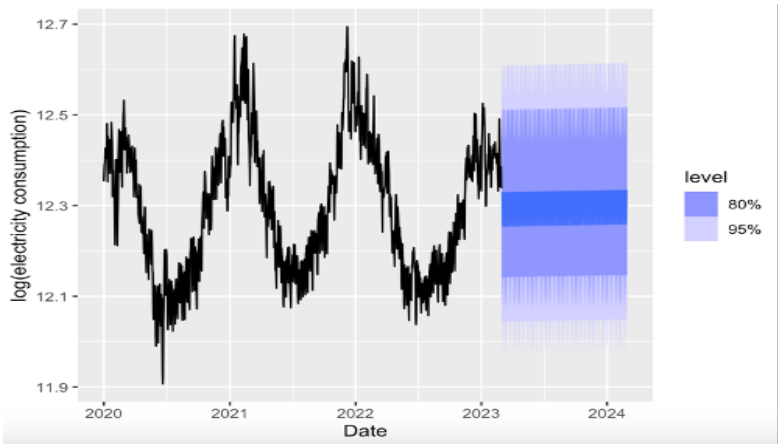

Figure 3: forecast plot under TSLM model

3.3 Transition to Exponential Smoothing Model (ETS)

Considering the limitations of TSLM in capturing the data's inherent patterns, we transitioned to an Exponential Smoothing State Space Model (ETS), which is better suited for modeling non-linear trends and seasonality more effectively. We chose an ETS model with an additive error (A), no trend (N), and an additive seasonality (A), denoted as ETS(ANA), to better accommodate the apparent seasonality. The choice of ETS(ANA) was driven by the lack of a significant trend in the TSLM model and the clear seasonality observed in the historical data.

3.4 Model validation and final choice

The residuals from the ETS(ANA) model (Figure 4) displayed a more random pattern, which suggests that the model is accounting for the underlying patterns in the data reasonably well. Besides, most of the autocorrelation bars lie within the blue dashed confidence intervals, indicating that there is no sign of autocorrelation in the residuals. The histogram of residuals seems to be normally distributed.

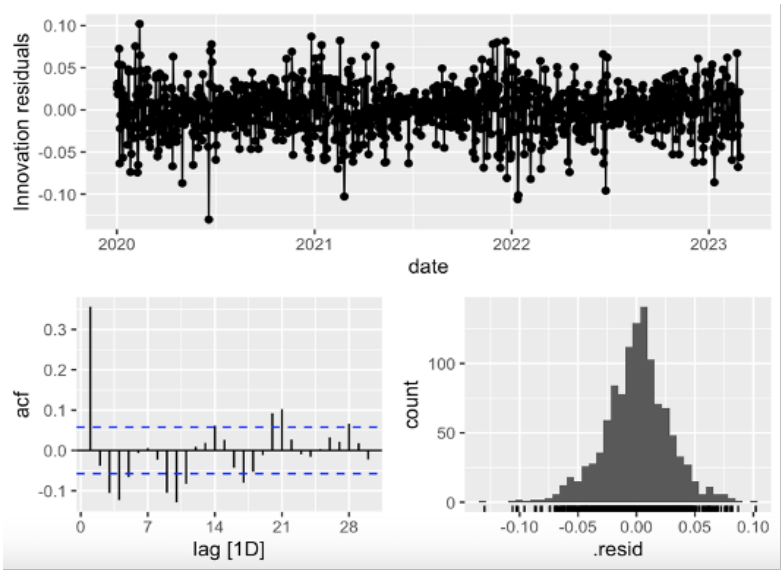

Figure 4: residuals plots of ETS(ANA) model

The ETS(ANA) model's forecast (Figure 5) showed narrower prediction intervals, which implies increased confidence in the forecasts and suggests that this model is more adept at understanding and predicting the underlying pattern in electricity consumption.

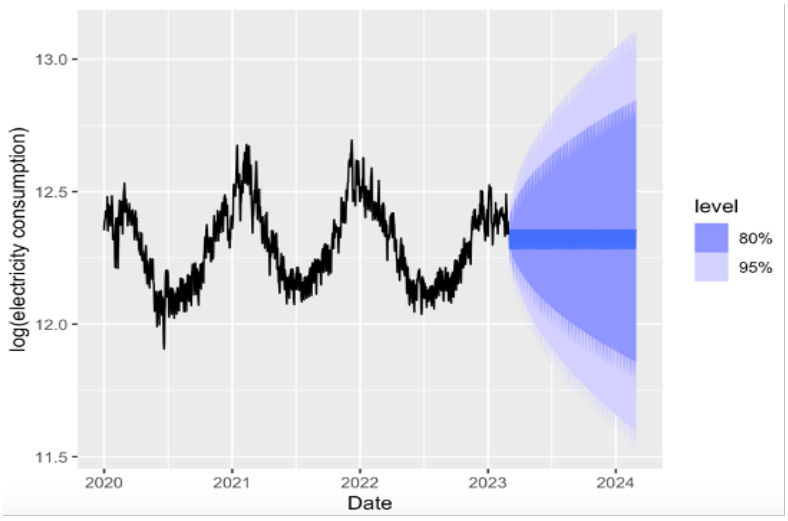

Figure 5: forecast plot of ETS(ANA) model

In conclusion, after careful evaluation, the ETS(ANA) model was chosen as the superior approach for our electricity demand forecasting, striking a balance between complexity and forecasting performance.

4. Adaptation of forecasting approach

4.1 Second Round: Incorporation of weather data

The first round model is ETS(ANA) that does not include the weather information from three cities in Finland. To forecast the electricity more accurately, the incorporation of such data should be considered. A TSLM model of electricity consumption as dependent variables and temperature data in three cities as independent variables is conducted to examine if temperature information significantly affects the electricity consumption.

```
Model: TSLM

Residuals:
     Min       1Q   Median       3Q      Max
-45849.4  -9189.2   -201.5   9300.4  43637.6

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          242309.3      765.6 316.487  < 2e-16 ***
helsinki_temperature   -2105.4      268.2  -7.849 7.97e-15 ***
tampere_temperature     -403.6      269.7  -1.497    0.135
rovaniemi_temperature  -1036.4      102.0 -10.165  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14220 on 1480 degrees of freedom
Multiple R-squared: 0.8413,     Adjusted R-squared: 0.841
F-statistic:  2615 on 3 and 1480 DF, p-value: < 2.22e-16
```

Table 2: TSLM Model Summary Statistics

From Table 2, there is a negative relationship between the electricity consumption and weather data. Two coefficients of Helsinki and Rovaniemi temperature are significant; Adjusted R-squared is relatively high. It’s decided that the temperature data should be included into the second round model.

Because ETS(ANA) model does not directly allow for the inclusion of extra variables, ARIMAX, an extension of the ARIMA (AutoRegressive Integrated Moving Average) model that incorporates external or exogenous variables into the time series forecasting process, is used. In the ARIMAX model, the forecasts are based on the past values of the time series itself and external factors. In our model, we run the ARIMAX model for the log of electricity consumption as the time series and the temperature from three cities as exogenous regressors.

Although the ARIMAX model does not outperform ETS(ANA) across error measures (Table 4), we believe that weather information is relevant to the forecasting task. From Figure 6, the forecast seems to be rational based on past consumptions.

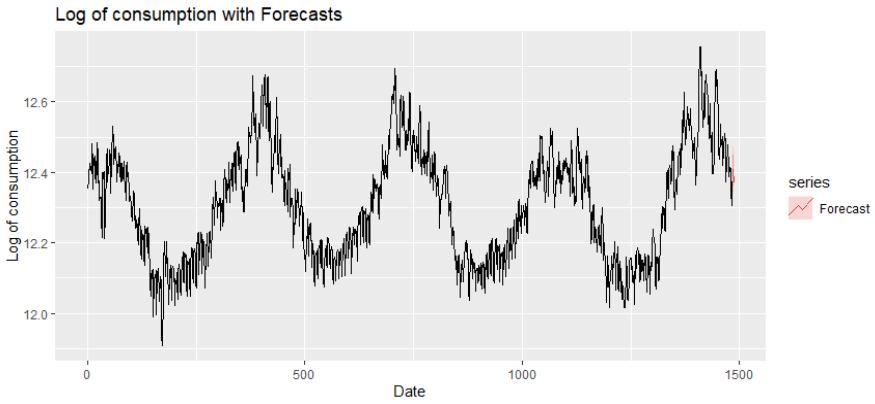

Figure 6: forecast plot of ARIMAX model

4.2 Third Round: Final adjustments, incorporating all previous learnings and data analysis

In the final round, we decided to continue with the ARIMAX model due to the model's ability to incorporate extra variables directly. We theorized that the day of the week would have a significant impact on electricity consumption, our thought process was: we consider "weekend" as Saturday and Sunday, if it's the weekend, the electricity demand would be lower since office and public spaces would be either not in use, or close earlier, thus require less electricity for heating. We tested this theory using the TSLM model, we adapted the TSLM model from the previous round and added "weekend" as an independent variable.

```
Model: TSLM

Residuals:
     Min       1Q   Median       3Q      Max
-37584.6  -9328.8   -507.1   8565.4  42007.8

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             246133.12     738.94 333.089   <2e-16 ***
helsinki_temperature     -2279.75     246.35  -9.254   <2e-16 ***
tampere_temperature       -225.61     247.70  -0.911    0.363
rovaniemi_temperature    -1054.91      93.56 -11.275   <2e-16 ***
Weekend.x               -12535.43     750.89 -16.694   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13050 on 1479 degrees of freedom
Multiple R-squared: 0.8664,     Adjusted R-squared: 0.8661
F-statistic:  2399 on 4 and 1479 DF, p-value: < 2.22e-16
```

Table 3: TSLM Model Summary Statistics with Weekend

From Table 3, it can be seen that there is a clear negative relationship between the electricity consumption and the "weekend" data. Out of the previously incorporated variables, "weekend" data had the highest coefficient; Adjusted R-squared is higher than the previous round's analysis. Thus it's decided that "weekend" data should be included into the model for the final round.

5. Performance Comparison

This section includes the analysis of the forecast accuracy across the submission rounds using their error metrics and discussion of the model's strengths and weaknesses.

Model	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
ETS(ANA)	2.84E-05	0.0282	0.0211	-1.28E-04	0.171	0.437	0.357
ARIMAX (weather data)	-4.480692e-06	0.0344190a6	0.02807488	-0.0007187151	0.2284978	0.8658139	0.05487988
ARIMAX (weather + weekend)	5.133971e-05	0.02416561	0.0181131	-2.244855e-06	0.1473396	0.5585981	-0.003990767

Table 4: Error metrics of models for each round

To determine the best model, we look for lower values of error metrics such as RMSE, MAE, MAPE, and MASE, and closer to zero values for ME, MPE, and ACF1. ARIMAX (weather + weekend) has the lowest RMSE, MAE, and MAPE among the three models, suggesting better accuracy in predictions. ETS(ANA) has the lowest MASE, followed by ARIMAX (weather + weekend), and then ARIMAX (weather data). Based on these comparisons, ARIMAX (with weather + weekend data) appears to be the best-performing model overall, as it has the lowest RMSE, MAE, and MAPE, and a relatively lower MASE compared to the other models.

Next, we proceed to provide strengths and weaknesses among the three models: ETS(ANA), ARIMAX (weather data), and ARIMAX (weather + weekend).

Model	Strengths	Weaknesses
ETS(ANA)	● simple and easy to interpret, suitable for quick analysis and forecasting ● automatically capture and adjust for seasonality in the data	● do not explicitly incorporate external variables or complex relationships between variables ● limited ability to capture all relevant factors influencing the time series.

ARIMAX (weather)	• include external variables (weather data) => can improve forecast accuracy • capture complex relationships between variables, including nonlinear and lagged effects	• require sufficient historical data for both the target variable and the exogenous variables
ARIMAX (weather + weekend)	• include additional variables such as weekend indicators => can capture more of the underlying patterns and improve forecast accuracy • Weekend indicator can reduce residual autocorrelation and improve the overall fit.	• Include additional variables => increase the risk of overfitting the model to the training data, which can lead to poor performance on unseen data

Table 5: Strengths and weaknesses of models for each round

6. Reflections on potential areas for improvement

- We could consider using other modeling techniques such as machine learning or time series neural networks as well as other data sources including population demographic, economics factors, and policies to further enhance future forecasts.
- We could consider assigning weights to the temperatures from these cities. The weights would reflect each city's relative importance based on factors such as population size, economic activity, policies, incentives, or their contribution to national electricity consumption.
- Our analysis reveals a negative correlation between temperature and electricity consumption. This insight may prompt us to focus on the minimum temperature as a key variable, rather than the average temperature.
- Our data did not account for other variables such as important holidays or events, considering the weekend has a fairly significant impact on the forecast, it is projected that these variables would have similar effects on the forecast.