TU DORTMUND

CASE STUDY

Project: Forecasting the electricity load Approach using linear models and seasonal filtering

Lecturers:
Prof. Dr. Matei Demetrescu,
Dr. Paul Navas

Author: Mukta Ghosh

Matriculation no: 231720

October 24, 2025

Contents

1 Introduction

The evolution of the electricity market reflects broader global trends towards sustainability and renewable energy adoption. Germany, as one of the leading economies in the whole of Europe, has been at the forefront of this evolution, implementing aspiring energy policies aimed at fostering a more sustainable energy landscape. Accurate prediction of electricity consumption allows providers to optimize generation, plan investments in infrastructure, and meet consumer demand beneficially.

This report uses linear and seasonal forecasting techniques to develop and evaluate forecasting models for one-hour-ahead electricity load data in Day Time Saving(DST) on the German market. Our analysis encloses the period from January 1st, 2015 to March 15th, 2024. The raw data is sourced from (ENTSOE, 2015), the European Network of Transmission System Operators for Electricity, which is the association for the cooperation of the European transmission system operators (TSOs).

This report systematically explores the load series' temporal structure, inspecting serial correlation and variance patterns through Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots which provide crucial insights into the seasonal and trend components of the data. According to information criteria, Akaike information criterion (AIC), optimizing 10 minimum AIC-based Autoregressive models with a simulation over one year of data will give fruitful insights for base AR model selection. Then selected 10 models run over full sample data and finalize one best-fitted AR model with the lowest AIC and Residual Standard Error (RSE). Next, integrate the best external predictors that enhance forecasting accuracy to predict a one-hour-ahead load, using linear predictive models and the forward selection method. Furthermore, the utility of deseasoning techniques is explored to remove periodic fluctuations from the load series, thereby refining our forecasting models and improving predictive performance.

The second section provides a more detailed overview of the dataset, data quality, data preprocessing and description of variables. The necessary statistical methods are presented and explained in the third section. In the fourth section, the statistical analysis, and interpretation of the results are presented. Finally, in the fifth section, the main findings of this project are summarized.

2 Problem statement

2.1 Description of the dataset

The data used for analysis is provided by TU Dortmund, a Case Study program that is sourced from ENTSO-E, launched in 2015. It's a Central collection of Electricity generation, transportation, and consumption data for the Pan-European market (ENTSOE, 2015).

The provided dataset encompasses the period from 2015 to 2024, with observations collected from over 30 countries across Europe. According to the aim of this project, only data related to Germany is separated for analysis. There are a few hourly time-series datasets and one is the Load dataset that is needed to predict one hour ahead. Other datasets will be used as external datasets corresponding to the average temperature in Germany, carbon and CO2 prices, Imports and exports data, as well as day-ahead forecasts for wind and solar generation, allocated transfer capacities, and maintenance schedules. The day-ahead datasets are adjusted to hourly time series by duplicating day values for 24 hours. Data quality for all datasets is very rough as these are real datasets.

2.2 Preprocessing of the dataset

2.2.1 DST adjustment

As only data related to Germany have to be accounted for, data should be DST adjusted by adjusting the electricity consumption data for the spring forward and fallback changes. During the spring forward transition, the clocks are set forward by one hour, resulting in a missing hour in the dataset, and during the fallback transition, the clocks are set back by one hour, resulting in a duplicate hour in the time series dataset. The spring forward adjustment involves duplicating the load from the previous hour, while the fall back adjustment removes the duplicated hour, thereby conserving the integrity of the dataset.

2.2.2 External Dataset preprocessing

Among other external predictors, certain variables, such as the future price of $carbon$, are recorded on a daily basis. To align with our target variable, $Total.load$, which is recorded hourly, the data for these other predictors is adjusted to an hourly frequency. For actual generation per type dataset, there have output and consumption aggregated data for different gen-

eration type. This dataset typically contains information about the production of electricity categorized by different types, such as wind, solar, nuclear, fossil fuels, etc. Also, the consumption data provides information about the usage or consumption of electricity by different sectors. For predicting one hour ahead of load data, only output data is accounted for. Generation output have in total 17 columns with different subcategories which are categorized as $Fossil$, $Hydro$, $Solar$, $Wind$, $Waste$ and $Other$ for working convenience.

The allocated Transfer Capacity dataset gives insights on the maximum amount of electricity that can be transferred between two different countries over an interconnection line. More specifically, how much capacity Germany and other countries have to transfer Load by one to one. For convenient, data is aggregated per hour which gives information about net capacity of transfer per hour. Another dataset Net Physical Flow, is basically import and export load data per hour for different countries. By aggregation one column is created to show net export load per hour. Maintainance dataset have available and installed capacity per generation type. Again for convenience all available and installed capacity is aggregated and $total\ install\ capacity$ and $total\ available\ capacity$. Another dataset gives data about the forecast of wind and solar power generation (MW) per bidding zone, per each market time unit of the following day. Carbon Future Price and Temperature data is also accounted for external predictors.

2.3 dataset Quality

Most of the selected dataset's quality to predict one hour ahead load, are not clean enough. All dataset have in total 81096 observations. Several columns have NA or 0 values which are omitted. Missing data for certain days are filled by using the last available value.

2.4 Project objectives

The primary objective of this project is to create precise one-hour-ahead electricity load predictions for the German market. This will be achieved by employing of linear and seasonal forecasting models, considering various predictors. The performance of the fitted models will be assessed using information criteria, and the mean squared forecasting error (MSFE) will be calculated to find out forecasting accuracy. Additionally, the project aims to identify the most promising predictors through a selection process and forward selection method. The results obtained from these methods will be compared to those of the base autoregressive (AR) model. Finally, the models will be fitted with decomposed load data and promising predictors, and their performance will be evaluated using information criteria.

3 Statistical methods

In this section various statistical methods are represented that will be used to analyze the provided dataset according to the objectives of this report. To perform analyses and visualizations the statistical software R, version 4.2.3 (R Development Core Team, 2023) is used.

3.1 Simple and Multiple Linear regression

Linear regression aims to model the linear relationship of explanatory variables $x_1, ..., x_k$ on Y of primary interest. Y is also called a response or dependent variable and the explanatory variables are called covariates, independent variables, or predictor variables. For a single predictor variable, simple linear regression is a straightforward approach for predicting a response. It represents an approximate linear relationship between X and Y (James et al., 2013, p.61). Mathematical formula for single linear regression,

$$Y = \beta_0 + \beta_1 x + \epsilon$$

where, x = independent or predictor variable and Y = dependent variable.

Generally, we often have multiple predictor variables in a linear model. Extending the simple linear regression model is the practical approach rather than multiple simple linear models for each predictor. For this, a separate slope coefficient for each predictor is accounted for in a single model. Multiple linear regression is a statistical method that predicts the outcome of a dependent variable by using several independent or predictor variables(James et al., 2013, p.71).

A mathematical formula for multiple linear regression,

$$Y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_K + \epsilon$$

where Y = dependent variable, β_0 = intercept, β_k = slope coefficient, X = independent or predictor variable and ϵ = residual/model error

For all observations $i = 1, 2, \ldots, n$ of the response variable, as well as the error term into separate column vectors, $\mathbf{Y}$ and $\boldsymbol{\epsilon}$, and the explanatory variables expressing as a design matrix denoted by $\mathbf{X}$, each row corresponds to an individual observation (Fahrmeir et al., 2013, p.74).

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix} = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{bmatrix}$$

Hence, the linear regression formula can be demonstrated in matrix notation as follows,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

3.2 Autoregression (AR)

Autoregression is a process, using a linear combination of past variable values to calculate the coefficients of interest, denoted as AR(p) where p is the order of the autoregressive process. Here, ϵ_t will be white noise that are with a mean of 0 and variance of σ^2 like a multiple regression. (Chen, 2022) A mathematical formula for AR regression,

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \epsilon_t$$

3.3 Ordinary Least Square (OLS)

In the OLS model, the historical data is considered as a sample and the combined past and future data is considered as a total. Thus, the sample is used to estimate the total. The OLS model is combined with the theoretical analysis of the one-dimensional linear regression model.

3.4 Estimation of parameters

The derivation of the true population values, β and σ^2 is nearly impossible. For estimating, several methods can be used. One method is the least square method that is applied to find the estimates of them, written as $\hat{\beta}$ and $\hat{\sigma}^2$ (Fahrmeir et al., 2013, p.77). The least-square approach can be expressed as follows:

$$LS(\beta) = \sum_{i=1}^{n} (y_i - x_i'\beta)^2 = \sum_{i=1}^{n} \epsilon_i^2 = \epsilon'\epsilon = (Y - X\beta)'(Y - X\beta)$$

To estimate the β values, the first derivative of the LS formula with respect to β is set to zero and the second derivative of this equation is shown as positive. Therefore, the LS method aims to estimate the regression parameters by minimizing the sum of the squared deviation. Based on the LS estimator $\hat{\beta}$, the mean of y can be estimated (Fahrmeir et al., 2013, p.105-108).

$\hat{\beta}$ and $\hat{\epsilon}$ can be found as follows:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

$$y = X\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = H\mathbf{y}$$

$$\hat{\epsilon} = y - \hat{y} = y - H\mathbf{y} = y - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

where H is called the prediction matrix or hat matrix which is an $n \times n$-matrix and $\hat{\epsilon}$ is a residual calculated by the difference between the actual value of y and estimated value $\hat{y}$. The formulas for covariance matrix of residuals and variance of residuals as follows,

$$\text{Cov}(\hat{\epsilon}) = \text{Cov}(I - H)y = (I - H)\sigma^2 I(I - H)' = \sigma^2(I - H)$$

$$\text{Var}(\hat{\epsilon}_i) = \sigma^2(1 - h_{ii})$$

where h_{ii} represents the ith diagonal element from the hat matrix(Fahrmeir et al., 2013, p.122).

As the true value of σ^2 is unknown, $\hat{\sigma}^2$ is an estimator of the true value of σ^2. The formula is as follows:

$$\hat{\sigma}^2 = \frac{\hat{\epsilon}'\hat{\epsilon}}{n - p}$$

Here, P is the rank of the matrix which is calculated as k+1, and n is the number of observations.

The maximum likelihood (ML) estimator can be computed by assuming that the errors are normally distributed i.e., $\epsilon \sim N(0, \sigma^2 I)$. And it leads to $y \sim \mathcal{N}(X\beta, \sigma^2 I)$. The following equation for maximum likelihood estimation (Fahrmeir et al., 2013, p.107),

$$L(\beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right)$$

3.5 Significance of parameter estimation

In a multiple regression model, we want to find out if an x variable is helpful to explain the y variable or not. For example, if the model has two independent variables, we can test if

variable x1 is a meaningful predictor variable in the model.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

The standard hypothesis test is equivalent to:

$$H_0 : \beta_j = 0 \qquad \text{[null hypothesis]}$$
$$H_1 : \beta_j \neq 0 \qquad \text{[alternative hypothesis]}$$

To test the null hypothesis for β_1 where j=1, we need to see if $\hat{\beta}_1 x_1$ is sufficiently far from zero which is determined by the standard error of estimate $SE(\hat{\beta}_1)$. If β_1 = 0, the model will be changed by reducing one predictor. If $SE(\hat{\beta}_1)$ is small, $\hat{\beta}_1$ may provide strong evidence that $\beta_1 \neq 0$ hence there is a relationship between X and Y . Alternatively, if $SE(\hat{\beta}_1)$ is large, then $\hat{\beta}_1$ must be large in absolute value to reject the null hypothesis. Statistical software will generate p-values for all coefficients in the model to perform the test, which measures the number of standard deviations that $\hat{\beta}_1$ is far from zero. Each p-value will be calculated by t-statistic (James et al., 2013, p.67).

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

3.6 Several model selection criteria

Model selection criteria are quantitative measures that are used to compare different statistical models. These criteria help in the process of choosing the most appropriate and best model from a set of candidate models, where each independent variable describes the relationship with the dependent variable. The main goal is to balance the trade-off between model complexity and goodness of fit to the data.

3.6.1 Akaike information criteria

One approach is to use probabilistic statistical measures that attempt to quantify not only the model performance on the training dataset but also the complexity of the model. The Akaike information criterion (AIC) is one of the most often used methods for scoring and selecting the models. It uses a model's maximum likelihood estimation (log-likelihood) to provide a

score. The model with a lower score indicates a better model fit. AIC is defined as,

$$AIC = -2.l(\hat{\beta}_M, \hat{\sigma}^2) + 2(|M| + 1)$$

where σ^2 is the error variance and M is the number of covariates included in the model. $l(\hat{\beta}_M$ indicates the maximum value of log-likelihood. In a linear model with Gaussian errors, AIC is defined as,

$$AIC = n \log(\hat{\sigma}^2) + 2(|M| + 1)$$

(Fahrmeir et al., 2013, p.148).

3.6.2 Bayesian information criteria

Another method is the Bayesian information criterion(BIC) for scoring and selecting a model. Like AIC, it also uses a model's maximum likelihood estimation; smaller values indicate a better model fit. However, BIC penalizes the additional parameters much more than AIC, which means that more complex models will be less likely to be selected. BIC is generally defined as (Fahrmeir et al., 2013, p.149),

$$BIC = -2.l(\hat{\beta}_M, \hat{\sigma}^2) + \log(n)(|M| + 1)$$

Assuming Gaussian errors,

$$BIC = n. \log(\hat{\sigma}^2) + \log(n)(|M| + 1)$$

3.7 Variable selection

Practically dataset may have a lot of covariates. The selection of covariates that have a better relationship with the dependent variable is necessary for neglecting the complexity of the model. Variable selection is the method of identifying independent variables or covariates that have better links with the dependent variable, to construct a singular model that incorporates only those covariates.

There are four popular variable selection methods: forward selection, backward selection, stepwise selection, and best subset selection. Forward selection begins with the model having no variable and gradually includes variables to the model, whereas backward selection begins with a full model that takes into account all of the variables to be included in the model, and

stepwise selection is a combination of two selection methods, forward and backward (James et al., 2013, p.78 - 79). The best subset selection method involves testing every possible combination of predictor variables and by this, selecting the best model (James et al., 2013, p.205).

3.8 Stepwise Selection with AIC

In this report, the stepwise selection method is chosen to find the suitable subset from the model because it is compatible to conduct automatically in most statistical packages. This method is a combination of forward and backward selection procedures that allow moving in both sides, adding and removing variables at different steps with AIC as the selection criteria. The AIC criterion is defined for a large class of models fit by maximum likelihood. In general, a small value indicates a model with a low test error and better fitted, therefore, the model with the lowest AIC value is selected as the best model (James et al., 2013, p.212).

Mathematical equation-

$$AIC = \left(RSS + 2d\hat{\sigma}^2\right)/n\hat{\sigma}^2 \tag{1}$$

where RSS = Residual Sum of Squares, n = number of observations, σ^2 = estimated variance of the residuals, and d = number of parameters (coefficients) in the model.

3.9 Residuals vs. fitted plot and model evaluation

The residual plot is a scatterplot that represents the difference between the real observed and fitted response values. The residuals vs. fitted plot is a scatter plot where the fitted response values and their residuals are plotted with a horizontal reference line. The fitted values are plotted on the x-axis, on the other hand, the residuals are plotted on the y-axis. This plot is used to clarify the linear model's assumption of homoscedastic error variances. If the points are scattered randomly around the reference line with a constant variance, this indicates a homoscedastic error variance. For the opposite scenario, the error variances are heteroscedastic (Fahrmeir et al., 2013, p.183)

Model evaluation includes some tests like normality, linearity, heteroskedasticity, and Multicollinearity, etc. The QQ plot shows if the residuals are normally distributed by checking the points that lie on the straight reference line in the plot. Multicollinearity occurs when covariates in a regression model have a high correlation with one another, and the regression

estimate becomes unstable with a high standard error. To detect multicollinearity in a dataset, we can compute the Variance Inflation Factor (VIF) for each independent variable.

4 Statistical analysis

This section illustrates all the methods and procedures applied. For the calculation and visualization, the R software (version 4.2.3) is used (R Core Team, 2022). The R packages ggplot2 (Wickham, 2016) and plyr (Wickham, 2022), "tseries", "readr", "urca","tidyverse", "rstatix", "olsrr", "ggfortify", "leaps", "lubridate", "dplyr", "forecast", "stats", "stlplus", "dynlm","zoo","MASS" are used for this task.

4.1 Analyzing Load data structure with ACF PACF

Initially, a benchmark model is fitted on load data to observe the load data distribution from Jan 2015 to March 2024. Load data is about 30,000 MW to 80,000MW. The frequency of load data is high from 45,000MW to 70,000MW.

In the ACF plot, each bar represents the correlation between the current observation and its lag. For example, The ACF at the 1st lag is approximately 0.9 which shows a strong positive correlation between the current value and the value 1 hour ago. A stationary time series is one whose statistical properties are such as mean, variance, and autocorrelation structure do not change over time and the ACF of the series remains constant over time. Appendix Figure 6 shows current load data strongly correlate with 1-hour, 2-hour, 3-hour, and 24-hour previous load data. By analyzing yearly PACF lag 144, 168, and 336 are found strongly correlated.

A seasonal pattern is also followed that at winter load generation increase than summer. On weekends load generation decreases so there have strong correlation with weekends (lag 167,168). For 2015 to 2024, data have a nearby similar trend over all years.

4.2 AR model based on information criteria

4.2.1 AR model selection

In time series data analysis, selecting the appropriate lag structure for an autoregressive model based on the dataset is important for accurate forecasting. To determine the optimal lag struc-

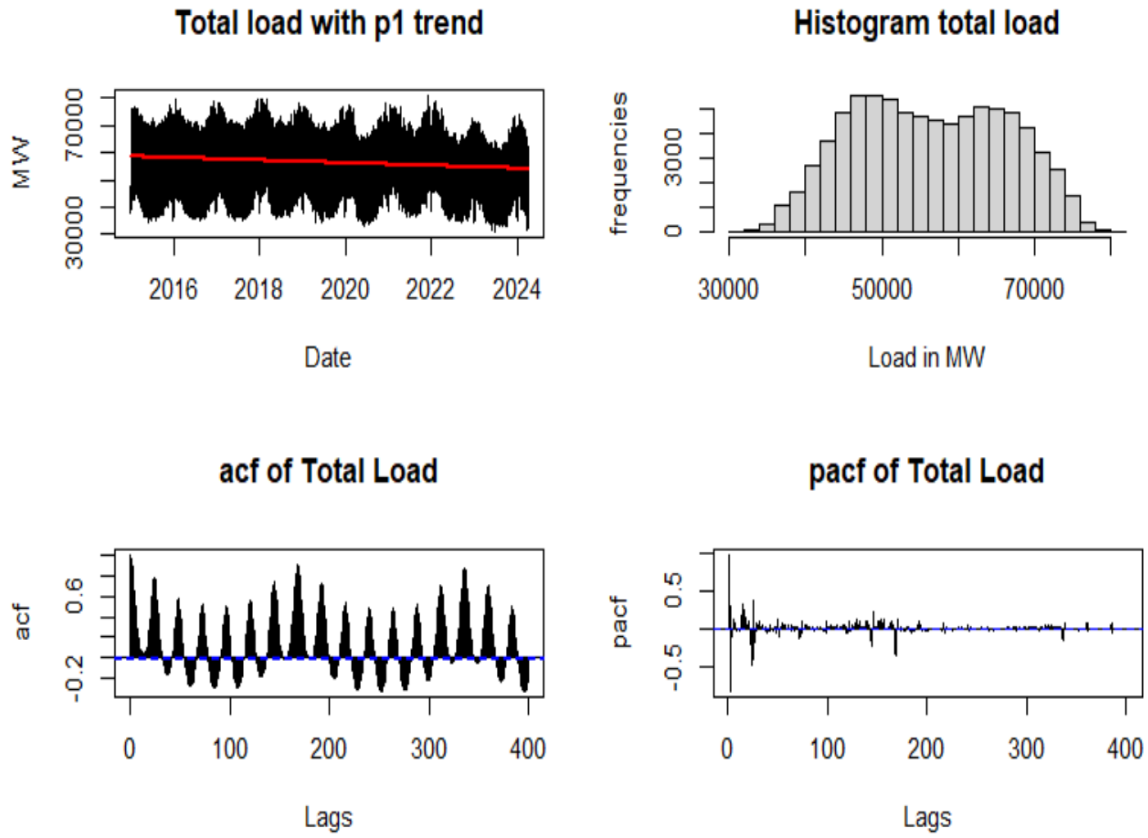

Figure 1: Load data Distribution, Histogram of Load data, ACF of Load data, PACF of Load data

ture, information criteria such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) is calculated to balance model complexity.

For more robust AR model specification, a simulation is run over a different period and generates different models with all possible combinations of 13 lagged predictors(lag matrix). This is a process of AR model specification based on lagged variable distribution. For the limitation of GPU on personal devices, choose 13 lags according to PACF. For lagged load predictors, hourly, daily, and weekly lagged variables up to specific intervals are included according to PACF for different frequencies. For more specification, different lags outside of the threshold are selected. Linear regression models for each combination of lagged predictors are fitted, the AIC and BIC for each model are calculated, and select top 10 models are selected for further analysis. Here the simulation is run over the period 1st March 2015 to 29th February 2016, a year of data. The following AR model is selected as the best model within the 10 best models from the simulation.

Model 1:

Formula: $\sim$ sampleload hourlag1 + hourlag2 + hourlag3 + daylag1 + daylag6
+ weeklag1 + weeklag2 + day1_hlag1 + day1_hlag2
+ day1_hfwd1 + week1_hlag1 + week1_hlag2 + week1_hfwd1

AIC: 136876

Here, the models predictors means, one hour lag, two hour lag, three hour lag, one day lag(24), six days lag (144), one week lag(168),two week lag(336), 25 hours lag with, 26 hours lag, 23 hours lag, one week and one hour lag(169), one week and two hour lag(170),one hour less than one week lag(169). In Figure 2 Aic of best 10 selected models and their MSE is showing.

4.2.2 AR model on full sample forecast

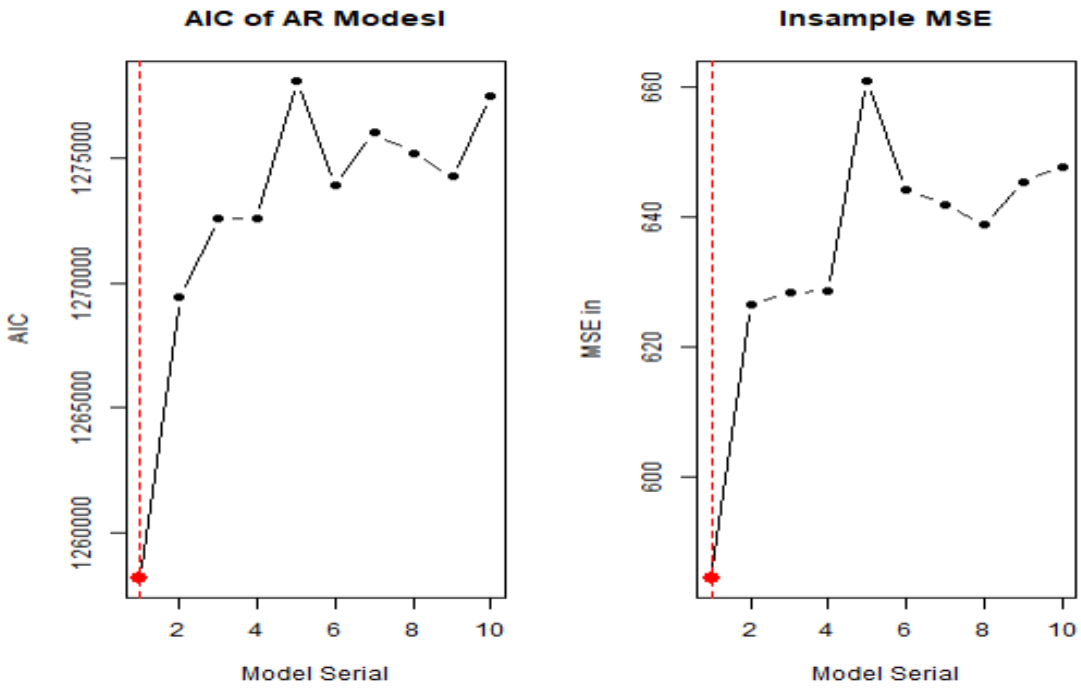

Figure 2: A correlation plot for descriptive analysis of variables

The selected 10 AR models are fitted over the full load dataset and calculate the AIC and mean squared error for all models. Over the full-load dataset Model 1 gives the lowest AIC and MSE values. In Figure 3 the fitted values for AR model and real load data are visualized.

Table 1: Best AR model

Model	AIC	BIC	MSFE
AR	1258220	1258360	584.5

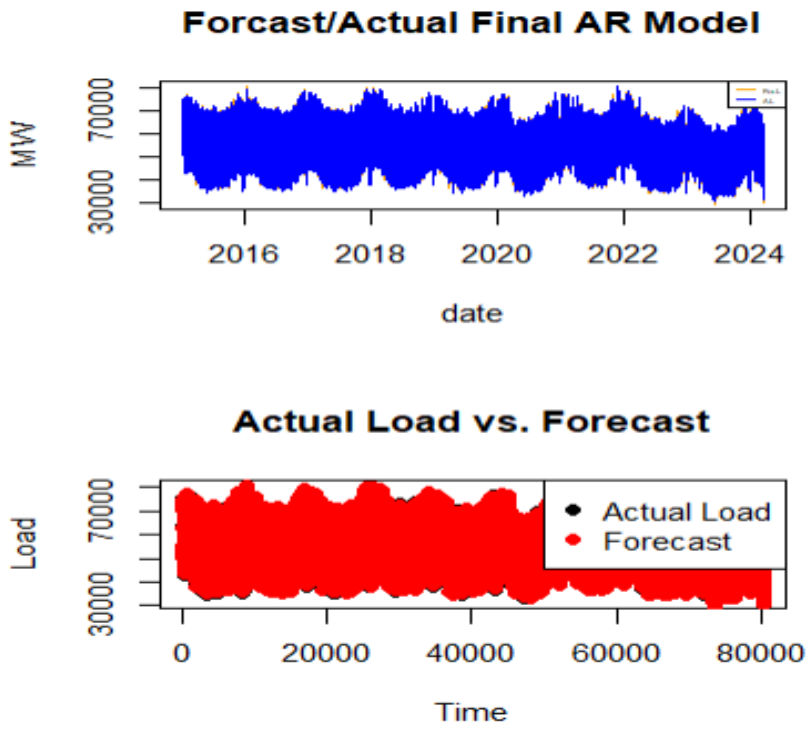

Figure 3: AR forecast and Real Load data

In real life time series data is not stationary often. Although sufficient and important load-lagged predictors are fitted into model, some special effects like holidays, most complex dynamics, and global reasons for load variation are not counted for this model. So in real life, it's not possible to fit all possible predictors other than the most prominent ones.

4.3 External Predictor selection

In this project, for further analysis and prediction correctly the load data, some external predictors are counted. To extend the analysis beyond AR models by incorporating various external predictors, such as temperature, carbon prices, output per generation type, imports and exports, day-ahead forecasts for wind and solar generation, allocated transfer capacities, and maintenance schedules.

Adding every single predictor with the Base AR model a clear summary is visible with forecast data, MASF, AIC, etc. The external predictors that fit the model are lagged one hour. Compared to AR model most of the predictors are insignificant. Only in Generation type $Fossil$, $Hydro$, $Wind$, $Waste$, $Solar$ categories are significant. These predictors' p-values are lower than 0.05. The $Other$ category doesn't exhibit significant predictive power. Its MSFE is comparable to that of the Base AR model. Therefore, the base model seems to provide sufficient forecasting accuracy.

Table 2: Summary of models with individual external predictors

Models	MSFE	Pr(>|t|)
AR	584.5	0.3481433
AR + Temp	583.45	insignificant
AR + Generation Per Type(Fossil,Hydro etc)	582.8	significant
AR +Maintanence	583.5	insignificant
AR + Carbon	583.5	insignificant
AR +Day Ahead wind,solar forcast	583.4	insignificant
AR + Allocated transfer capacity	583.5	insignificant
AR + net export	583.4	a bit significant

4.4 Model with all predictors

Models with individual predictors have nearby MSFE values with the Base AR model. Which indicated a very low effect on predicting future load. For the individual predictor models, Generation type $Fossil$, $Hydro$, $Wind$, $Waste$, $Solar$, and $netexport$ had a significant effect (p-value<0.05) on one-hour ahead load prediction.

Predictive models using all predictors are also analyzed to find out the effect on load together. While individual predictors may have limited predictive power, integrating them into a complex model allows for a more holistic understanding of load behavior and leads to improved forecasting performance. The external predictors, carbon price ($eua\ futures\ price$), $Fossil$ and $Hydro$ generation type, $day\ ahead\ wind$ and $net\ allocated\ transfer\ capacity$ are the most significant found.

Table 3: model with all external predictors

Model	**AIC**	**MSFE**
AR+ all external predictors	1278820	618.6

4.5 Model with promising predictors

The external predictors, carbon price ($eua\ futures\ price$), $Fossil$ and $Hydro$ generation type, and $net\ allocated\ transfer\ capacity$ are the most significant found. In table5, shows the AIC and MAFE for model with promising predictors and model with all predictors. By fitting the most promising predictors with the Base AR model, found an MSFE value of 619.2 and an AIC value of 1278953, Which are a little higher than the AR model. Multicollinearity checking with VIF can assess multicollinearity among predictor variables in a regression model. Figure 7 and Figure 8 have multilinearity VIF values for a model with the most significant predictors. Lag 144,lag 336 and other external predictors have low value of VIF rather than others. By not counting these low valued predictors could help for a better prediction. For time shortage could not proceed further.

4.6 Model with forward stepwise predictor selection

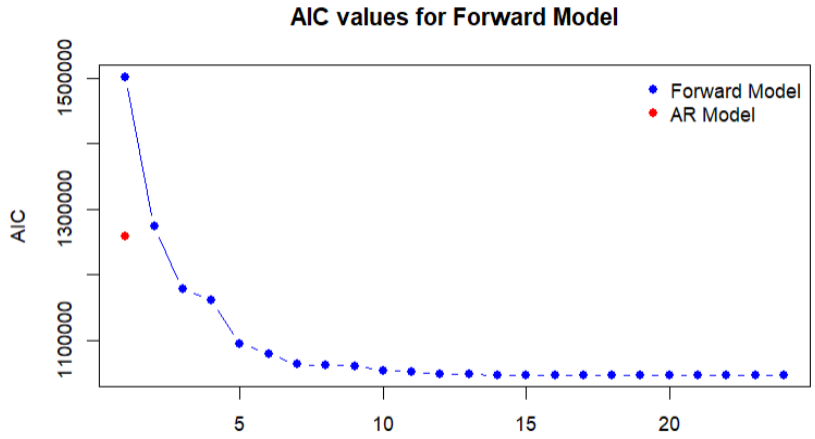

Figure 4: AIC for AR model and Forward model

Stepwise forward selection can be a useful tool for exploratory analysis and model building, but it also has limitations, such as overfitting, selection bias, and sensitivity to the selection criterion. In 5 showing all steps of the forward stepwise method for predictor selection with AIC values. For adding $installedcapacity$, $temperature$, and $solar$ the AIC values increased. So predictors should select until $netexport$ with AIC 1047561. In Table 4 shows the AIC of the forward model(with final selected steps) and the Base AR model.

Table 4: model with most promising external predictors

Model	AIC	MSFE
AR+ most promising external predictors	1279027 (high)	619.2
forward model	1047561 (low)	
Decomposed model	1094021 (medi)	199

4.7 Decomposed model

Next, conducts a seasonal decomposition and regression modeling to forecast electricity load based on promising predictors. It starts by decomposing the original load data into seasonal components at different frequencies. Also deseasoned previously selected promising predictors like Fossil, Hydro, Carbon, and Allocated Transfer Capacity. Fossils decomposed daily and weekly. Other predictors decomposed as daily by checking PACF.

Next, a multiple regression model is fitted using lagged load data and deseasoned predictors. The fitted values from the regression model are combined with the trend and seasonal components to derive the final fitted values. Residuals are calculated to evaluate model accuracy through Mean Squared Forecast Error (MSFE) which is a value of 199, shown in table 4. In figure 5 the AIC values for the AR model, AR and all external models, AR and specific predictors, Forward model, and Decomposed model are shown all together. The decomposed model has a relatively low AIC value compared to other models. The decomposed model performed well on full-sample data which are used for model fitting. A trade of need to generalize on out-of-sample data. Also further investigation by considering cross-validation, out-of-sample performance, and practical implications, etc is needed.

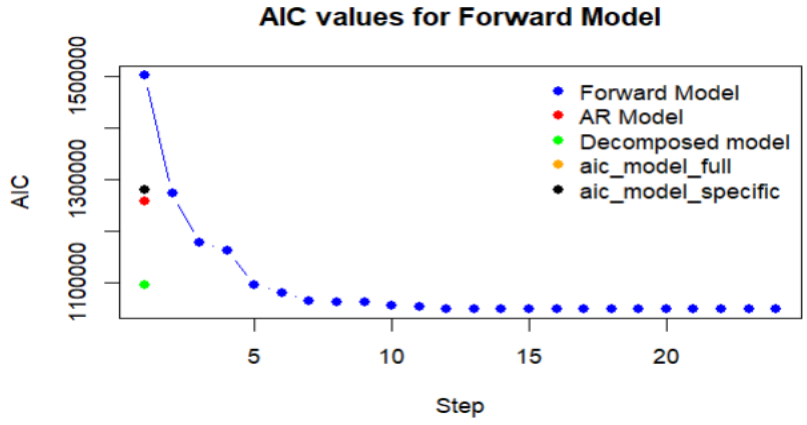

Figure 5: AIC for all model

5 Summary

Financial and economic data are very complex and unpredictable type because many factors always influence them, and their use in reality often requires consideration of their endogenous variable relationships, which is the core idea of most time series models. As the autocorrelation structure of provided data is not perfectly constant over time and includes trend and seasonal effects over data, these are not purely stationary. The trend and seasonal effect found by ACF, and PACF with different frequencies. Also, AR models based on information criteria capture the best-fitted model considering the temporal dependencies and patterns in time series data.

One simulation over a 1 year of data was run to find out the 10 best models according to information criteria. With analysis lag 1,2,3 24, 144, 168, 336, 25, 26, 23, 169, 170,167 correlate with load at time t. Similar models have run over full sample data (year 2015 to 2024) and found best fitted with the lowest AIC value. To improve the prediction of one-hour-ahead load data, some external predictors are evaluated one by one according to their AIC and MSFE values. Some generation types (especially $Fossil$, $Hydro$,) have found significant effects on prediction.

Later on, evaluating another model with the Base AR and with all external predictors, found MSFE higher than individual models. From the summary of the full model, select the most promising predictors and find not promising results. By analyzing VIF multicollinearity, we can find which variables have most higher colinearity. Lag 144, lag 336, and other external predictors have lower values of VIF rather than others. By applying forward model selection criteria found better results with AIC 1047561.

As some practitioners think about better-fitted models by decomposing trends and seasonal effects, an analysis also has on this basis. By decomposing seasonal effect on real load data and all promising external predictors and fit model with these deseasoned series and found a new final forecast by adding components that are far better than other models with MSFE 199.3. Although it is, by cutting on a specific step on forward stepwise model, possible to get a good fitted model.

Bibliography

Jiapeng Chen. *Comparison Between ARIMA Model and OLS Model Based on the Economic Representations*, 2022. URL https://www.researchgate.net/publication/366297281$_C omparison_B etween_A RIMA$

ENTSOE. *Central collection of Electricity generation, transportation and consumption data for the Pan-European market*, 2015. URL : https://transparency.entsoe.eu/.

Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, and Brian Marx. *Regression: Models, Methods and Applications*. Springer Berlin, Heidelberg, 2013.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*. Springer New York, NY, 2013.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. SpringerVerlag New York, 2016. ISBN 9783319242774. URL https://ggplot2.tidyverse.org.

Hadley Wickham. *plyr: Tools for Splitting, Applying and Combining Data*, 2022. URL https://cran.r-project.org/web/packages/plyr.

Appendix

A Additional tables

Table 5: Summary of models with individual external predictors

steps	AIC
1	1501858
+ Total load lag1	1273348
+ Total load lag2	1177872
+Total load lag167	1162315
+ Total load lag169	1094696
+Total load lag168	1080006
+ Total load lag170	1063349
+ Total load lag3	1062128
+ Total load lag23	1061215
+Total load lag25	1054287
+ Total load lag 24	1052683
+ Total load lag26	1048194
+Total load lag336	1047977
+HYdro	1047809
+ Fossil	1047779
+Total load lag144	1047762
+Day ahead wind	1047745
+ Day ahead solar	1047715
+net Allocated transfer capacity	1047656
+ eua futures carbon price	1047584
+ available capacity	1047575
+ other generation type	1047562
+ net export	1047561
+ Wind	1047561
+ installed capacity	1047562
+ temperature	1047563
+ solar	1047564

B Additional figures

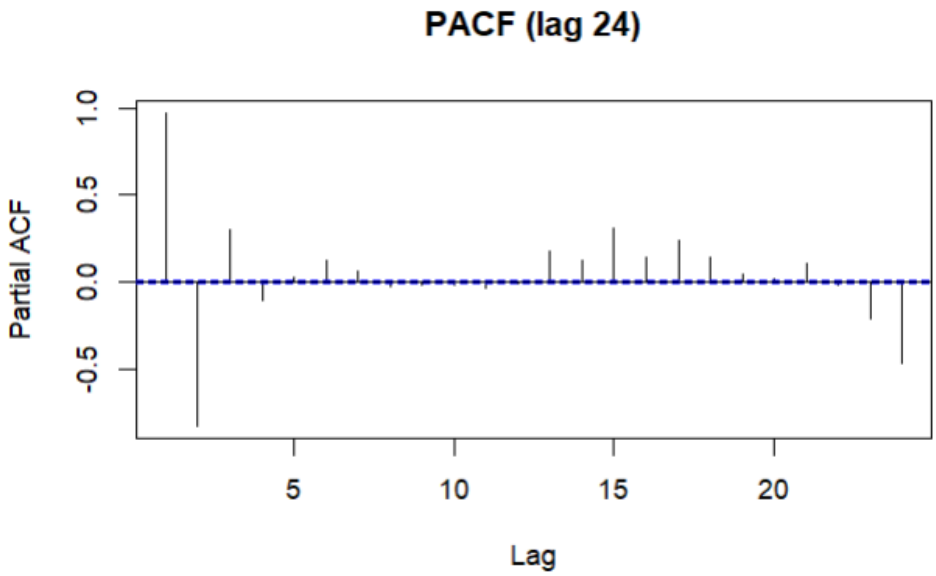

Figure 6: PACF of 24 lag with frequency 1

Variables <chr>	Tolerance <dbl>	VIF <dbl>
Total_load_lag1	0.004795	208.535
Total_load_lag2	0.002855	350.257
Total_load_lag3	0.015137	66.065
Total_load_lag24	0.001805	554.135
Total_load_lag144	0.340057	2.941
Total_load_lag168	0.001760	568.185
Total_load_lag336	0.219729	4.551
Total_load_lag25	0.001735	576.268
Total_load_lag26	0.006794	147.187
Total_load_lag23	0.007665	130.466

Figure 7: Multicolinearity for model with most promising predictors

Variables <chr>	Tolerance <dbl>	VIF <dbl>
Total_load_lag169	0.001571	636.613
Total_load_lag170	0.004734	211.247
Total_load_lag167	0.007570	132.104
lagged_data$Hydro	0.613935	1.629
lagged_data$Fossil	0.353014	2.833
lagged_data$eua_futures_price	0.675281	1.481
lagged_data$Day_Ahead_wind	0.388793	2.572
lagged_data$net_Allocated_transfer_capacity	0.455162	2.197

Figure 8: Multicolinearity for model with most promising predictors

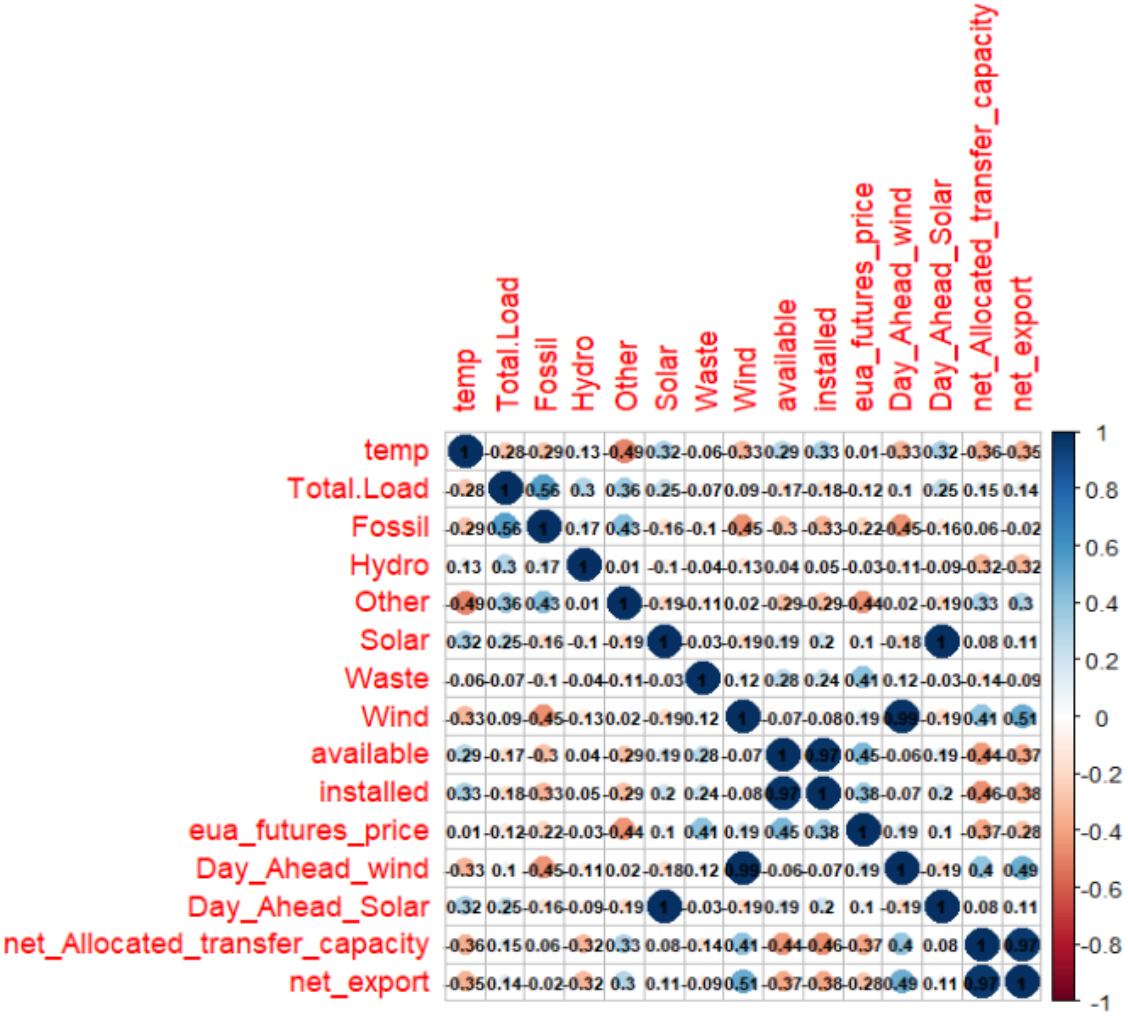

Figure 9: Correlation Matrix on predictors with TotalLoad