HELLENIC OPEN UNIVERSITY

School of Social Sciences

Postgraduate Program of Studies in

Supply Chain Management – Energy Systems

Postgraduate Dissertation

"Time Series Analysis of Electricity Demand"

Gerasimos Minetos

Supervisor: Nikolaos Thomaidis

Patras, Greece, June 2023

“Time Series Analysis of Electricity Demand”

Gerasimos Minetos

Supervising Committee

Supervisor*:*

Nikolaos Thomaidis

Co-Supervisor:

Alexandros Diamantidis

Patras, Greece, June 2023

Abstract

One of the most important factors for economic and social development is energy availability, since different forms of energy are used for manufacturing, transportation, electricity generation and heating. In academic literature, electricity is cited as the form of energy that has penetrated the most in almost all aspects of modern life, being essential non only for commercial uses but also for residential uses. During the last couple of years, the significant supply chains disruptions in the international energy markets of crude oil and natural gas, caused by the Russia-Ukraine military conflict, created significant variations in the electricity markets and ignited the interest for examining the security of the electricity grids. An essential part of the academic research focuses on research about electricity load demand and forecasting of electricity load as it enables the design of immediate and efficient responses in demand variation. Increasing electricity demand requires the availability of raw energy sources as well as the good operation of the electricity grid, while period of low electricity demand should be matched with decreasing electricity production for reducing production costs. The majority of existing models for electricity modeling and forecasting in literature are characterized by increasing complexity making their use from policy makers and market experts difficult. The purpose of this postgraduate thesis is to examine the statistical behavior of the suggested electricity load models, examine their out of sample forecasting capabilities and performance. The models that will be examined in this dissertation are estimated using hourly observations for electricity load, day-ahead price and temperature in Spain during 2019. Postestimation assessment and comparison of the examined models will be conducted with the use of forecasting accuracy metrics.

Keywords: Electricity load modeling, forecasting, regression models, time series analysis.

“Ανάλυση Χρονοσειρών Ζήτησης Ηλεκτρικής Ενέργειας”

Γεράσιμος Μινέτος

Περίληψη

Ένας από τους σημαντικότερους παράγοντες για την οικονομική και κοινωνική ανάπτυξη ενός κράτους είναι η διαθεσιμότητα ενέργειας καθώς διαφορετικές μορφές ενέργειας είναι απαραίτητες στην βιομηχανία, τις μεταφορές, την παραγωγή ηλεκτρικής ενέργειας καθώς και για την θέρμανση. Στην ακαδημαϊκή βιβλιογραφία, η ηλεκτρική ενέργεια αναφέρεται ως η μορφή ενέργειας που έχει διεισδύσει περισσότερο σε σχεδόν όλους τους τομείς της καθημερινότητας, και είναι απαραίτητη όχι μόνο για την χρήση της από την βιομηχανία αλλά και για τις καθημερινές ανάγκες των νοικοκυριών. Τα τελευταία δύο χρόνια, οι σημαντικές διαταραχές στις εφοδιαστικές αλυσίδες ενέργειας στις αγορές αργού πετρελαίου και φυσικού αερίου που προκλήθηκαν από την στρατιωτική σύγκρουση μεταξύ Ρωσίας και Ουκρανίας, δημιούργησαν σημαντικές διακυμάνσεις στην αγορά ηλεκτρικής ενέργειας και αναθέρμαναν τον ενδιαφέρον για την έρευνα της διασφάλισης της σταθερότητας στην λειτουργία του δικτύου ηλεκτρικής ενέργειας. Ένα μεγάλο μέρος της ακαδημαϊκής έρευνας επικεντρώνεται στην μελέτη της ζήτησης για ηλεκτρική ενέργεια και την πρόβλεψη της μελλοντικής της κατανάλωσης προκειμένου να διασφαλιστεί η κάλυψη της μεταβολής της μέσω του σχεδιασμού αποτελεσματικών ενεργειακών πολιτικών. Η αύξηση της ζήτησης για ηλεκτρική ενέργεια απαιτεί την διαθεσιμότητα των πρώτων πηγών ενέργειας που χρησιμοποιούνται στην παραγωγής της, καθώς και την εύρυθμη λειτουργία του δικτύου διανομής ηλεκτρικής ενέργειας. Μία περίοδος με χαμηλή ζήτηση ηλεκτρικής ενέργειας θα πρέπει να συνδυαστεί με μείωση της παραγωγής ηλεκτρικής ενέργειας προκειμένου να μειωθεί το κόστος λειτουργίας των σταθμών παραγωγής ενώ μια περίοδος υψηλής ζήτησης πρέπει να συνοδεύεται με αύξηση της προσφοράς. Η πλειονότητας των μοντέλων που συναντάμε στην βιβλιογραφία για την εξέταση της συμπεριφοράς αλλά και την πρόβλεψη της ζήτησης για ηλεκτρική ενέργεια χαρακτηρίζεται από μοντέλα αυξημένης πολυπλοκότητας, καθιστώντας δύσκολη την χρησιμοποίηση τους από τις ρυθμιστικές αρχές καθώς και τους συμμετέχοντες στην αγορά. Σκοπός της παρούσας μεταπτυχιακής διατριβής είναι η εξέταση της στατιστικής συμπεριφοράς των προτεινόμενων από την βιβλιογραφία μοντέλων ζήτησης ηλεκτρικής ενέργειας καθώς και η εξέταση της ικανότητας των μοντέλων αυτών να εξάγουν προβλέψεις για την μελλοντική ζήτηση από το δείγμα των παρατηρήσεων. Οι μεταβλητές των μοντέλων που θα χρησιμοποιηθούν στην παρούσα μεταπτυχιακή εργασία θα εκτιμηθούν με την χρήση ωριαίων δεδομένων κατανάλωσης ηλεκτρικής ενέργειας, τιμής day-ahead και θερμοκρασίας για την Ισπανία κατά την διάρκεια του έτους 2019. Τα εξεταζόμενα μοντέλα ζήτησης ηλεκτρικής ενέργειας θα συγκριθούν μεταξύ τους με την χρήση δεικτών ακρίβειας πρόβλεψης (forecasting accuracy metrics).

Λέξεις – Κλειδιά

Μοντέλα ζήτησης ηλεκτρικής ενέργειας, θεωρεία προβλέψεων, μοντέλα παλινδρόμησης, ανάλυση χρονοσειρών.

Table of Contents

List of Figures

List of Tables

List of Abbreviations & Acronyms

AR model: Autoregressive model

ARIMA model: Autoregressive Integrated Moving Average model

CDD: Cooling Degrees Day

HDD: Heating Degrees Day

MAD: Mean Absolute Deviation

MAPE: Mean Absolute Percentage Error

MSE: Mean Squared Error

SARIMA model: Seasonal ARIMA model

SMA: Simple Moving Average

1. Introduction

Modern developed and developing economies and societies are relying heavily in energy consumption for the majority of production activities and almost for every daily living activity. The most common energy form that is used in production of several commodities as well as for common residential purposes such as cooking, lighting and heating; is electricity. As the global population is expected to continue the increasing trend during the next years; the demand and use of electricity is also expected to follow a similar increasing trend.

Electricity is distinguished from the other forms of energy in the sense that is derived by other primary sources of energy such as coal, natural gas, nuclear, geothermal, hydropower, photovoltaics, and wind turbines. The two unique characteristics of electricity are that it is a just-in-time form of energy (since there is no delay between electricity generation in the production facilities and electricity consumption by the final consumer), and that electricity with the existing technology cannot be stored in significant quantities for future use (no ability for inventory building). Traditionally electricity is generated in production generation facilities where primary energy sources such as carbon, natural gas, and nuclear energy are burned to create heat that in turn is used for heating water. The heated water in turn produces high pressured steam that moves the blades of a turbine, and consequently generates electricity.

Electricity eventually is supplied into the electricity grid and is distributed to the final consumer. The greatest share of electricity generation nowadays is derived by energy sources that are exhausting (coal, natural gas, uranium) and their use has significant environmental consequences. As the power stations are using primary energy sources as fuel for the electricity generation turbines and there is time needed for powering up additional turbines, the management of electricity demand and supply is crucial for decreasing operational and environmental costs (Sadler, 2022). Electricity power stations are usually located far from the final consumer since they usually located near the primary sources of energy. From the production facilities to the final consumer, electricity is transmitted and distributed through the electricity grid of a country. The electricity grid is the interconnected system of operating and back up electricity generation plants, electricity transmission substations, transmission lines able to transfer high voltage electricity, local electricity substations that are transforming high voltage electricity to lower for being able to be used by the final customers through distribution lines. In addition, there are several smaller electricity grids or micro-grids that are connected in the national electricity grid such as regional grids, community grids, university campus grids or even industrial area grids (Sadler, 2022).

Each final user in the electricity system removes electrical power for operating business or residential machines. This electricity removed from the grid is called electricity load or electricity demand. The electricity demand has significant variations during a day or a week or even a whole year since there are different factors affecting the final consumer on using electrical devices. Electricity generation facilities are feeding the system with always with a minimum level of electricity, the base load, while they implement additional resources for facing periods of increased demand (high peak load). Nowadays, the electricity generation facilities are large investment projects since they use natural gas and nuclear power as a fuel, providing the base load electricity in the grid while there are several back-up electricity generation facilities that operate during peak times. The back-up generation plants require significant resources for powering up and starting generating electricity, adding significant cost in the additional high peak load. The optimal operation of the electricity grid requires taking decisions about the efficient allocation of energy resources and reducing electricity generation costs. Important role in decision making in electricity generation and distribution plays the understanding and forecasting the electricity load. Knowing the periods of low electricity demand could allow periodical shut down of generation plants for maintenance or security assessment, while knowing the periods of high electricity demand allows the on-time operation of back-up units, the scheduling of resources or increasing the generation capacity with renewable sources of energy. Understanding and forecasting electricity load plays an important role in the liberalized competitive electricity markets not only because ensures the system stability but also affects the decisions for expansion possibilities, changes in companies' market shares, and create profits from trading excess electricity capacity in near countries through the interconnection of grids.

Forecasting techniques and models are being used extensively in academic literature in many areas. In economics and finance, forecasting models have been designed in predicting inflation, exchange rates and the future behavior of macroeconomic variables. In supply chain management forecasting has been used in inventory management and demand prediction, while forecasting methodology has also been used outside the area of business and economy, such as medicine. In electricity load forecasting there are several different methodologies and forecasting horizons developed without a clear consensus about the forecasting superiority of a model or forecasting accuracy of complex modeling. The models developed in academic literature range from traditional causal relationship models such as regression and multiple regression models, time series models, exponential smoothing, artificial neural networks, and fuzzy logic models.

Developing complex models with the use of many different variables has not provide evidence of a better understanding of the behavior of electricity demand or producing more accurate forecasts compared to more simple models. This thesis will try to understand the electricity load behavior by focusing in simple econometric and time series techniques. The research question is how simple models that could be used by a broad range of users such as market experts and policy makers capture the behavior of electricity load. In addition, can these models produce reliable forecasts in the short-term using real data for Spain? The structure of the thesis is organized as: Chapter 2 Literature Review provides a brief review of existing literature on energy and electricity forecasting, Chapter 3 Methodology presents the different models and the data that will be used in the analysis, and Chapter 4 Conclusion will discuss the statistical results, accuracy measures and concluding remarks.

2. Literature Review

2.1 The Spanish Electricity Market

The Spanish electricity market has been subject to continuous transformation since the early 20th century, not only in the regulation that govern it but also in the energy mix used for electricity production. In addition, different social and economic events such as the period of the Spanish Civil war, the European deregulation of electricity markets, the economic and oil crises during the 1970s affected the energy decisions about electricity generation through the years (Chaparro-Peláez et al., 2020).

The electricity market in Spain before the 1997 was regulated by the Ministry of Industry and Energy which was setting the electricity prices in an effort to achieve efficiency in electricity market and financial stabilization. The State through the state-owned company Red Electrica de Espana was controlling the electricity system and transmission network. In electricity generation side there was not competition present and electricity generation was made from a couple of fully integrated private-owned companies. The generation capacity of each company and electricity generation station was known and the risk was significant low. The principle of price formation in state regulated electricity markets was the profit maximization by taking into account the investment required and the operational costs from electricity generation and distribution. The Spanish State, thus, was setting the electricity tariff and was controlling the efficient allocation of energy resources and investment allowances (Chaparro-Peláez et al., 2020).

The Spanish energy mix during this period included coal and hydropower energy sources as the largest shares used for the electricity generation. Coal was the dominant energy fuel during this period since Spain, as the majority of European countries, had not any oil and gas reserves and hydropower had volatility in generation capacity due to weather conditions and especially long dry periods during summer. Also, the oil crises during the 1970s together with the growing electricity demand affected the government decisions in the electricity market in an attempt to increase capacity through investments in additional coal burning and nuclear plants.

The liberalization of the Spanish electricity market started in 1994 with the creation of a regulatory and transparency commission for the entire electricity system, and was concluded in 1997 with the adoption of the European Directive about electricity markets. State intervention in the electricity supply is not anymore needed since the basic principle dictates that competition among the existing companies and the companies that are free to enter in the market would result to the efficient match of supply and demand, stimulate technological innovation and promote efficiency of new investment decisions. Adding to the changes in the regulation of the electricity market, significant changes observed also in the Spanish energy mix. The environmental consequences of coal burning power plants, European policies to reduce the CO2 emissions, and the introduction of combined heat and power (CHP) using natural gas as energy fuel changed the primary energy sources included in the Spanish energy mix. In the energy mix there is an increasing share of renewable sources of energy due to the cost decrease and competitiveness of wind turbines and photovoltaics. Nowadays, the electricity generation in Spain is derived mainly from the use of natural gas fueled generation facilities, nuclear plants, hydropower plants and renewable sources of energy while the existed coal-burning power plants are operating during periods of high electricity demand as a back-up plan.

Spain is one of the largest European countries in terms of population and electricity consumption. The electricity demand in the country shows significant variation due to different weather conditions that are observed during the year (Moral-Carcedo & Vicéns-Otero, 2005). A large percentage of the Spanish residential population has not heating utilities that operate with natural gas or oil, having as a result the use of electricity for heating systems such as radiators and electricity heat pumps (Blázquez et al., 2012). As population in Spain has been constantly growing, the electricity demand has also been increasing. This increasing electricity demand in the country raises concerns not only for the increasing emissions into the environment from the power plant facilities, but also about the energy security and autarchy of the country. Disruptions in the availability of electricity could affect negatively the industrial production and economic growth of the country. In addition, excess electricity demand according to the economic theory has as a consequence high electricity prices, which in turn affect key macroeconomic indicators and household incomes.

Electricity nowadays is viewed as a commodity that is also traded in a similar way with the other traditional commodities in organized markets, the Power Exchange Market. The market liberalization of the electricity market led in the competition of mainly two large private-owned companies, Endesa and Iberdrola that generate the majority of electricity capacity and few smaller private-owned competitive companies, Gas Natural Fenosa, EGL, EDP Hidrocantabrico Energia, and Acciona. Electricity price formation is set every day in the Power Exchange, an organized market in which electricity producers and electricity purchasers are negotiating different selling and buying prices. The electricity power exchange is settling the hourly prices of electricity and the respective quantities of electricity sold and bought at that price level. The auctions in the daily power exchange are settled by the market operator. The operator is matching the selling and buying orders, calculates the marginal electricity price, distributes the generation and demand share among the participating agents, and examines different technical restrictions in the production or transmission of electricity. The difference of electricity power exchange with other commodity organized markets is that prices are set for the next day of the occurrence of the negotiations (day a-head prices). Obviously, there is risk inherited in the process since different events could occur affecting the generation facilities or the transmission lines, creating supply disruptions in electricity supply chain.

2.2 Forecasting in Electricity Markets

During the last decades, there are several researches and concerns about energy management since the availability of energy resources is decreasing and several actions have been taken for the decarbonization of the national energy mixes. Electricity is generated through the use of other energy sources with carbon being constantly replaced from other alternative fuels with less CO2 emissions. However, the availability of electricity generated by hydropower, wind farms, and photovoltaics is subject to the weather conditions and cannot reliably face unanticipated peak electricity load periods. Also, natural gas and nuclear plants dependent on energy fuels that their reserves are decreasing and their prices exhibit significant volatility. Electricity capacity is, thus, dependent in the availability of energy sources and the risk of supply chain disruptions in the value chains of those energy sources could affect the stability of the whole electricity power system.

The disruptions in the supply of energy fuels such as natural gas (for natural gas electric generation plants), Uranium (for nuclear electricity generation plants) or the unavailability of water resources (for hydropower generation stations) and the dependance of renewable sources of energy in weather conditions, are not the only unforeseen events that could create disruption in electricity systems. Significant risk is also inherited in the physical infrastructure of the electric grid. Possible breakdowns in the step-up or step-down voltages substations, load imbalances that could create a black out in the whole system, or damages in the transmission cables create threats in the optimal power system operations.

An important difference between electricity and other sources of energy or other commodities is that electricity is a just-in-time form of energy since the electricity generation plants create electricity, simultaneously fed into the electricity grid and consumed by the final residential, industrial and commercial user. This property of electricity of non-storage availability and inventory building for future use does not allow an easy solution in the problems mentioned above. In general terms, electricity demand presents strong seasonal variation since there are periods of significant low electricity load and periods of peak electricity load. Electricity generation plants feed constantly the system with the minimum electricity necessary. However, there are periods during a day that additional utilities are operating for matching increased electricity demand by feeding additional load in the system (intermediate load). The efficiency of the system is tested during small periods of time when sudden demand increases require the operation of back-up electricity power generation facilities since the base and intermediate load suppliers cannot face the peak electricity load. In the other hand, oversupplying the system with electricity load creates significant operational costs and wastes of energy fuel resources. Thus, efficiency and optimal operation of the electricity grid requires a matching in the electricity generation with the electricity load.

Electricity demand forecasting plays an important role in the electricity supply and demand management for all the agents that participate in the electricity market. Available information about the electricity load behavior and forecasted electricity load allows the electricity generation facilities (supply side) to efficient allocate their resources and minimize operational costs. Electricity generation facilities can efficiently schedule their power generation from the available energy sources, reduce the cost from the wastage energy source consumption, take advantages of periods when there is increased electricity generation from renewable sources of energy and ensuring a reliable electricity supply in the grid. In addition, accurate forecasts allow the electricity system operator to change the electricity generation to match the forecasted demand (secondary regulation system).

Accurate load forecasts are beneficial not only for the electricity generation companies and system operators, but also for policy makers and distributors of electricity. Long-term forecasts about electricity demand allow the policy makers to take decisions for supporting additional investment on production capacities or about the energy mix of the country in the short-term or in the long-term. Periodic maintenance of production facilities or transmission facilities can be performed without creating disruptions in the grid with the use of electricity load forecasts. Forecasting electricity load provides information about potential expansion of the electric grid or creating additional substations. Electricity modelling and forecasting could provide valuable information to electricity grid operators and distributors for the effective and resilient grid operation. The continuous monitoring of electricity load pattern and the behavior of the factors that affect electricity consumption allows the grid operators to detect anomalies in the electricity grid, identify potential problems and take precaution measures for preventing blackouts, brownouts and electricity supply disruptions. Another important application of forecasting methodology in electricity markets is on the side of electricity price. Accurate forecasts about the wholesale electricity price could provide a significant benefit for mitigating uncertainty and risk on generation production companies (Lindberg et al., 2019).

Significant academic research has been devoted also in the electricity price modeling and forecasting. As the electricity production companies are participating in a market where electricity prices and electricity production quantities are being settled for the next day, mitigating the operational risk needs being able to forecast the electricity price. Electricity load and price forecasting plays an important role in electricity Power Exchange since it enables all the participant in the market to make informed decisions about buying and selling electricity in the spot and future market. Electricity prices present similar seasonal variation due to the variation of electricity load (according to the law of demand and supply), variation in the economic activity (working days and weekends), and regional climate factors. In addition, the time series process of electricity prices exhibits several spikes, negative values and long memory (Seitaridis et al., 2021). Accurate forecasts allow the optimization of market participants' portfolio, the efficient management of risk in electricity trading markets and profit maximization.

2.3 Overview of Electricity Load Models

Forecasting or predicting the future values of behavior has been applied in a wide range of human interests. Among the first forecasting attempts that are cited are the weather prediction, dating back millennials in Ancient Greece and more formally weather forecasting research during the 19th century. Several forecasting models and tools have been developed for forecasting weather, earthquakes, political results, the spread of virous in medicine, macroeconomic trends, financial performance, stock exchange returns, demand and sales in supply chain management, and energy consumption. The definition of forecasting lies in the utilization of all the available information or historical data as input in an attempt to make predictions about the trends in the future behavior of the variable of interest. Forecasting techniques are divided in two categories, qualitative forecasting techniques that are based on the personal opinion of experts through interviews or analysis of data and quantitative forecasting techniques that use historical statistical data for projecting the past statistical behavior and patterns into the future. The forecasting models are also divided in categories depending the forecasting horizon, the number of periods for which the forecast is generated.

In energy management and electricity consumption forecasting methodology plays an important role since it is essential to decision making for the efficient consumption of energy sources, investment in renewable energy sources and the optimal operation of energy systems. The electricity load forecasting has drawn significant attention in academic literature and in energy business debates. As population is increasing worldwide and technology is progressing, the more electricity is expected to be consumed for cooking, heating, and operation of electrical appliances. In addition, new consumer norms and preferences such as the gradual substitution of petrol automobiles with electric and the increase in the number of electric consuming devices that make everyday life more comfortable are expected to increase the electricity demand.

The electricity load forecasting depending on the forecasting horizon can be classified into three categories; short-term forecasting (forecasted values up to a couple of weeks ahead), medium-term forecasting (forecasting horizon ranges from one month up to a year), and long-term forecasting (forecasting electricity load for a period more to one year ahead). Long-term forecasts about electricity demand are useful for managing the energy resources, taking decision about the future energy mix, examining possibilities of capacity expansion or shutting down electricity generation plants, and expanding the electricity grid. Medium-term forecasting of electricity load is used for the planning the supply chain of energy resources in electricity generation facilities, building stocks of energy fuels in period of potential supply disruptions, examining the potential revenues of the generation companies, and designing future tariffs or distribution fees. Finally, the short-term electricity load forecasting is important for facing peak load demand, managing the efficient operation of the electricity network, reduce the uncertainty for the production companies in the amount of revenues that will make in short term, and profit generation from electricity exchange with different electricity systems.

The majority of the quantitative forecasting models that we can see in academic literature can be categorized into three categories; time series models, regression analysis models, and soft computing models using simulation and artificial intelligence.

2.3.1 Time Series Models

Time series models are among the simplest models used in electricity load forecasting. Essential in this modeling and forecasting process is the availability of electricity load time series, meaning the continuous in time sequence of observations. The forecasted values of the variable of interest in time series models is made from the behavior of the previous observations. In this category we can find simple averaging models such as Moving Averages models, Exponential Smoothing models, Holt-Winters Exponential Smoothing model, and Autoregressive Moving Average Models (ARMA) (Ghalehkhondabi et al., 2016).

Time series models have been tested in academic literature for all the forecasting horizons. ARMA models were used by (Ediger and Akar, 2006) for generating the future demand for several energy sources in Turkey, including electricity, based on annual historical data from 1950 to 2004. Different autoregressive models were compared including the Autoregressive Integrated Moving Average (ARIMA) and Seasonal Autoregressive Integrated Moving Average (SARIMA) models for energy demand by source and total energy demand. Ediger and Akar concluded that forecasts for the total energy demand using the ARIMA and SARIMA models performed better that the summation of forecasts by energy sources based on criteria such as the goodness of fit and forecasting error. In short-term forecasting modelling, (Pappas et al., 2010) applied the ARMA methodology in hourly electricity load observations for two years from the Greek power system in order to forecast hourly electricity load in 2006. The proposed ARMA model was fitted using three different criteria (Akaike Information Criterion, Corrected Akaike Information Criterion, and Schwarz's Bayesian Information Criterion) finding satisfactory results in predicting week ahead hourly electricity load (short-term horizon). The ARIMA and Generalized Autoregressive Conditional Heteroskedasticity (GARCH) methodology was used by (Hor et al., 2006) to predict electricity load daily patterns by taking into account also the distribution of residuals (Ghalehkhondabi et al., 2016).

2.3.2 Regression Models

The second category of electricity modeling methodology includes regression models, one of the most used statistical methodologies for examining the causal relationship between variables and in some instances used for generating forecasts. Building a regression model requires constructing the relationships between the dependent or response variable and other independent variables that affect the dependent variable. In the case of electricity load forecasting, regression models use the historical values of the electricity load (dependent or response variable) and one or several influence or independent variables (price, weather, humidity) that have a causal relationship with electricity consumption. Regression models are distinguished into linear and nonlinear regression models accordingly to the relationship between the variables. In the case of the linear relationship between the variables of the model, regression models can be univariate models (simple linear regression models or multiple linear regression models) and multivariate regression models (the regression model examines the linear relationship between multiple dependent and independent variables) (Ghalehkhondabi et al., 2016).

In short-horizon electricity load forecasting using hourly observations, (Ramanathan et al., 1997) developed a multivariate regression model with different regression equations for each hour of day. In their approach, twenty-four different regression equations occurred for the weekdays and another twenty-four different regression equations emerged for each hour during a weekend. Adjusting the regression equations with the forecast error of the previous hour equation produced extremely good forecasting performance against a wide range of alternative models including ARMA techniques with few limitations in the presence of extreme temperature events.

An important property of regression, and in general econometric models, is the correlation of macro-economic and social variables with electricity demand. A multiple regression equation for forecasting electricity demand in N. Cyprus was examined by (Egelioglu et al., 1999). In this model, different economic variables including the electricity price, the number of customers, the number of tourists and annual electricity consumption examined highlighting the importance of developing weather sensitive models that could produce more accurate results. Adjustments of traditional regression models with time series approaches have also made. In (Harris and Liu, 1993) the dynamic relationship between electricity load and several independent variables was examined using monthly data in a multiple-input model together with ARIMA as a baseline transfer function for the error term. The synthesis of these two approaches according to the authors produced models with better understanding of the behavior of the data and the economic theory.

The nonlinear relationship between energy load and temperature was studied by (Halepoto et al., 2014). The authors found that the nonlinear behavior of electricity load at any specific hour of the day and several exogenous variables affect the forecasting performance of models. They proposed that both linear and nonlinear modeling techniques would generate more realistic results. In their paper except from the traditional simple and multiple linear regressions included quadratic and exponential regressions for modelling the effect of temperature on hourly electricity load and validating their model for different 4-day samples from the historical data.

In the methodology used in academic literature, we can find significant work in the use of panel data time series models for electricity load and electricity prices modeling. Panel data time series allows for clustering among time, and thus could provide insights about the hourly difference on the behavior of electricity load. The estimation of a linear panel data regression models assuming fixed effects was examined by Thomaidis and Biskas (2021) for analyzing the key fundamental drivers of electricity prices in Greece showing that there is heterogeneity in dynamic behavior of hourly electricity prices. The panel date time series analysis allows the examination of the dynamic relationship between the variables and captures the memory of the electricity time series process.

2.3.3 Soft Computing Models

In academic literature except of the two broad energy modeling and forecasting methods described above, there are several models developed based on computing modeling and artificial intelligence. The advantage of those models is their ability to capture the complex behavior of electricity systems, mimic energy phenomena and examine different scenarios. Several soft computing methodologies are suggested in academic literature such as Genetic Algorithms, Fuzzy Logic, Neural Networks, and Evolutionary Algorithms (Ghalehkhondabi et al., 2016). However, soft computing models are having the disadvantage that are not easily understood and cannot be used from an audience with limited programming knowledge. As in this dissertation electricity load modeling and forecasting capabilities will be performed with the use of time series and regression models analysis, soft computing modelling is not analyzed further.

3. Methodology

3.1 Data Description and Analysis

In this section a analysis of the sample data for the key variables that are going to be used is presented. Real data for electricity load are provided from the European Network of Transmission System Operators for Electricity (ENTSO-E) transparency platform in which thirty-five national transmission operators are reporting electricity generation, transportation and consumption data for ensuring the security of the European electricity grid. Actual hourly electricity load data for Spain are used for the analyzed time period covering from 01/01/2019 at 01:00 until 31/12/2019 at 24:00 (https://transparency.entsoe.eu/load-domain/r2/totalLoadR2/show, accessed:01 June 2023). The data set includes 8.760 hourly observations for the actual total electricity load in Spain with the minimum observed electricity consumption being 17.168 MW and the maximum electricity load being 40.107 MW (mean value 28.537,41 MW and standard deviation 4.524,51). Table 1 presents the descriptive statistics for the total Electricity load by hour and for the whole time series. On average, peak electricity load is observed at 13:00, while the larger standard deviation is observed at 09:00. Performing the Shapiro-Wilk and Skewness-Kurtosis test for normality in the Actual Total Electricity Load, we could not reject the null hypothesis that the variable is normally distributed at 5% level of significance for each hour of the day (with exception the 02:00 and 20:00). Thus, electricity forecasting models that are based on the assumption of non-normal distributed variable are expected to poorly perform.

Table 1 Descriptive Statistics: Electricity Load over the period 01/01/19-31/12/19

Hour	Mean	St. Deviation	Skewness	Kurtosis
01:00	25138.81	2027.491	.126546	2.456702
02:00	23792.78	1747.039	-.0457488	2.564048
03:00	23000.88	1606.076	-.2045823	2.747808
04:00	22650.34	1581.456	-.325611	2.939562
05:00	22719.36	1627.052	-.4626522	3.079485
06:00	23729.73	2051.77	-.5124323	2.658191
07:00	25673.54	2921.983	-.613535	2.19648
08:00	27952.92	3813.568	-.585012	2.151791
09:00	29752.4	4087.034	-.47225	2.36221
10:00	31021.34	3814.594	-.3896329	2.574172
11:00	31730.44	3566.857	-.3101583	2.619569
12:00	32100.55	3488.666	-.3249996	2.559539
13:00	32155.35	3447.513	-.2853374	2.532362
14:00	31780.11	3364.096	-.1587558	2.498789
15:00	30979.33	3350.07	-.2130735	2.636585
16:00	30411.38	3539.759	-.2370169	2.593737
17:00	30087.37	3619.856	-.2303898	2.532592
18:00	29917.62	3537.917	-.1842982	2.436762
19:00	30417.08	3512.613	.0132285	2.498083
20:00	31368.35	3362.576	.0941747	2.621341
21:00	31990.97	3093.376	.2518188	2.670597
22:00	30899.44	3088.565	.4683471	2.598752
23:00	28800.91	2833.951	.4284667	2.458061
24:00	26834.93	2307.325	.2235331	2.4662
Whole Sample	**28537.94**	**4524.502**	**.0599045**	**2.133783**

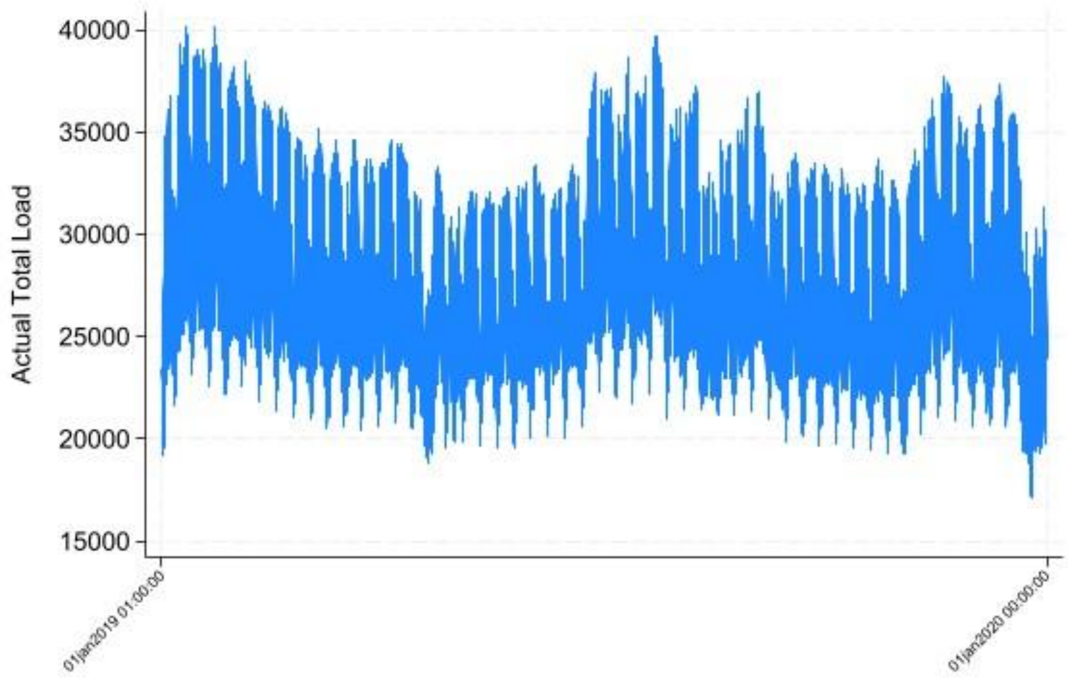


Figure 1. Hourly Electricity Load – Spain 01/01/19 at 01:00 to 31/12/19 at 24:00.

One first observation from the above time plot of electricity load we can detect seasonal periodicities. Electricity load appears to exhibit periodicities within a day, within a week and within months. Particular hours of the day appear to have significant lower electricity load compared to other, while similar patterns appear between the weekdays and the days of the weekend. Also, there is significant higher electricity consumption during summer and winter months compared with the remaining months. The seasonality in electricity load behavior is in accordance with traditional economic theory since the economic activity during the working hours of the day and the working days of the week is higher. In addition, the electricity for residential cooling and heating is higher during winter and summer months respectively. Thus, seasonality in electricity load is a factor that should be taken into account in modeling and forecasting with the adaption of dummy variables. Dummy variables are usually taking the value 1 for a specific hour/day/month or season and 0 elsewhere (with 1 dummy variable less than the number of observed facts, i.e. 6 dummy variables for the day of the week seasonality). We can see the periodicity in electricity load described, in the next box plots by hour of the day, by day of the week (1 for Monday to 7 for Sunday) and by month of the year.

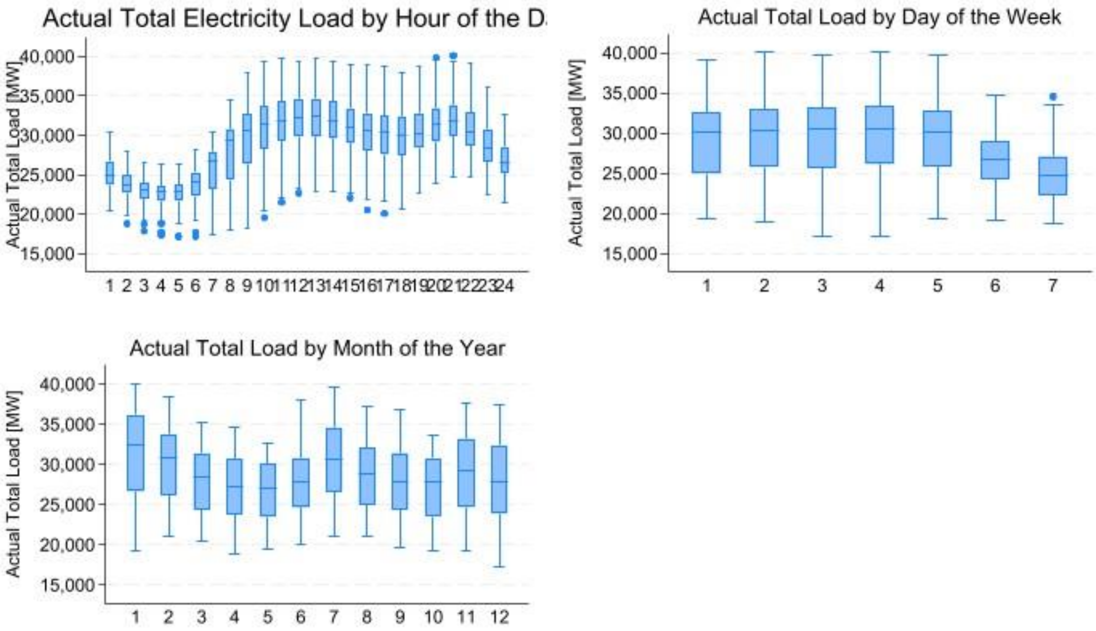


Figure 2: Box-Plots of Electricity Load by Hour of the day, by Day of the week and by Month.

In general terms, electricity load as provided by ENTSO-E will be used as the dependent variable in the majority of the models that will be examined in the next session. The main independent variables that will be examined are the day-ahead electricity price (Euros/MWh), day ahead scheduled total electricity generation and average temperature. All the variables are available on ENTSO-E transparency platform. Also, the hourly average temperature in Spain is provided by ENTSO-E transparency platform and as plotted in the graph below we can see that in Spain the temperature variates from 1 degree of Celsius to 34 degrees of Celsius on average for the whole country.

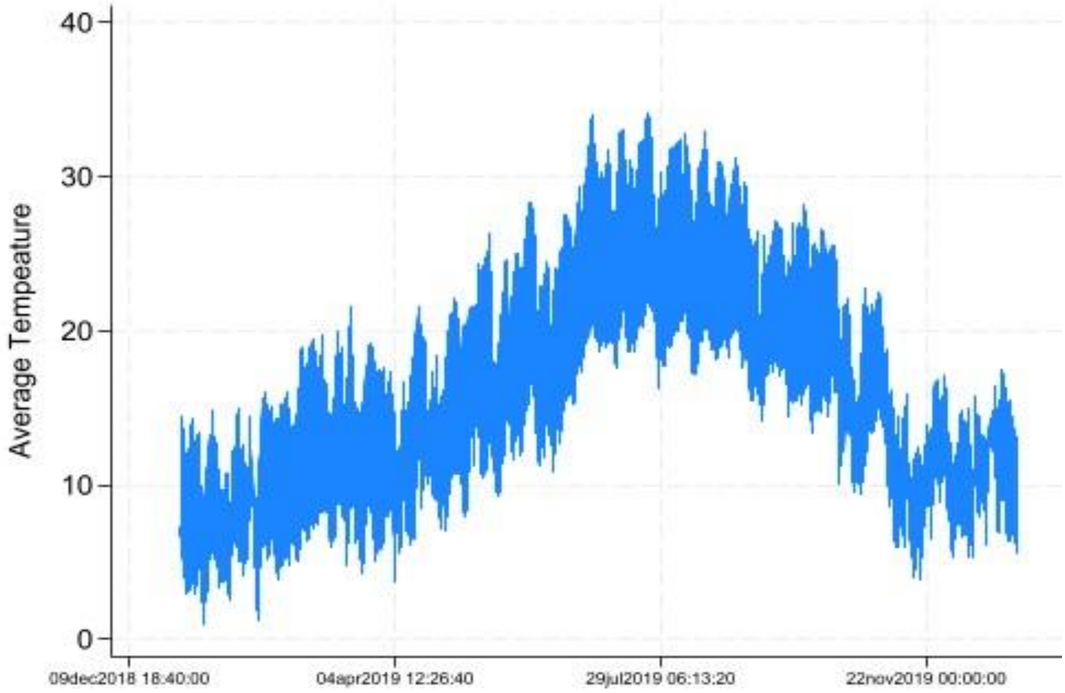


Figure 3. Average Hourly Temperature in Spain from 01/01/2019 01:00 to 31/12/2019 24:00

There are several arguments in academic literature about the effect of average temperature on electricity consumption since most models assume a linear relationship between the two variables. However, lately there have designed and examined electricity load models arguing for a non-linear relationship between electricity consumption and temperature should be used since a marginal increase from 18 degrees of Celsius to 19 degrees of Celsius or a marginal decrease from 19 degrees to 18 is not expected to have a significant result in electricity load. The non-linear relationship between temperature and electricity consumption is introduced in many research papers of the academic literature with the use of variables capturing the deviation from the temperature that will result in an increase in electricity consumption for heating or cooling. This non-linear realization of temperature is commonly named as Heating Degree Days (HDD) and Cooling Degree Days (CDD) and take into account the degrees difference in observed temperature outside from the human comfort zone. These variables are calculated as:

$$HDD_t = \max(0, T^H - T_t)$$

$$CDD_t = max(0, T_t - T^C)$$

In the formulas above T_t denotes the observed average temperature in time t while T^H and T^C denotes the threshold temperature for heat and cold, the temperature over and under which is accompanied with increase in electricity consumption because of the use of electricity appliances. In the analysis of this dissertation, the threshold temperature is set to 20 degrees of Celsius for the HDD and 15 degrees of Celsius for CDD as suggested by the Spanish Technical System Operator (Moral-Carcedo & Vicéns-Otero, 2005).

3.2 Electricity Load Modeling and Forecasting

3.2.1 Time Series Models

Among the simplest methods for generating electricity load forecasts are the averaging methods and the exponential smoothing method. These models are particular useful because they use only the available information of the historical time series of electricity load and are simple in their calculations.

3.2.1.1 Moving Average Method

One of the most common averaging methods used in forecasting is the Simple Moving Average, which takes into account the historical information of the time series to generate a forecast value. The simple moving average does not use the whole available past observations for creating a forecast but a few numbers of previous observations. One of the properties of this method is that the simple moving average is rolling over time and it “drops” information that is not relevant. The formula of the Simple Moving Average has the form:

$$Forecasted\ Electricity\ load_t = \frac{\sum_1^n Electricity\ Load_{t-i}}{n}$$

In the above equation, n is the number of periods used for generating the forecast.

3.2.1.2 Simple Exponential Smoothing Model

Simple exponential Smoothing is the second approach used in forecasting and has been applied in energy and electricity consumption based only on past observations of the time series. The exponential smoothing models are similar with the moving average models described above with the difference that a weight in the more recent observations is imposed. Also, the weights in the observations are different and exponentially decrease as more far to the history of the time series we are going (Islam et al., 2020). By denoting with α the smoothing coefficient ($0 < \alpha < 1$), the formula for the exponential smoothing moving average has the form:

$$Forecasted\ El.load_t = a * \ Elec.Load_{t-1} + (1-\alpha) * \ Forecasted\ El.Load_{t-1}$$

3.2.1.3 Holt-Winters Exponential Smoothing Model

The Holt-Winters exponential smoothing method generalized the above approach of simple exponential smoothing methodology to deal with the presence of seasonal and trend behavior in the time series. Trend can be seen as the long-term tendency of a time series, and in case of electricity we expect to be the long-term increase. Seasonality is the tendency of electricity load to exhibit a repeating behavior. In Holt-Winters model there are three exponential parameters; the smoothing coefficient, the seasonal component coefficient, and the trend component coefficient. In the short-term electricity forecasting using few periods (less than a couple of years), trend is not really observed and thus Holt-Winters seasonal model should be preferred (additive seasonality or multiplicative seasonality) (Islam et al., 2020).

3.2.2 Autoregressive Models

The advantage of Autoregressive models is that the behavior of electricity load and forecasts are captured only from the historical observations of electricity load. Autoregressive Models are built on the properties of the Exponential Smoothing and Moving Average models described above. An Autoregressive Model (AR) assumes that the current value of electricity load is a linear combination of its previous observed values (Islam et al., 2020). The estimation of AR model of order p is simple linear regression with the current value of electricity load as the dependent variable and p-lag values of electricity load as the independent variables. The representation of an $AR(p)$ model has the form:

$$Lnload_t = \beta_0 + \sum_{i=1}^{p} \beta_i * Lnload_{t-i} + \ \alpha_t$$

In the above model, β_0 is a constant, β_i are the regression coefficients of the lag values of electricity load and their number is equally to the number of the order of the AR process p, and α_t is the error term of white noise (an important property and building block in autoregressive models).

In electricity load forecasting more common are autoregressive moving average models (ARMA) in which the forecasted value of electricity load is a linear combination of lag values of the electricity load and previous white noises. An $ARMA\ (p, q)$ model without a constant term has the form:

$$Lnload_t = \sum_{i=1}^{p} \varphi_i * Lnload_{t-i} + \sum_{j=1}^{q} \beta_j * \varepsilon_{t-j}$$

In the above model, φ_i are the coefficients of the AR process of order p, β_j are the coefficients of the MA process of order q, and ε_t are the white noise terms.

The ARMA models as described above require a time series that is stationary. However, stationarity is a property that rarely exist in time series. In the case of non-stationary time series, the modelling process follows the Box-Jenkins methodology of differencing the time series to remove the non-stationarity. The methodology of integration enters the AR and MA processes, creating the Autoregressive Integrated Moving Average (ARIMA) models of order p for the AR process, q for the MA process and d for the differencing times of the time series. In available forecasting models used in academic literature for energy and electricity forecasting we can also find some other similar models such as the Seasonal ARIMA (or SARIMA) and Autoregressive Moving Average with Exogenous variables (ARMAX) (Islam et al., 2020).

The time series models and autoregressive models described have a significant advantage that they require only the knowledge of the past values of electricity load for generating forecasts. However, the most important disadvantage of these methods is that they lack in understanding and explaining the behavior of electricity load over time. Electricity load changes not only because of time but also because of other factors influencing the demand for electricity. Hence, forecasting models should examine not only the effect of time but also the effect of other variables in electricity load.

3.3 Regression and other Econometric Models

3.3.1 Multiple Linear Regression

Regression models that describe the linear relationship between one dependent or response variable and more than one independent or explanatory variables are called multiple linear regression models (Yildiz et al., 2017). In this case, the natural logarithmic value of total electricity load (lnLoad) is the dependent variable and as independent variables are used the seasonal dummies, the natural logarithmic values of day-ahead price (lnPrice) and the temperature variables (HDD and CDD). The representation of a multiple linear regression model has the form:

$$LnLoad_t = \beta_0 + \beta_1 * LnPrice_t + +\beta_2 * HDD_t + \beta_3 * CDD_t + \sum_{4}^{9} \beta_i * D_i + \varepsilon$$

In the regression equation above the β_i are the regression coefficients (including a constant term), D_i are the seasonal dummies (taking the value 1 for a specific day of the week and 0 for the remaining days of the week) and ε_t is the error term. Alternative variations in the regression equation above are including as independent variables the hour of the day, the number of the day, or/and the number of the month to capture the deterministic part of electricity load instead of dummy variables. Estimation of the regression model using the historical data provides estimates for the coefficient of each independent variable, and thus we can generate the forecasted value of lnLoad. Usually, the past observations of the variables are divided into two parts with the first part being used for the model estimation while the second part for evaluation of the forecasting accuracy (Islam et al., 2020). If the data set is relatively large, it allows rolling division of the data set into estimation and forecasting testing regions.

3.4 Postestimation Model Evaluation and Forecast Accuracy

In the previous sections was highlighted the importance of forecasting of electricity load and described the forecasting models presented in academic literature without presenting a former definition of the term forecasting. Forecasting is developing the mechanism for making predictions about the future performance of variables or occurrence of events based on the historical and current information. The forecasting process starts with the data evaluation, continues with the selection of the most suitable forecasting model (or models), the decision about the forecasting periods and the generation of the forecasts, and concludes with the monitoring and evaluation of the forecasting performance when the actual future variables will become available. Electricity load forecasts are essential for the generation facilities and electricity system operators since they are used for ensuring the electricity grid stability and are essential for the profitability of the agents in the electricity market. In this context, the accuracy of the forecasting process in really important. Usually, forecasts about demand are evaluated regularly and improvements are made in case that there are changes in the variables that degrades the forecast performance over time (Islam et al., 2020).

Important factor for the evaluation of the forecasting models is the size of the data set since a larger data set does not necessarily means a higher forecasting accuracy. After the forecasting model is chosen, the historical data are divided into two parts with the first part being used for the estimation of the models' parameters (in-sample period) and forecast generation and the second part being used for evaluating the forecasts generated with the actual values of the historical data (out-of-sample period or postestimation testing). If the forecasting accuracy measures suggest that the model captures satisfactory the future values of electricity load, then the forecasting windows are rolling for generating the forecasts for the future values. Forecasting accuracy evaluation is performed by graphing the forecasted and actual values on the data set (visual performance), and by calculating forecasting accuracy metrics. Forecasting accuracy metrics are using the forecast error which is defined as the difference between the actual value and the forecast at time t. If the difference is positive, the forecasting model underestimated the future electricity load while if the difference is negative the forecasting model overestimated the future electricity load (Islam et al., 2020). Few of the most used forecasting accuracy metrics are presented below.

Mean Absolute Deviation (MAD)

The Mean Absolute Deviation (MAD) metric is the average of the absolute values of the forecast errors. The absolute value in the calculation of this forecast accuracy measure guarantees that possible changes in the polarities of the forecast errors will not cancel each other, and as a result the MAD shows the size of the error (Islam et al., 2020). The larger the MAD, the less accurate the forecasting model since it systematically out- or over-estimate the future values of electricity load.

$$MAD = {}^{|\sum(Actual\ Value - Forecast)|}\!/_{n}$$

Mean Square Error (MSE)

The Mean Square Error (MSE) is the average of the squared errors. The larger the MSE, the less accurate the forecasting model. This performance metrics gives a bigger magnitude to large errors through squaring them.

$$MSE = {\sum(Actual\ Value - Forecast)^2}/{n}$$

Mean Absolute Percent Error (MAPE)

The Mean Absolute Percent Error (MAPE) metrics uses the mean value of the absolute percent forecast error (the division of forecast error with the actual observed value). This accuracy metric has the advantage that show as the percentage error with respect with the actual variable and thus give us better information about the quality of the error.

$$MAPE = \frac{\left|\sum(Actual\ Value - Forecast)\ /Actual\ Value\right|}{n} * 100$$

4. Results

4.1 Time Series Analysis

4.1.1. Simple Moving Average Models

The first models that will be examined for their ability to model electricity load are the simple averaging models; models that are considered among the simplest and are extensively applied for forecasting. Firstly, the simple moving averages will be calculated using all the hourly electricity load observations by dividing the sample in two parts. The simple moving average of 12-hour, 24-hour, 36-hour, 48-hour, 72-hour, 96-hout, 120-hour, 144-hour and 168-hour are calculated. Following the description of the simple moving average models described in Chapter 3, the 12-hour simple moving average or SMA(12) uses the previous 12 hourly observations for producing the forecast for the next period. Similarly, the SMA(24) used the previous 24 hourly observations (previous day) while the SMA(168) uses the 168 previous hourly observations (7-days). The SMA mentioned are made from the beginning of the sample and forecast accuracy metrics are calculated for the SMA electricity load forecasts for the period 01/08/2019 01:00 to 31/08/2019 24:00.

A first assessing of simple average models is usually plotting the forecasts created against the actual observations when they are becoming available or by dividing the sample in two parts (estimation sample and forecast testing sample). In Figure 4 below, the Simple Moving Averages (SMA) and Actual Electricity Load for the period 01/08/2019 00:00 to 07/08/2019 23:00 are presented. We observe that using all the available hourly data from the previous periods, the larger the number of previous observations used (or the longer the model's memory), the smoother the forecasts created and the less are following the patterns of the original time series.

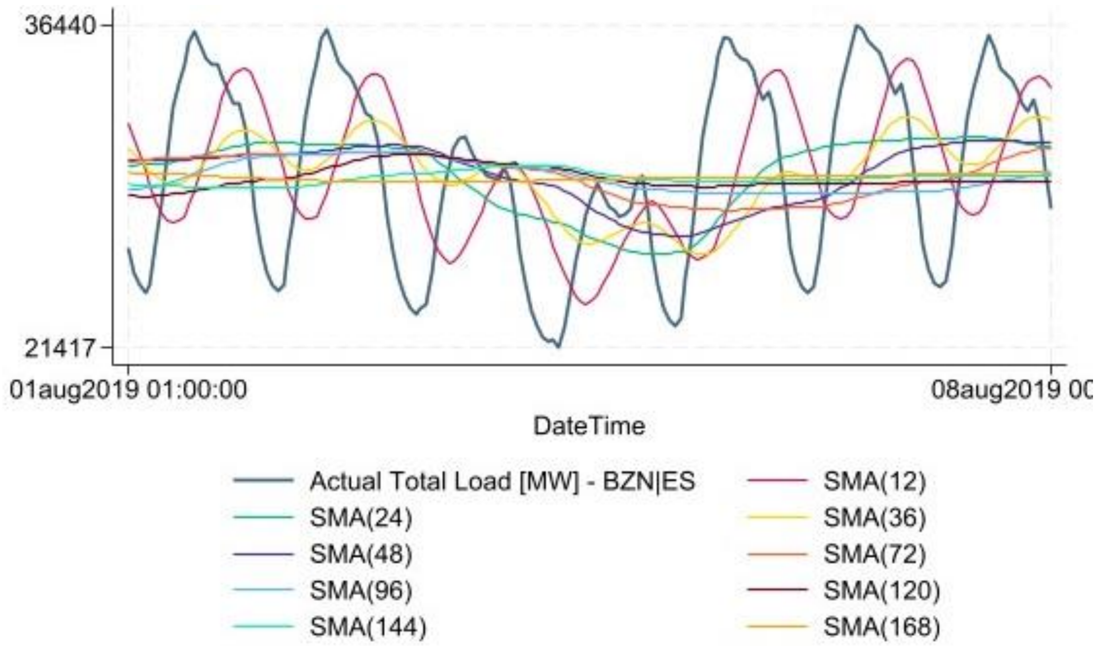


Figure 4.: SMA Forecasts and Actual Electricity Load

A second assessment of each SMA models' forecasting performance and ability to capture the behavior of the time series is through the calculation of the forecasting accuracy measures. For each SMA model, the forecasting error (difference between actual electricity load observed and the forecasted electricity load) were calculated. The forecasting accuracy metrics of Mean Absolute Deviation (MAD), Mean Absolute Percentage Error (MAPE), and Mean Square Error (MSE) for each SMA are presented in the Table 2 below. The SMA(168) generated the forecast for electricity load on August with the smaller forecasting accuracy metrics value since in average the forecast deviation from the actual electricity load was 2.634 MWh and the forecast error compared to the observed actual value was 9,82%.

	MAD	MSE	MAPE
SMA(12)	3566,22	18394500	13,26%
SMA(24)	2975,50	14706050	11,46%
SMA(36)	2882,53	14117199	11,18%
SMA(48)	2840,85	14556145	11,05%
SMA(72)	2859,46	14085077	11,07%
SMA(96)	2808,67	13292444	10,82%
SMA(120)	2808,31	13169701	10,79%
SMA(144)	2683,51	11312169	10,12%
SMA(168)	2634,02	10552668	9,82%

Table 2: Forecasting accuracy measures of the SMA

Another approach that could provide more accurate forecasts is calculating the simple moving average only for the specific hour of the day, thus calculating 24 SMA forecasts one for each hour of the day. For the hourly simple moving average forecasts, we are calculating the SMA(2), SMA(3), SMA(4), SMA(5), and SMA(6). The SMA(2) at 01:00, for example, generates the forecast for the specific hour of the day by computing the average of electricity load at 01:00 of the previous two days. Figure 5 below, illustrates the simple moving averages at 01:00 for the period 01/08/2019 to 31/08/2019. As we can see, the simple moving averages mimic the pattern of the actual electricity load at 01:00 with a delay. As we increasing the number of the observations used for deriving the SMA, the smoother the forecast. In Appendix A, Graph A.1 we followed the same procedure for more representative hours of the day.

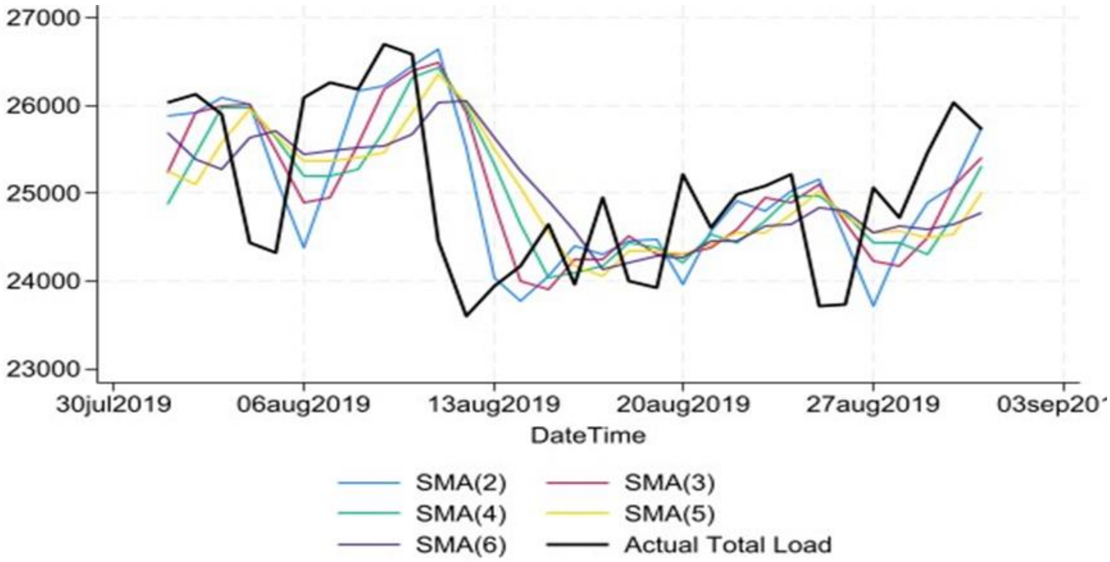


Figure 5: Hour Specific SMA forecast

4.1.2. Autoregressive Moving Average (ARMA) models

Another modeling approach for electricity load are the models that belong in the Autoregressive family methodology in which the response variable is dependent on its previous time observations (lag values). The autoregressive models come in many variations, as described in the literature review with ARMA (Autoregressive Moving Average models) being the most common for electricity load modeling and forecasting. The base rationale behind ARMA models for electricity load modeling is the idea that electricity load behavior can be considered as random and influenced by different non-deterministic factors and processes that cannot be modeled. However, one of the challenges in Autoregressive models is choosing among the various modelling options (AR, MA, ARMA, ARIMA, ARMAX, SARIMA) and model identification (optimal lags for the AR, MA, and Seasonal Differentiation).

One basic assumption for applying ARMA modelling approach in electricity load is the stationarity property of the time series. A time series process is stationary when exhibits a constant mean and a constant variance. The procedure that will follow for examining the ARIMA (p,d,q) approach on modeling electricity demand is using the hourly electricity load observations for Spain from 01/01/2019 01:00 to 31/07/2019 23:00 for model identification and estimation of parameters, and the observations from 01/08/2019 01:00 to 31/08/2019 24:00 will be used for postestimation testing of forecasting accuracy. Testing the sample for stationarity by the Augmented Dickey-Fuller test gives as a Z(t) statistic of -10,850 and a corresponding p-value of 0,000 (reject the null hypothesis that the Total Electricity Load follows a random walk process with or without a drift). The same conclusion we get and from Phillips-Perron unit-root test which uses Newey-West standard errors for serial correlation (the unit-root test shows a Z(t) statistic of -15,710 at 9 Newey-West lags and a corresponding p-value of 0,000 providing us with sufficient evidence for rejecting the null hypothesis of a random walk with or without drift).

From the unit-root tests we conclude that the hourly electricity load time series is a stationary process and thus there is no need for differencing and thus the parameter d=0 in the ARIMA model specification. For determining the number of the parameter p for the AR process and the parameter q for the MA process, statistical software such as STATA determines the optimal number of lags by fitting several different models and suggests the best model according to the minimization of information criteria. Based on Akaike’s Information Criterion (AIC), Schwarz’s Bayesian Information Criterion (BIC), and the Hannan and Quinn Information Criterion (HQIC) an ARIMA (2,0,2) model can be used for modeling electricity load. Plotting the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) should agree with the lags specified from the minimization process of the information criteria.

In the ACF graph in GA.3 of the Appendix A, we observe that the autocorrelation function for natural logarithm of electricity load for the estimation part of the sample does not decay exponentially but there is clear evidence of seasonal behavior on the time series since there is a repeating pattern in the autocorrelation during a stable number of lags. Thus, a model that is taking into account the seasonal behavior of electricity load during a day should be preferred. The $SARIMA\ (p, 0, q, s)$ has been used in a large number of academic researches, with s being the parameter for seasonal differencing. From the PACF plot we can see that the third lag is falling with the interval and hence a 2 lag AR and MA process might be sufficient for the estimation of the SARIMA specification. A $SARIMA(2,0,2,12)$ is estimated for the data sample from 01/01/2019 to 31/07/2019 and forecasts for August 2019 are plotted in Graph G.A.3 of the Appendix A. We observe that the SARIMA model is following the seasonal effects of the actual electricity load data behavior very closely. In addition, the $SARIMA(2,0,2,12)$ produced forecasts with a MAD of 1.927 MWh and MAPE of 7,06%.

4.2 Regression Models

4.2.1 Multivariate Linear Regression (MLR) Models

MLR with HDD/CDD as variables for the weather effect on electricity load

There are several different versions of regression models presented in academic literature varying in the choice of dependent/independent variables, statistical specifications and assumptions. As the aim of this dissertation is to analyze and forecast electricity load; the response variable will be the natural logarithm of total electricity load. The independent variables that will be used are the day-ahead price, the 24-hour average day a-head price, the temperature effect (average temperature or the HDD and CDD temperature variables), and dummy variables for modeling the seasonal effects. The electricity load exhibits daily seasonality with lower consumptions during weekends. Thus, in the model will be included six dummy variables taking the value 1 or 0 according to the day of the week for Tuesday to Sunday. For the effect of each specific season of the year (winter, spring, summer and autumn) on electricity load, since the observations are for one year and we divide the sample in estimation and out-of-sample forecasting samples, we will include in the model the variable month taking the values 1 to 12 for each month of the year. Thus, the model can be written as the equation below:

$$LnTotalLoad_t = \beta_0 + \beta_1 * LnPriceDA_t + \beta_2 * LnAver24Price_t + \beta_3 * LnHDD_t + \beta_4 * LnCDD_t + \beta_5 * D_{Tuesday} + \beta_6 * D_{Wednesday} + \beta_7 * D_{Thursday} + \beta_8 * D_{Friday} + \beta_9 * D_{Saturday} + \beta_{10} * D_{Sunday} + \beta_{11} * Month + \varepsilon_t$$

For the estimation of the coefficients β_i in the above linear regression model we will use the hourly observations of each variable from 01/01/2019 00:00 until 31/07/2019 24:00 and then we will test the model's forecasting ability by testing out-of-sample forecasts for August 2019 with the actual observations. Using 5.064 hourly observations we get the regression output using ordinary least squares (OLS) from STATA as shown in Table 3 below.

Source	SS	df	MS
Model	73.8699307	11	6.71544825
Residual	56.2260521	5,052	.011129464
Total	130.095983	5,063	.025695434

Number of obs	=	5,064
F(11, 5052)	=	603.39
Prob > F	=	0.0000
R-squared	=	0.5678
Adj R-squared	=	0.5669
Root MSE	=	.1055

lnTload	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
lnDAPrice	.3697038	.0094241	39.23	0.000	.3512285	.3881792
lnaver24	-.1620117	.0149977	-10.80	0.000	-.1914136	-.1326098
lnhdd	-.0461783	.0019224	-24.02	0.000	-.049947	-.0424095
lncdd	.0886959	.0022402	39.59	0.000	.0843041	.0930877
d_tue	.0351896	.0056756	6.20	0.000	.024063	.0463162
d_wed	.0431589	.0056	7.71	0.000	.0321804	.0541373
d_thur	.040418	.005607	7.21	0.000	.0294259	.0514102
d_fri	.0299117	.0056311	5.31	0.000	.0188723	.0409512
d_sat	-.0543751	.0056582	-9.61	0.000	-.0654677	-.0432825
d_sun	-.1135582	.0056687	-20.03	0.000	-.1246714	-.1024451
month	-.0360607	.0011562	-31.19	0.000	-.0383274	-.033794
_cons	9.602672	.0511542	187.72	0.000	9.502387	9.702956

Table 3.: OLS Regression Output MLR models with HDD/CDD

The results of the estimation show a causal linear positive relationship between the electricity load and the price at time t, while there is a negative relationship with the 24-hour average electricity price. In addition, the model suggests that during the weekend electricity load is decreasing (seasonal effect). Overall, 56,78% of the variance of electricity load can be predicted from the independent variables used, according to the R-squared statistic. The

F-statistic of 603,39 and the p-value of 0,00 suggests that we can reject the null hypothesis that the independent variables jointly do not predict the behavior of the dependent variable at 0,05% level of significance. Also, the p-values of the coefficients of each variable are suggesting that we can reject the null hypothesis that are not statistically significant at 0,05 level of significance. Hence, the regression equation can be written as:

$$LnTotalLoad_t = 9{,}60 + 0{,}369 * LnPriceDA_t - 0{,}162 * LnAver24Price_t - 0{,}0461 * LnHDD_t + 0{,}0351 * LnCDD_t + 0{,}0351 * D_{Tuesday} + 0{,}0431 * D_{Wednesday} + 0{,}044 * D_{Thursday} + 0{,}0299 * D_{Friday} - 0{,}0543 * D_{Saturday} - 0{,}113 * D_{Sunday} - 0{,}036 * Month$$

From the above estimation of the model, forecasts for the hourly total electricity load for the period 01/08/2019 00:00 to 31/08/2019 were produced and compared against the actual total electricity load. Plotting the forecasted and actual values of electricity load for August in Figure 6 below provide us with a first assessment of the forecasting ability of the model.

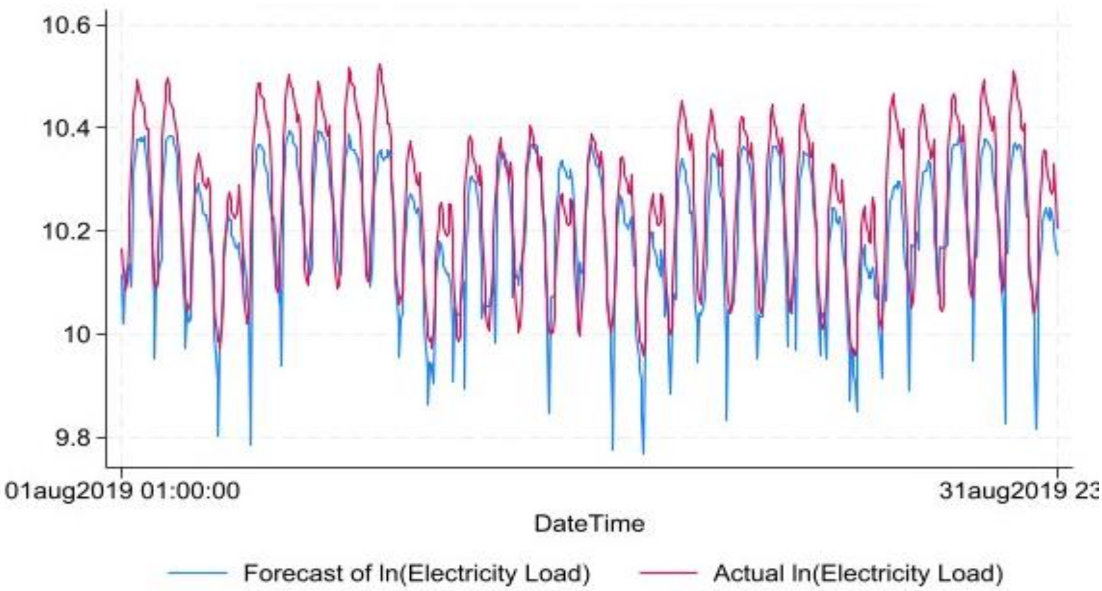


Figure 6: Postestimation Forecasted and Actual Electricity Load for August 2019

However, the above graph with the forecasts created by the estimation of the time series of hourly electricity load might appear that fits the actual data pattern on August, if we give a closer look to specific single-hour we see contradicting results. For example, plotting the forecasts at 06:00, 09:00, 15:00 and 22:00 on August 2019 in the following Figure 7, reveals different forecast error behavior subject to the hour of the day.

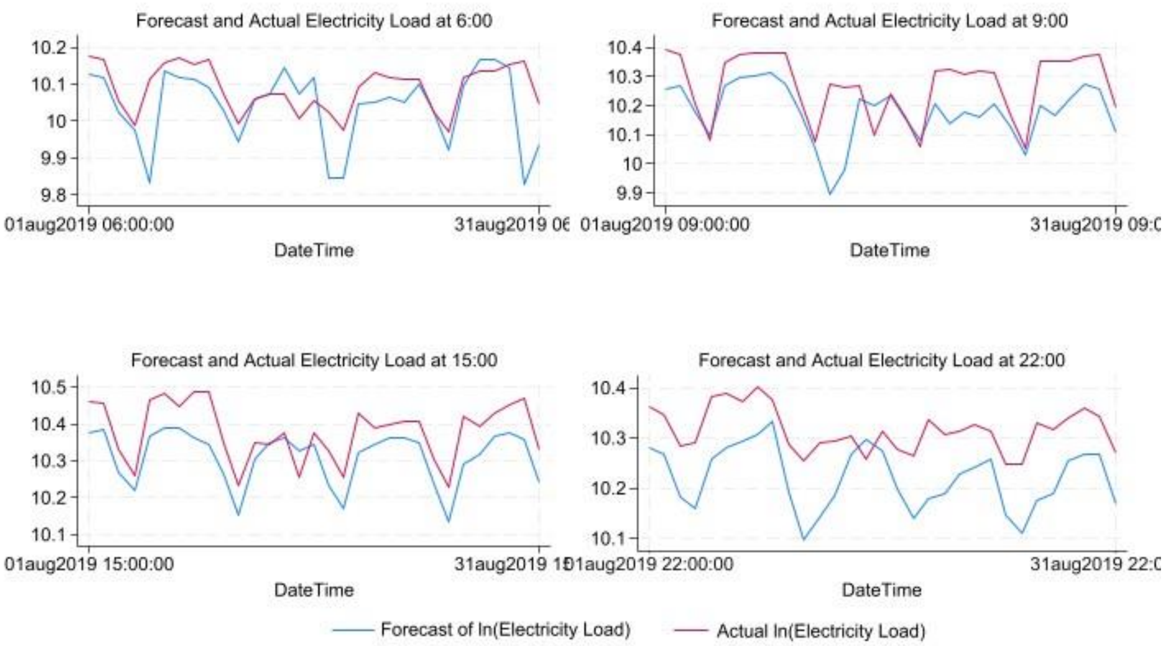


Figure 7: Hour Specific postestimation forecasts of the OLS model with HDD/CDD August

Since we have enough data for rolling the estimation and forecast testing window, a second estimation of the parameters of the linear regression model from 01/01/2019 at 01:00 until 31/08/2019 at 00:00 follows. The regression output from the statistical software is shown in the Table B1 in Appendix B. The relationships between the dependent and independent variables are similar with a small variation in the size of the estimated coefficients. The number of hourly observations is 5,807 and the R-squared is 0,5813. Again, the F-statistic of the model has a p-value of 0,000 mining that we have sufficient evidence to reject the null hypothesis that all the coefficients of the model are no-statistically significant. Also, the linear regression output for the coefficients of the model shows the same relationship between the dependent and independent variables as the previous estimation with the smaller sample.

In terms of forecasting ability, the model is tested by creating forecasts for the period 01/09/2019 00:00 until 31/09/2019 23:00 and tested with the actual values of electricity load observed during this period. In Graph A2 on Appendix A are plotted the actual and forecasted values of electricity load. The forecasts on September show similar behavior with the previous model (as for specific hour every day) and larger forecast error compared. The forecasting accuracy metrics for August is a MAD of 2.062 MW and a MAPE of 7,04% while for September the MAD is 3.141 MW and the MAPE is 10,85%. We can conclude that the estimated model is exhibiting a similar relationship between the variables t the testing sample.

MLR with Average Temperature as a variable for the weather effect on electricity load

In the above estimations of the model, we modeled the effect of the weather conditions (variation in the temperature) in the electricity consumption using the Heating Degrees Day (HDD) and Cooling Degrees Day (CDD) following the non-linear causal relationship suggested in literature. However, we should examine if the average temperature has a linear effect on electricity consumption. Thus, in the regression model for the estimation window January-July we estimate the model by dropping from the analysis the HDD/CDD variables and using the hourly average temperature observed in Spain (natural logarithm). The regression output from the STATA statistical software is shown in Table 4.

Source	SS	df	MS			
				Number of obs	=	5,064
				F(10, 5053)	=	603.25
Model	70.7953071	10	7.07953071	Prob > F	=	0.0000
Residual	59.3006757	5,053	.011735736	R-squared	=	0.5442
				Adj R-squared	=	0.5433
Total	130.095983	5,063	.025695434	Root MSE	=	.10833

lnTload	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
lnDAPrice	.3666247	.009689	37.84	0.000	.34763	.3856194
lnaver24	-.0928263	.0152571	-6.08	0.000	-.1227368	-.0629158
lntemp	.1840704	.0044585	41.29	0.000	.1753299	.192811
d_tue	.0300658	.0058265	5.16	0.000	.0186433	.0414882
d_wed	.0400901	.0057508	6.97	0.000	.0288161	.0513641
d_thur	.0403311	.0057577	7.00	0.000	.0290435	.0516187
d_fri	.0319135	.0057832	5.52	0.000	.020576	.043251
d_sat	-.0561783	.0058104	-9.67	0.000	-.0675691	-.0447874
d_sun	-.1119141	.0058195	-19.23	0.000	-.1233228	-.1005054
month	-.0426788	.0013019	-32.78	0.000	-.045231	-.0401266
_cons	8.887452	.051229	173.48	0.000	8.787021	8.987883

Table 4. Regression Output of OLS with average temperature as independent variable.

The R-squared statistic in the new linear regression estimation is 0,5442 showing that 54,42% of the variance in electricity load is explained from the model independent variables. The model has smaller R-squared statistic from the previous one with a positive relationship between average temperature and electricity load. We can write the estimated model as:

$$LnTotalLoad_t = 8{,}887 + 0{,}366 * LnPriceDA_t - 0{,}092 * LnAver24Price_t + 0{,}184 * Lntemp_t + 0{,}03 * D_{Tuesday} + 0{,}040 * D_{Wednesday} + 0{,}040 * D_{Thursday} + 0{,}031 * D_{Friday} - 0{,}056 * D_{Saturday} - 0{,}111 * D_{Sunday} - 0{,}042 * Month$$

After the estimation of the parameters, we generate the out-of-sample forecasts for August 2019 and we compare them against the actual observations. From even the simple plot we can see that this model produces less accurate forecasts compared with the model using the HDD and CDD as independent variables for the same period. In terms of forecasting accuracy metrics, we have a MAD of 3.120 MW and a MAPE of 10,29%.

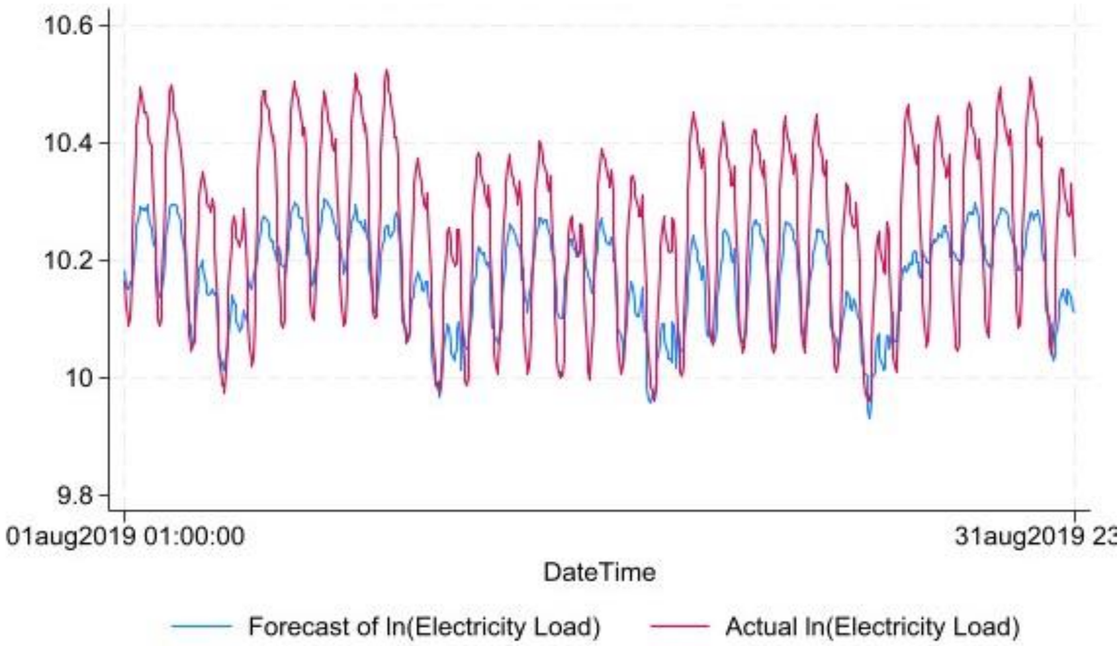


Figure 8: Postestimation Forecasts of Electricity Load on August 2019 Model 2

MLR Model for each hour of the day

As we observed in the above analysis, the multiple linear regression time series models using the hourly observations from 01 January 2019 01:00 until 31 July 2019 23:00 produced forecasted electricity load values that exhibited an hour-specific forecasting error variability. Thus, models that are taking into account the specific hour characteristics of electricity load are expected to perform better. From the two specifications examined, the model that included the HDD and CDD temperature variables performed better compared to the alternative model that used the average temperature in terms of R-squared value but also has lower MAD and MAPE in the forecast testing region. The model for i= 1, 2, ..,24 takes the form as shown in the equation bellow:

$$LnTotalLoad_{it} = \beta_{0i} + \beta_{1i} * LnPriceDA_{it} + \beta_{2i} * LnAver24Price_{ti} + \beta_{3i} * LnHDD_{ti} + \beta_{4i} * LnCDD_{ti} + \beta_{5i} * D_{Tuesday} + \beta_{6i} * D_{Wednesdayt} + \beta_{7i} * D_{Thursday} + \beta_{8i} * D_{Friday} + \beta_{9i} * D_{Saturday} + \beta_{10i} * D_{Sunday} + \beta_{11i} * Month + \varepsilon_{ti}$$

Following the similar procedure as above, we estimate the 24 different hourly models for the time period 01 January 2019 to 31 July 2019 using Ordinary Least Squares for each hour independently. In the table below, there are reported the overall model fit statistics.

HOUR	NUMBER OF OBSERVATIONS	F-TEST	R-SQUARED	ADJUSTED R-SQUARED	ROOT MSE
1	211	16,57	0,4780	0,4492	0,06328
2	211	14,40	0,4432	0,4125	0,05921
3	211	15,10	0,4550	0,4249	0,05458
4	211	18,03	0,4992	0,4715	0,05096
5	211	23,06	0,5603	0,5360	0,04792
6	211	37,42	0,6741	0,6561	0,04944
7	211	59,18	0,7659	0,7529	0,05746
8	211	81,22	0,8178	0,8078	0,06193
9	211	94,67	0,8396	0,8307	0,05785
10	211	109,59	0,8583	0,8505	0,04857
11	211	104,28	0,8522	0,8440	0,04599
12	211	91,27	0,8346	0,8254	0,04761
13	211	70,74	0,7964	0,7851	0,05205
14	211	50,65	0,7368	0,7223	0,05903
15	211	45,71	0,7164	0,7008	0,06338
16	211	49,06	0,7306	0,7157	0,06594
17	211	60,57	0,7700	0,7573	0,06275
18	211	96,94	0,8427	0,8340	0,05103
19	211	89,07	0,8312	0,8218	0,05264
20	211	101,8	0,8491	0,8408	0,04712
21	211	141,7	0,8868	0,8805	0,03571
22	211	97,39	0,8433	0,8347	0,04135
23	211	41,00	0,6938	0,6769	0,05728
24	211	25,74	0,5872	0,5644	0,05922

Table 5: Test Statistics for each OLS estimated Hourly Equation

The F-value in all models provided evidence allowing us to reject the null hypothesis that the independent variables do not reliably predict the behavior of the response variable (electricity load) at α=5% level of significance. The individual hypothesis testing for the significance of the estimated coefficients of each variable varied within each model. As we seen from Table 5 above, the estimated linear regression model for 10:00, 11:00 and 21:00 exhibit an R-squared values suggesting that more than 85% of the variance in electricity load can be predicted by the independent variables included.

After the estimation of the 24 linear equations, post-estimation forecasts were extracted for the month of August. In Table 6 below are presented the forecasting accuracy metrics for each equation. As we observe, the models have a MAPE considerably lower than the single equation model using all the hourly observations of the variables. The overall average MAPE for August 2019 from the 24 equations is 3,84%.

HOUR	MAD	MSE	MAPE
1	1113,22	1596882,83	4,44%
2	869,98	1111467,62	3,64%
3	659,34	566895,06	2,84%
4	587,57	471766,54	2,58%
5	487,80	395318,23	2,15%
6	773,63	1268501,66	3,25%
7	972,85	2258561,57	3,93%
8	1208,31	2498519,14	4,63%
9	820,36	2142644,83	2,94%
10	881,48	2182381,04	3,02%
11	981,23	2027852,40	3,15%
12	1190,71	2452962,87	3,64%
13	1353,00	2743570,91	4,05%
14	1961,03	4652071,18	5,90%
15	1845,17	4217468,21	5,68%
16	1727,63	3820210,57	5,41%
17	1480,15	3049431,18	4,72%
18	1086,27	2028323,08	3,57%
19	966,36	1583011,67	3,23%
20	904,03	1316571,06	3,05%
21	1229,32	1998831,18	3,95%
22	1007,19	1621100,91	3,38%
23	1351,33	2393769,58	4,81%
24	1072,73	1994124,18	4,08%

Table 6: Forecast Accuracy Metrics for Each Hourly estimation of Model 1

Plotting the out-of-sample forecasted value of electricity load at specific hours during August 2019 (as we did in the single equation estimation above) for comparison are presented in Figure 9 below. The forecasts of electricity load at 6:00 (from Equation 6), at 09:00 (from Equation 9), at 15:00 (from Equation 15) and at 22:00 (from Equation 22) are produced from the statistical software. Clearly, compared with Figure 7 we can see that the distance between the forecast electricity load and the actual electricity load lines are smaller (smaller forecast error). Also, the forecasts are following closer the time series patterns of the actual observations compared to the multiple linear regression model of the whole hourly range.

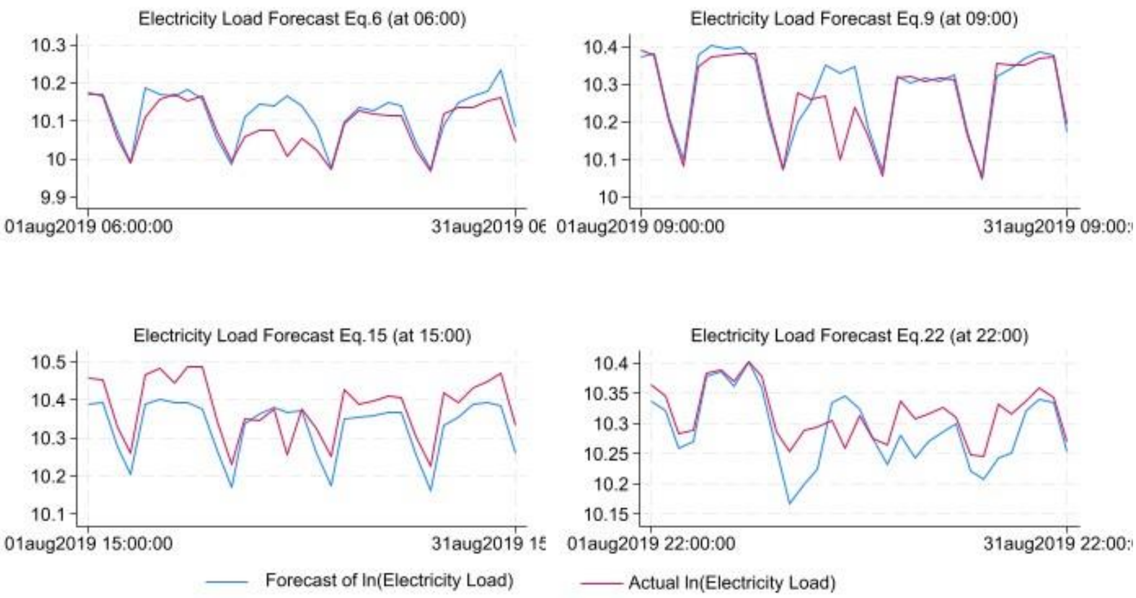


Figure 9: Forecasts of Electricity Load at Different hours

Examining again the alternative model for the climate effect as we did in the single equation analysis, the 24-hour single regression equations using as an independent variable the average hourly temperature instead of the HDD and CDD produced the following overall model fit statistics presented in Table 10 below. As we expected, the single equation models provided different results depending on the hour of the day. For example, at 09:00 the regression output suggests that 80% of the variance of the response variable (electricity load) can be predicted from the independent variables. However, overall, the models have lower R-squared values and higher Root Mean Squared Error comparing with the other 24

Multivariate Linear Regression equations provided clearly evidence that the first should be preferred.

HOUR	F-TEST	R-SQUARED	ADJUSTED R-SQUARED	ROOT MSE
1	17,47	0,4662	0,4395	0,06383
2	15,57	0,4378	0,4097	0,05935
3	16,5	0,4521	0,4247	0,05459
4	19,37	0,4920	0,4666	0,05120
5	24,48	0,5504	0,5279	0,04834
6	42,09	0,6779	0,6618	0,04903
7	66,75	0,7694	0,7579	0,05688
8	85,5	0,8104	0,8010	0,06302
9	87,16	0,8131	0,8040	0,06224
10	73,26	0,7855	0,7748	0,05961
11	58,11	0,7439	0,7311	0,06038
12	46,66	0,7000	0,6849	0,06397
13	40,12	0,6673	0,6507	0,06636
14	31,27	0,6099	0,5904	0,07168
15	28,16	0,5847	0,5639	0,07651
16	30,86	0,6068	0,5871	0,79470
17	32,79	0,6211	0,6022	0,08034
18	36,68	0,6471	0,6295	0,07624
19	46,72	0,7002	0,6852	0,06997
20	44,97	0,6921	0,6768	0,06714
21	44,35	0,6892	0,6737	0,05902
22	32,03	0,6156	0,5964	0,06461
23	25,4	0,5595	0,5375	0,06853
24	19,29	0,4909	0,4655	0,06560

Table 7: Regression Statistics of each OLS estimated hourly equation Model 2

4.2.2 Time Series Panel Data Analysis

As we highlighted in the time series analysis and the simple regression using the whole hourly observation data set, the electricity load does not follow exactly a time series process since different hour within the day have different characteristics. Thus, the electricity load can be seen as 24 different time series. In this case, estimation of the linear relationship between the dependent variable and the independent variables could be made through panel data analysis which allows each hour of the day to be a separate time series. Also, panel data time series analysis examines the dynamic relationship between the variables by fitting the fixed-effects or random effects instead of separated OLS as we did in the section above.

The dependent variable is the total electricity load (natural logarithmic value) and the set of independent variables used are the natural logarithm of the day ahead price, the average 24hour price, the weather variables (Heating Degrees Days and Cooling Degrees Day) and the dummies used for modeling the seasonality effect. The panel time series is unbalanced since its missing one value of the average 24hour price from each panel. Balancing the panel by dropping the observations of the first day of the year and performing the Hausman specification test, the p-value suggests fitting a fixed effects model. The estimation results for the same choice of variables with the simple linear regression above from 02.01.2019 until 31.07.2019 are presented in the next table.

```
Fixed-effects (within) regression               Number of obs     =      5,063
Group variable: hour                            Number of groups  =         24

R-squared:                                      Obs per group:
     Within  = 0.6051                                         min =        210
     Between = 0.7578                                         avg =      211.0
     Overall = 0.3680                                         max =        211

                                                F(11, 5028)       =     700.37
corr(u_i, Xb) = 0.1099                          Prob > F          =     0.0000

------------------------------------------------------------------------------
      lnTload | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
   lnDAPrice |   .1341006    .006992    19.18   0.000     .1203933    .1478079
       lnhdd |   .0185341   .0016728    11.08   0.000     .0152548    .0218134
       lncdd |   .0364919   .0017111    21.33   0.000     .0331375    .0398464
    lnaver24 |   .0985795   .0103083     9.56   0.000     .0783708    .1187883
       d_tue |   .0194727   .0036961     5.27   0.000     .0122267    .0267188
       d_wed |   .0254683   .0036511     6.98   0.000     .0183105    .0326261
      d_thur |   .0261684   .0036507     7.17   0.000     .0190115    .0333253
       d_fri |   .0152723   .0036664     4.17   0.000     .0080846      .02246
       d_sat |  -.0789016   .0036937   -21.36   0.000    -.0861428   -.0716604
       d_sun |  -.1342004   .0037007   -36.26   0.000    -.1414555   -.1269454
       month |  -.0038553   .0009365    -4.12   0.000    -.0056912   -.0020193
       _cons |    9.35217   .0337903   276.77   0.000     9.285926    9.418413
-------------+----------------------------------------------------------------
     sigma_u |  .11052215
     sigma_e |  .06859921
         rho |  .72189276   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(23, 5028) = 300.87                 Prob > F = 0.0000
```

Table 8. Panel Data Regression Output (Fixed-Effects)

After the estimation of the parameters for the 24-hourly clusters, we can produce the forecasts for the August 2019. A visual representation of the forecasts of electricity load and the actual observations from the historical data are presented in the next graphs.

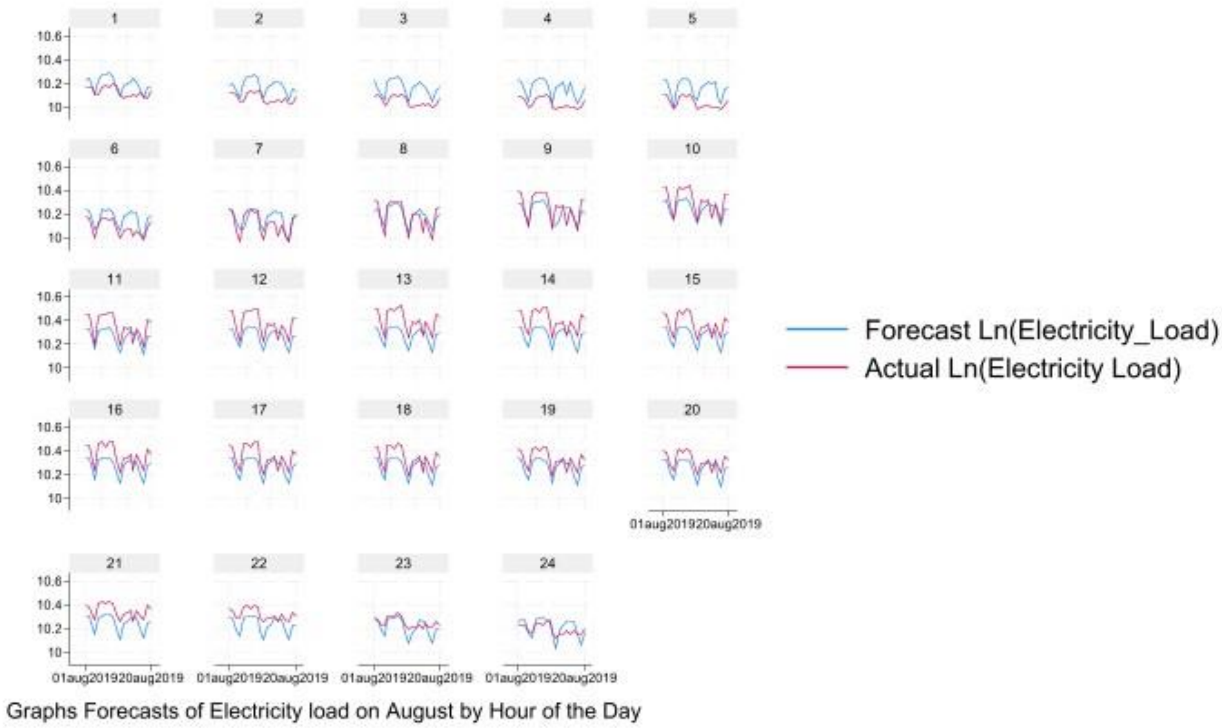


Figure 10: Postestimation forecasts of the Panel Data

Similarly, the panel model described above is estimated for the same period using instead of the Heating Degrees Day and Cooling Degrees Day non-linear realization of temperature conditions, the hourly average temperature (natural logarithm). Again, forecasts are created for model comparison in Forecasting Accuracy section.

4.3 Forecasting Accuracy Comparison

In the previous section we developed several different models for analyzing the behavior of electricity load in Spain using hourly observations. We used the first 7 months of the observations to calibrate the models and in turn we created the electricity load forecasts for August. The next step in forecasting methodology is comparing models' forecasting accuracy by calculating the forecast error from the historical observations of electricity load on August. A first comment for each model forecasting behavior was made in the previous sections from the graphical representation of actual and forecasted values of electricity load.

4.3.1 Simple Moving Average Models

The first forecasting methods that are examined are the simple moving average models. The first SMA that calculated from the whole data set of hourly electricity load used the previous 12, 24, 36, 48, 72 and 96 observations. The SMA that calculated for the specific hour at each day used the previous 2, 3, 4, 5 and 6 previous hourly observation of electricity load. Among the eleven SMA according to the Mean Absolute Deviation (MAD) forecasting accuracy metric, the smaller value is reported from the SMA(6) calculated from the same hour observed electricity load. In addition, all the SMA models calculated from previous specific hour averages performed better compared to the SMA calculated from the whole set of hourly observations.

4.3.2 Multiple Regression Models

The second forecasting method that was examined in results was the simple regression of the hourly electricity load observations. Two different regression models were estimated with the key difference between the two models being the temperature variable (HDD and CDD in the one model and average hourly temperature in the other). The models were estimated for the period 01/01/2019 01:00 to 31/07/2019 24:00 and then forecasts were created for August 2019. The model with the smaller MAD of 2.075,108 was the model with the model that were using the non-linear temperatures of HDD and CDD. The model using the average hourly temperature created forecasts that gave a MAD of 3.135,17. The Mean Absolute Percentage Error of the regression model using HDD/CDD in the independent variables was 7,06%, while the alternative model tested exhibit a MAPE of 10,31%.

4.3.3 Panel Time Series Models

Two versions of panel data were examined, one with the inclusion of the hourly average temperature as independent variable and one with the inclusion of Heating Degree Days and Cooling Degrees Day. Forecasts were created from the estimated models from historical data from 02.01.2019 to 31.07.2019 with hourly clusters and compared against the actual electricity load observations on August (forecasting window of historical data). Again, the model that used the HDD/CDD temperature variable created better forecasts in terms of MAD and MAPE criteria. The MAD for the model was 2.540,4 against the second that had a MAD of 3.476,86 (MAPE of 8,7% and 12,09% accordingly for each model). Interesting in the comparison between the regression models is that the pane data approach produced post estimation forecasts that lack against the simple multivariate regression models.

5. Conclusions

Nowadays, energy consumption and especially electricity consumption is important for economic activity and household everyday life since the expansion of technology has created many devices that operate only with electricity. The importance of securing the efficiency and reducing operational risk of the electricity grid has been cited from the academic literature, market experts and policy makers. Forecasting the electricity consumption has become crucial for the designing of long-term energy policies and the continuous decarbonization of global energy mix. As the integration of renewable sources of energy such as solar and wind power will continue to grow contributing in the decarbonization of the European energy mix, efficient electricity load models should take into consideration the intermitted and uncertain nature of renewable sources of energy. Thus, models that incorporate different weather data (such as temperature, solar radiation and wind speed) together with electricity generation capacity forecasts could improve electricity load forecasting accuracy. In addition, from a corporate point of view accurate electricity load forecasting is important for the maximization of the profitability of energy generation companies and distribution companies. Lately, the expansion of smart meters in households provides the opportunity of scheduling their electricity consumption for economically benefit of low-price periods.

The vast amount of research on forecasting methodology and the modern computational and statistical power are making electricity load forecasting more feasible. In this thesis, a categorization and presentation of the most used forecasting models was made. Instead of trying to develop a more complicated and sophisticated forecasting model, we tried to examine how simpler models are capturing the behavior of electricity load and could provide a reliable short-term forecast. Temperature effects were examined with a linear and a non-linear specification. Regression models appeared to perform well based on overall goodness of fit while the test hypothesis of the parameters showed that there is a causal relationship between electricity load, average price and temperature. In addition, the effect of temperature from the comparison of the R-squared of the two models appears to be non-linear (models that have as explanatory variables the HDD and CDD variables exhibited better estimation results and better forecast accuracy metrics in the postestimation part of the sample). However, the different behavior of electricity load at specific hours of the day suggests that models that take into account those different characteristics should be preferred.

In this dissertation, the econometric models examined appeared to perform well in creating short-term predicted electricity load values. However, there is a limitation in adopting such methodology for predicting the future of variables, since it requires information that is not available at the time that the forecast is being made. On the other hand, time series models such as the SMA and Autoregressive models could be used for creating forecasts since they use only the previous information inherited in the time series process of electricity load. The forecast of electricity load using the 168 previous hourly observations (SMA(168) or 7-day hourly observation) presented a MAPE at around 9% which is considered as a good score. A similar performance with a MAPE of 7% we observed with the SARIMA(2,0,2,12) model arguing that autoregressive models could provide a reliable and fast solution for short term electricity load forecasting.

As modern technology and artificial intelligence are increasing the computational capabilities of statistical software and make their use easier even for no-experts in econometrics, models that are combining the causal relationship between variables and time series properties should be examined. Models that offer the ability to capture non-linear dynamic relationships among variables, the memory of time series processes and have the ability to learn such as neural networks and artificial neural networks could capture the mechanisms of energy markets and create accurate forecasts.

Advanced machine learning techniques give the opportunity to researchers to apply advanced learning algorithms such as deep and reinformed learning to improve forecasting efficiency of electricity load and energy models. The expansion of the smart grid that offers real time and high-resolution electricity load data gained from the smart meters and Internet of Things (IoT) devices connected in the electricity grid offers opportunities for the adoption of Big Data Analytics techniques in electricity load modeling, leveraging to an extend the forecast error. The vast number of models in academic literature as also those examined in this dissertation are using historical electricity load and other variables data to generate forecasts for future periods. Big Data Analytics and real time availability of electricity data are offering a promising area for future research in adaptive real-time electricity load forecasting for supporting dynamic decision making.

References

1 Adams, F. G., & Shachmurove, Y. (2008). Modeling and forecasting energy consumption in China: Implications for Chinese energy demand and imports in 2020. Energy Economics, 30(3), 1263–1278. https://doi.org/10.1016/j.eneco.2007.02.010

2 Bianco, V., Manca, O., & Nardini, S. (2009). Electricity consumption forecasting in Italy using linear regression models. Energy, 34(9), 1413–1421. https://doi.org/10.1016/j.energy.2009.06.034

3 Bianco, V., Marchitto, A., Scarpa, F., & Tagliafico, L. A. (2020). Forecasting energy consumption in the EU residential sector. International Journal of Environmental Research and Public Health, 17(7), 2259. https://doi.org/10.3390/ijerph17072259

4 Blázquez, L., Boogen, N., & Filippini, M. (2013). Residential electricity demand in Spain: New empirical evidence using aggregate data. Energy Economics, 36, 648–657. https://doi.org/10.1016/j.eneco.2012.11.010

5 Central Collection and publication of electricity generation, transportation and consumption data and information for the pan-European market. (no date) Data view. Available at: https://transparency.entsoe.eu/load-domain/r2/totalLoadR2/show (Accessed: 01 June 2023).

6 Chaparro-Peláez, J., Acquila-Natale, E., Hernández-García, Á., & Iglesias-Pradas, S. (2020). The digital transformation of the retail electricity market in Spain. Energies, 13(8), 2085. https://doi.org/10.3390/en13082085

7 Chitsaz, H., Shaker, H., Zareipour, H., Wood, D., & Amjady, N. (2015). Short-term electricity load forecasting of buildings in microgrids. Energy and Buildings, 99, 50–60. https://doi.org/10.1016/j.enbuild.2015.04.011

8 Ediger, V., & Akar, S. (2007). Arima forecasting of primary energy demand by fuel in Turkey. Energy Policy, 35(3), 1701–1708. https://doi.org/10.1016/j.enpol.2006.05.009

9 Ghalehkhondabi, I., Ardjmand, E., Weckman, G. R., & Young, W. A. (2016). An overview of energy demand forecasting methods published in 2005–2015. Energy Systems, 8(2), 411–447. https://doi.org/10.1007/s12667-016-0203-y

10 Halepoto, I, & Uqaili, M. & Chowdhry, B.. (2014). Least Square Regression Based Integrated Multi-Parameteric Demand Modeling for Short Term Load Forecasting. Mehran University Research Journal of Engineering & Technology. 33. 215-226.

11 Hausman, J. A. (1983). Chapter 7 specification and estimation of simultaneous equation models. Handbook of Econometrics, 391–448. https://doi.org/10.1016/s1573-4412(83)01011-9

12 Herraiz-Cañete, Á., Ribó-Pérez, D., Bastida-Molina, P., & Gómez-Navarro, T. (2022). Forecasting energy demand in isolated rural communities: A comparison between deterministic and stochastic approaches. Energy for Sustainable Development, 66, 101–116. https://doi.org/10.1016/j.esd.2021.11.007

13 Islam, M. A., Che, H. S., Hasanuzzaman, M., & Rahim, N. A. (2020). Energy demand forecasting. Energy for Sustainable Development, 105–123. https://doi.org/10.1016/b978-0-12-814645-3.00005-5

14 Kuster, C., Rezgui, Y., & Mourshed, M. (2017). Electrical load forecasting models: A critical systematic review. Sustainable Cities and Society, 35, 257–270. https://doi.org/10.1016/j.scs.2017.08.009

15 Lindberg, K. B., Seljom, P., Madsen, H., Fischer, D., & Korpås, M. (2019). Long-term electricity load forecasting: Current and future trends. Utilities Policy, 58, 102–119. https://doi.org/10.1016/j.jup.2019.04.001

16 Moral-Carcedo, J., & Vicéns-Otero, J. (2005). Modelling the non-linear response of Spanish electricity demand to temperature variations. Energy Economics, 27(3), 477–494. https://doi.org/10.1016/j.eneco.2005.01.003

17 Nti, I. K., Teimeh, M., Nyarko-Boateng, O., & Adekoya, A. F. (2020). Electricity Load Forecasting: A systematic review. Journal of Electrical Systems and Information Technology, 7(1). https://doi.org/10.1186/s43067-020-00021-8

18 Ramanathan, R., Engle, R., Granger, C. W. J., Vahid-Araghi, F., & Brace, C. (1997). Short-run forecasts of electricity loads and Peaks. International Journal of Forecasting, 13(2), 161–174. https://doi.org/10.1016/s0169-2070(97)00015-0

19 Sadler, T. R. (2022). Energy economics: Science, policy, and Economic Applications. Lexington Books.

20 Seitaridis, M.I., Thomaidis, N.S. and Biskas, P.N. (2021) ‘Fundamental responsiveness in European electricity prices’, Energies, 14(22), p. 7623. doi:10.3390/en14227623.

21 Thomaidis, N.S. and Biskas, P.N. (2021) ‘Fundamental pricing laws and long memory effects in the day-ahead power market’, Energy Economics, 100, p. 105211. doi:10.1016/j.eneco.2021.105211.

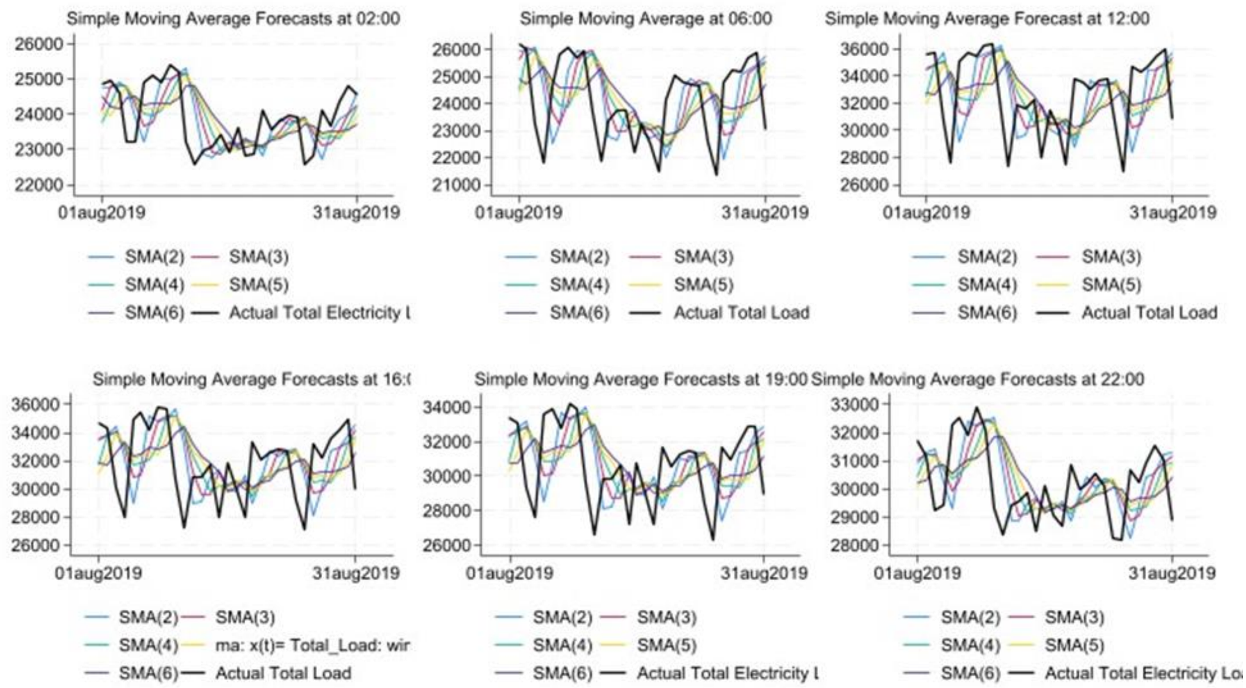
22 Thomaidis, N.S. and Dounias, G.D. (2010) ‘On detecting the optimal structure of a neural network under strong statistical features in errors’, Journal of Time Series Analysis, 32(3), pp. 204–222. doi:10.1111/j.1467-9892.2010.00693.x.

23 Yildiz, B., Bilbao, J. I., & Sproul, A. B. (2017). A review and analysis of regression and machine learning models on Commercial Building Electricity Load forecasting. Renewable and Sustainable Energy Reviews, 73, 1104–1122. https://doi.org/10.1016/j.rser.2017.02.023

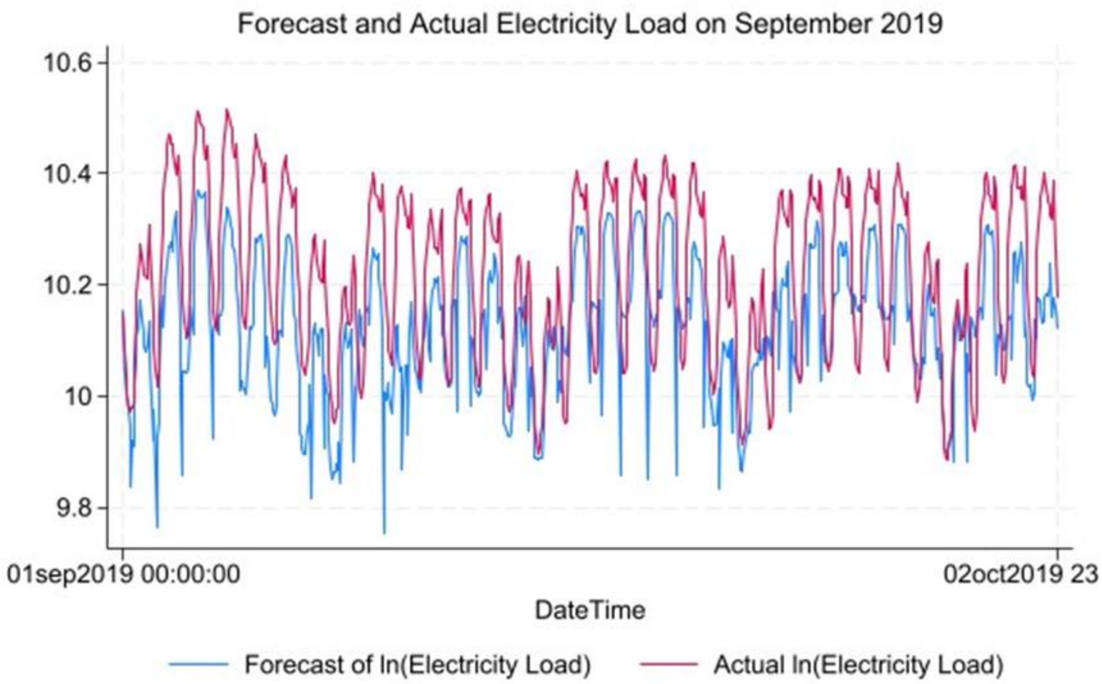
24 Zellner, A. (1962). Three-stage least squares: Simultaneous estimation of simultaneous equations. Netherlands School of Economics.

Appendix A: Graphs

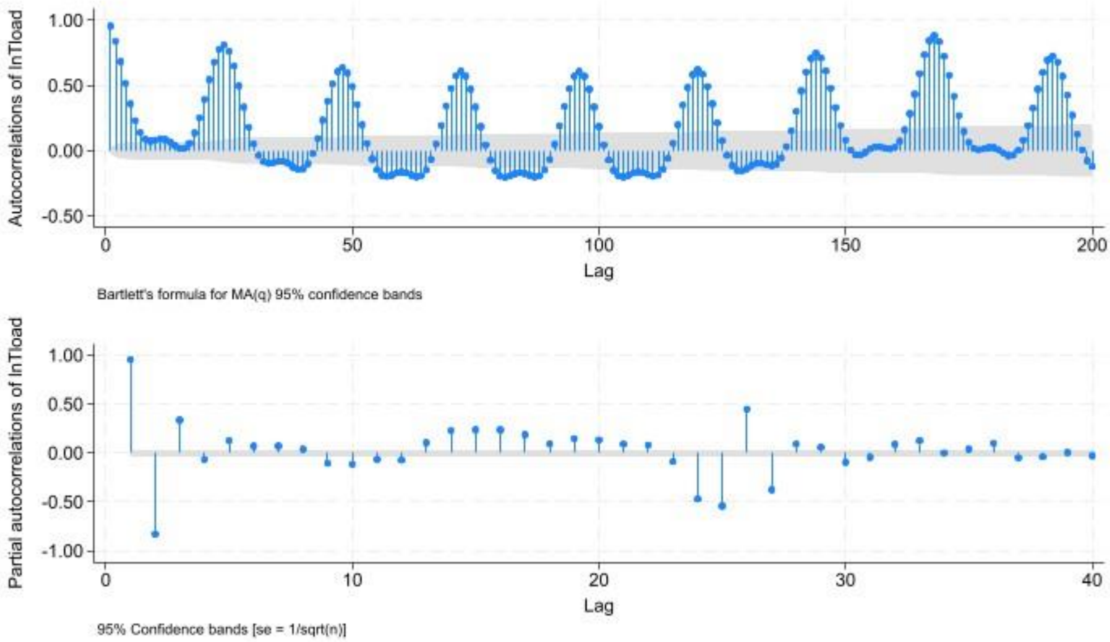
Graph A.1. SMA Forecast Plots at Specific Hour



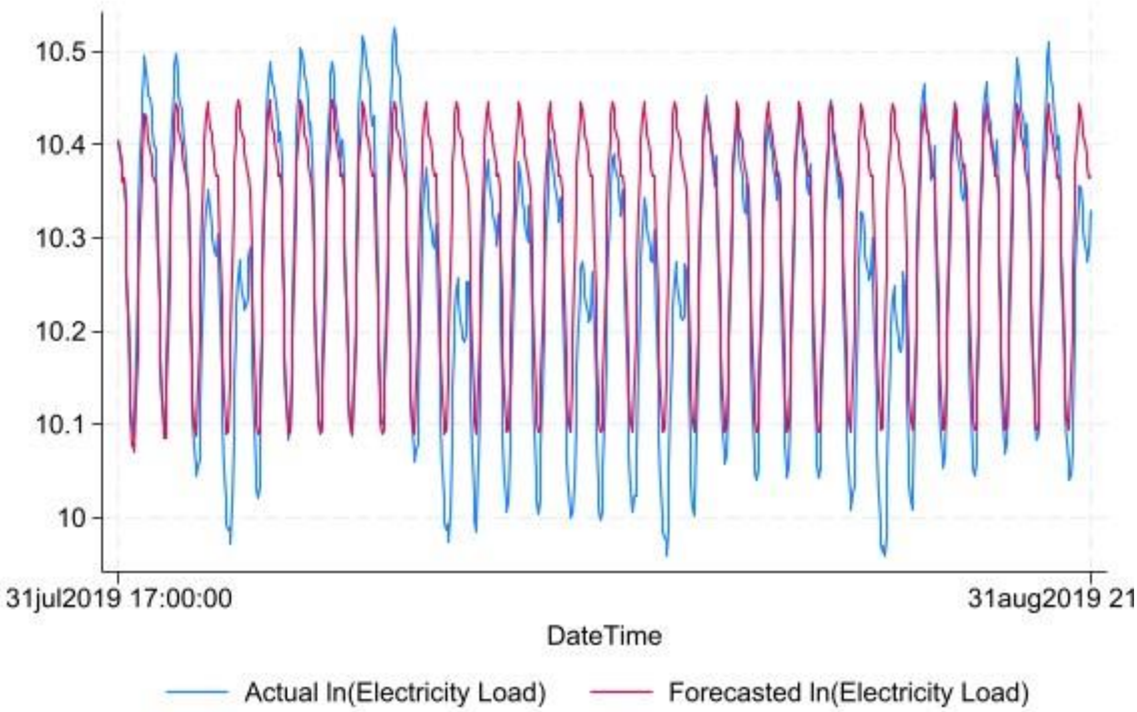
Graph A.2. Regression Forecasts for September 2019



Graph A.3. ACF and PACF for ln(Electricity Load)



Graph A.4. SARIMA(2,0,2,12) Electricity Load Forecast



Appendix B: Statistical Analysis Output

B1 Table: Regression Output from hourly observation sample until 31 August 2019

Source	SS	df	MS
Model	84.7769083	11	7.70699166
Residual	61.069859	5,795	.010538371
Total	145.846767	5,806	.025120008

Number of obs	=	5,807
F(11, 5795)	=	731.33
Prob > F	=	0.0000
R-squared	=	0.5813
Adj R-squared	=	0.5805
Root MSE	=	.10266

lnTload	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
lnDAPrice	.3752586	.0089916	41.73	0.000	.3576318	.3928855
lnaver24	-.1715216	.0141257	-12.14	0.000	-.1992132	-.1438301
lnhdd	-.0416294	.0018301	-22.75	0.000	-.0452171	-.0380416
lncdd	.088201	.0018169	48.54	0.000	.0846391	.0917629
d_tue	.0325524	.0051873	6.28	0.000	.0223834	.0427214
d_wed	.0371005	.0051269	7.24	0.000	.0270499	.0471511
d_thur	.0362713	.0050985	7.11	0.000	.0262764	.0462661
d_fri	.0275364	.0051133	5.39	0.000	.0175123	.0375605
d_sat	-.0556502	.005143	-10.82	0.000	-.0657325	-.0455679
d_sun	-.1152248	.0051796	-22.25	0.000	-.1253788	-.1050708
month	-.0298318	.0009624	-31.00	0.000	-.0317185	-.0279451
_cons	9.597166	.0479101	200.32	0.000	9.503244	9.691088

Author's Statement:

I hereby expressly declare that, according to the article 8 of Law 1559/1986, this dissertation is solely the product of my personal work, does not infringe any intellectual property, personality and personal data rights of third parties, does not contain works/contributions from third parties for which the permission of the authors/beneficiaries is required, is not the product of partial or total plagiarism, and that the sources used are limited to the literature references alone and meet the rules of scientific citations.