Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in
Statistica per l'Economia e l'Impresa

Relazione Finale

Forecast Combinations for Electricity Prices: The Case of the Italian Market

Relatore: Prof. Luigi Grossi
Dipartimento di Scienze Statistiche

Laureando: Luca Salvador
Matricola: 2001341

Anno Accademico 2022/2023

Contents

Introduction

The evolution of electricity markets in the last decades has been significant, moving from centralised monopolies to decentralised, competitive models. The evolution began in the 20th century with a single entity managing all phases of electricity supply, based on the belief that electricity was a natural monopoly best managed by a central entity. Starting in the 1970s, politicians began to push for the liberalisation of the market, reducing government interference and promoting competition. This movement brings potential benefits, such as increased innovation, improved products and lower prices, while promoting economic growth.

This process started in several countries in the 1990s, significantly influenced by the Chilean reform in the 1980s. The European Union embarked on this path in the mid-1990s, although member states had different speeds of implementation. These changes led to the entry of new competitors, increasing competition and further disrupting traditional market structures. In the 21st century, these trends continue, driven by technological innovation, policy reforms and market mechanisms, leading to a more diverse and competitive energy landscape.

Electricity prices are a linchpin in the market, influencing economic activities, thus making price forecasting a crucial aspect of investment planning and market strategy formulation. For market operators, accurate price forecasts are invaluable. In the short term, they inform buying or selling decisions, maximising profits. Medium to long-term forecasts guide future investment planning in new infrastructure and generation capacity, ensuring the stability of the electricity supply. Additionally, in the field of risk management, these forecasts are essential for mitigating volatility exposure, efficient resource allocation, and informed investment decision-making. However, single forecasts may occasionally be off-target, presenting potential financial risks for market participants. Herein lies the significance of forecast combinations or blending different models' forecasts. This approach provides more accurate predictions, reducing uncertainty in energy markets and risks associated with investment decisions. The quality of individual forecasts and the weight attributed to each in the combination are both extremely important for achieving an improvement in forecast combination. This technique, also known as model averaging, is designed to yield superior forecasts by leveraging the strengths of several individual forecasts, offsetting each model's errors and biases to generate a more accurate prediction. Forecast combination gained recognition with Bates and Granger's seminal 1969 work (Bates & Granger, 1969). They demonstrated that a blend of multiple forecasts could reduce mean-square error. This principle was further reinforced in the more recent M4 competition (Makridakis et al., 2020), a large-scale predictive analysis contest, where combinations showed impressive results.

In Chapter 1 the evolution and liberalisation of electricity markets is discussed. The shift began in the 1970s, driven by a desire for improved competition, efficiency, and consumer choice. This liberalisation involved the separation of monopolistic entities into distinct functions. The narrative continues by explaining the mechanisms of the marketplace, focusing on wholesale electricity trading, which encompasses day-ahead markets and intra-day markets, vital for balancing supply and demand in real-time.

In Chapter 2, the text introduces the literature on Forecasting Combinations and provides an overview of the methods used to evaluate the performance of forecasts. Moreover, the chapter explores five distinct methods for determining the weights for forecast combinations. These methods are crucial for optimizing the use of different forecasts in creating a combined, and typically more accurate, forecast.

In Chapter 3, various models used for time series forecasting are introduced. These models include ARIMA, Exponential Smoothing, Random Forest, and Spline. The chapter dives into the specifics of each model, illustrating their characteristics, benefits, and applications in time series forecasting.

In Chapter 4, an exploratory data analysis is conducted on 24 distinct time series. This analysis includes the application of Stationarity tests to these series. Moreover, forecasting plots are presented, juxtaposing actual data with predictions from selected forecasting models. Finally, the chapter concludes with a summary of the forecasting performance for each model, providing an informative comparison of their respective predictive capabilities.

Chapter 1

Electricity markets

The history of electricity markets has evolved significantly from a highly centralised and vertically integrated system to a decentralised and market-oriented model. At the beginning of the 20th century, the electricity sector was characterised by a monopolistic structure in which a single player managed the entire process of electricity generation, transmission and distribution to end consumers. This structure was based on the belief that electricity was a natural monopoly and that therefore one central entity was best placed to manage the entire system. In the 1970s, however, politicians began to question the effectiveness of this system and started to look for ways to introduce competition and liberalise the electricity sector. As the name suggests, liberalisation is a process of reducing or eliminating the government's presence in an industry to increase competition and create the conditions for a free market.

In general, a free market can be beneficial in that it allows competition between companies for the supply of goods and services, which can ultimately lead to higher quality products, lower prices and more innovation. Companies are then motivated to be efficient and innovative to attract customers and increase their market share, which can lead to increased productivity, economic growth and job creation. Consumers also benefit, as they have more choices and can make purchases that better suit their needs and budget, having the freedom to choose from a wider group of suppliers. Furthermore, the free market allows individuals and businesses to pursue their interests and allocate resources according to market demand, rather than relying on government intervention or central planning. This can lead to a more efficient allocation of resources and a better overall economic outcome.

On the other hand, in industries characterized by stringent regulations and monopolies, firms often find themselves bounded by inflexible rules that can impede innovation and limit their ability to adapt to market fluctuations. The process of deregulation and liberalisation provides corporations with the latitude to function more autonomously and base their decisions on market dynamics rather than governmental mandates. An additional benefit of deregulation is its capacity to boost competitiveness within a market. Indeed, in a regulated industry, the provision of specific products or services is typically restricted to a solitary entity or a limited number of companies, potentially escalating prices while compromising on quality. Deregulation simplifies the process for new entrants in the market, as it diminishes entry barriers and incites increased competition, which could ultimately translate to more affordable prices and superior quality goods and services for the end consumer. However, to make a market entry, certain prerequisites need to be fulfilled, for instance, an electricity utility is required to adhere to specific regulations and undertake its own set of obligations and responsibilities.

1.1 The Liberalisation Process

In the 1990s, several countries started to reorganise and liberalise their electricity markets, following in the footsteps of the pioneering Chilean reform of the 1980s. The United Kingdom and the Nordic countries of Europe were among the first to begin the transition, leading the way for many nations around the world to embrace deregulation as a means of creating competitive energy markets. In Australia, for example, the National Electricity Market (NEM) was established in 1998 to replace a previously state-run electricity market. New Zealand followed suit a year later with the introduction of the New Zealand Electricity Market (NZEM).

Nevertheless, the real impetus for deregulation was observed in Europe. The European Union (EU) began its path towards energy market liberalisation in the 1990s, starting with the publication of the first electricity directive in 1996. This was followed by a second electricity directive in 2003, which required all EU member states to have fully liberalised electricity markets by 2007. While some countries, such as Germany, liberalised their markets quickly, others, such as France and Italy, were slower to adopt these changes.

In Italy, the first steps were taken in 1992, when Enel was transformed into a joint-stock company, and then Italy liberalised its electricity market in February 1999 (Giulietti & Sicca, 1999), following the adoption of the European Union Electricity Directive in 1996. At the legal level though, the complete liberalisation of the entire Italian energy market was implemented with Decree Law 73 of 2007, later converted into Law 125 of 03/08/2007.

As already mentioned, the liberalisation process involved the separation of the integrated monopolies that previously managed the production, transmission and distribution of electricity and the introduction of competition through the creation of an electricity wholesale market. Formed under Law No. 481/95 in the year 1995, the Autorità per l'Energia Elettrica e il Gas (AEEG), or the Italian Regulatory Authority for Electricity and Gas, was instituted with the purpose of overseeing and regulating public utility services related to electricity and gas. The AEEG's task was to promote competition and efficiency in the sectors, to ensure fair remuneration for companies and to guarantee service through a transparent and geographically uniform tariff system. In particular, the AEEG played a significant role in the liberalisation of the Italian electricity market, operating from 1996 to 1998. Over time, the AEEG extended its regulatory scope to other aspects of the energy sector, eventually taking on the name ARERA. Today, ARERA continues to regulate the Italian energy market and protect the interests of users and consumers.

The liberalisation of the electricity market in Italy has led to the entry of new competitors and the emergence of new players such as wholesalers, producers, distributors, traders and consortia. Intermediaries and service providers have played an increasingly important role in the liberalised market and new specialised entities have emerged to provide such services. Enel S.p.A., a major electricity company, was obliged to divest 15,000 MW of capacity by 2003, according to Legislative Decree No. 79/99. The government set the guidelines for divestment, including the requirement that new generators have a mix of base and peaking plants, diversified primary energy sources, and adequate geographic distribution to avoid local monopolies. The divestment process involved the creation of three joint-stock companies by Enel S.p.A. and their subsequent sale to new operators. (ARERA Annual Report "The liberalisation of the market" 2000)

1.2 The marketplace

Since the 1990s, the deregulation and introduction of competitive markets in many countries have transformed traditionally monopolistic energy sectors under state control. Electricity is now traded through spot and derivative contracts, but cannot be economically stored in a cheap way and therefore requires a constant balance between production and consumption. Unlike other goods, electricity is difficult to store as many technologies for the storage of energy on a large scale are still very limited. As a result, electricity must be produced at the same moment it is consumed and the price of electricity varies throughout the day, depending on the balance between supply and demand. Excess supply and low demand lead to low prices, while high demand and low supply drive prices up. The power system must be kept in balance; electricity must be produced and consumed in equal amounts at all times.

This balance is granted on the market via trading electricity in one-hour cycles. In the electricity market, there are different trading products with different periods between purchase and actual delivery. On the forward market, electricity can be traded several years in advance, while long-term contracts are used by buyers to cushion the risk of price increases. By paying a premium for this planning certainty, buyers generate additional revenue for sellers, which can be used to finance new generation capacity. The closer the delivery date, the more accurate the consumption and production forecasts become, and the short-term spot market divides into two markets with different delivery times: a day-ahead market and an intra-day market. One of the largest markets for electricity exchange is the European internal energy market exchange and the majority of electricity trading takes place on day-ahead markets.

Day-ahead markets, a crucial component of the wholesale electricity markets, function as a forward market where electricity quantities and prices are determined a day in advance based on forecasted demand. Participants—generators, retailers, and large consumers—submit their bids and offers for each hour of the upcoming day, indicating how much electricity they plan to produce, consume, or sell and at what price. These bids and offers are then processed by the market operator using complex optimization algorithms to match supply and demand for every hour, resulting in a market clearing price, also known as the locational marginal price. This market price represents the marginal cost of electricity supply for each hour and serves as the reference price for all market transactions for that day. This mechanism ensures that the generation resources are scheduled economically, prioritizing the lowest-cost electricity first, while maintaining system reliability. The day-ahead market provides market participants with a degree of financial certainty and allows for efficient resource planning, thus playing a pivotal role in the functioning of modern electricity markets.

In most electricity exchanges, day-ahead markets are complemented by additional types of markets that allow the participants to trade electricity in real-time or in advance. These markets include intra-day markets, which allow participants to adjust their positions during the day, and futures markets, which allow participants to trade electricity for delivery in the future. Together, these markets offer market participants a range of options to manage their exposure to price and supply risks in the electricity market. In order to enable the market to be balanced at every time of the day, an important mechanism plays a role. To take on uncertainties derived from external factors such as weather conditions, the intra-day market allows indeed for operations once the day-ahead market closes. Trades can take place immediately when purchases and sales bids meet. All intra-day market activities need to be completed at least 30 minutes before the start of the electricity delivery hour. Participants in the day-ahead market trade electricity for the next day and submit bids before noon specifying the quantity and time of delivery. The exchange then determines the wholesale price for each hour of the following day and accepts the successful bids. This wholesale price is an important reference for the electricity market, similar to the closing price of a share on the stock exchange. On the intraday market, on the other hand, electricity can be traded up to 30 minutes before delivery, although within the regulation zone, it is only traded five minutes in advance. In the electricity market, there are different trading products with different time periods between purchase and actual delivery. On the forward market, electricity can be traded several years in advance, while long-term contracts are used by buyers to cushion the risk of price increases. By paying a premium for this planning certainty, buyers generate additional revenue for sellers, which can be used to finance new generation capacity. The closer the delivery date, the more accurate the consumption and production forecasts become, and the short-term spot market divides into two markets with different delivery times: a day-ahead market and an intraday market.

In the electricity sector, there is a wholesale market in which electricity producers compete to offer their electricity production to retailers who then reprice and sell it to consumers. In general, wholesale prices have been limited to large retail suppliers. However, there is a recent trend in markets such as New England to open up to end-users seeking to reduce unnecessary overheads in their energy costs by purchasing directly from generators. This shift by consumers towards buying electricity directly from generators is a relatively new phenomenon, as in the past it was believed that large retail suppliers were better equipped to handle the complexities of electricity markets. However, as end-users seek greater control over their energy costs, purchasing directly from generators is becoming increasingly attractive. The advantages of buying electricity directly from generators include lower prices and greater control over the energy sources used. By eliminating middlemen, consumers can save money and ensure that their electricity is generated from specific sources, such as renewable energy. In overall terms, the shift by end users to purchasing electricity directly from generators is part of a broader trend where consumers are seeking more control and transparency in their energy consumption and as markets continue to open up and technology advances, this trend is likely to continue.

The wholesale market is where commodities (electricity) are bought and sold and the trading process includes electricity generators, electricity suppliers (whose job it is to trade electricity for sale to end consumers), and traders, who carry out the daily transactions that contribute to the liquidity of the market. Competition is high in wholesale markets, where electricity is traded before it is sold to consumers. Electricity companies compete to sell their electricity at the lowest possible price but also take into account the reliability and stability of the grid. These markets typically operate in two phases: a day-ahead market, where transactions are made based on demand forecasts, and a real-time market, which responds to immediate demand and compensates for deviations from the day-ahead forecast. These markets, with their sophisticated auction mechanisms and pricing algorithms, ensure an efficient and economic allocation of electricity, always favouring the cheapest resources to meet demand, just like in any other efficient financial market. With the advent of renewable energies, these markets have become even more complex, as the fluctuations of solar and wind energy require innovative strategies to balance the grid. However, the further development of wholesale electricity markets plays a key role in maintaining grid reliability, promoting competition and facilitating the transition to a more sustainable energy future.

As we navigate the third decade of the 21st century, global electricity markets present a picture of dynamic transformation, marked by a confluence of technological innovation, policy reforms and market mechanisms. Traditional boundaries have been blurred by the proliferation of decentralised power generation, giving rise to a more diverse and competitive energy landscape.

Chapter 2

Forecast Combinations

As described in Chapter 1 electricity prices are an important component of the market, due to the crucial role they play in economic dynamics, so price forecasting is an indispensable component for companies and institutions when considering the macroeconomic scenario and planning future investments and strategies in the markets.

For market operators, price forecasts help in planning electricity supply and choosing trading strategies. For example, short-term price forecasts allow traders to decide whether to buy or sell electricity in the market to maximize profits, while medium and long-term price forecasts help the operators to plan future investments in new generation capacity and infrastructure, ensuring a long-term stable and reliable supply of electricity. Moreover, electricity price forecast plays a determining role while looking at the risk management side of market participants; indeed, it is extremely useful to reduce exposure to volatility, allocate resources efficiently and make informed investment decisions.

Single forecasts in the context of electricity prices can sometimes be inaccurate, leading to potential financial losses and risks for market participants. Therefore combining different forecasts coming from different models can result helpful as it can allow us to obtain a better and more accurate forecast.

In the context of electricity prices, the forecast combination can be useful, for instance, to mitigate uncertainty in energy markets and to reduce the risk associated with investment decisions. In addition, one model could outperform others while looking at a specific hourly timeframe but not for others. It is important to consider that to achieve a substantial improvement in forecast combination, it is not only the quality of the single forecasts that is important but also the estimation of the weights attributed to each of them.

Forecast combination, also known as model averaging, is a method used in statistical forecasting to generate improved forecasts by combining more individual forecasts. The underlying idea is that by combining forecasts from different models, the errors and biases inherent in each individual model can be reduced in order to obtain a more accurate forecast. One important assumption while considering the implementation of combining forecasts is that models taken into account are generally accurate enough. Indeed, if we are using specific models that are not suited for the data considered and thus inaccurate in forecasting, this may as well be leading to worse forecasts than just a single accurate model.

The work conducted by Bates and Granger in 1969 (Bates & Granger, 1969) is considered one of the seminal works in the field of forecast combinations. In this work, they explore the potential benefits of combining several forecasts to improve forecast accuracy. Considering a particular example in their work, they report that *" ... the composite set of forecasts can yield lower mean-square error than either of the original forecasts."*. In this paper, one of the various aspects analysed is that the variance of a combination of forecasts as simple as the arithmetic mean of two forecasts has a smaller variance than both forecasts taken alone.

Another major early contribution to this field was made by Clemen (Clemen, 1989) where he also concluded that *"forecast accuracy can be substantially improved through the combination of multiple individual forecasts."*. Moreover, an important finding of Clemen is that simple combinations perform relatively well compared to other more complex combinations and in line with this, more recent research found that during the M4 competition Makridakis et al. (2020) the simple combinations achieve good performances and are sufficiently competitive. The M4 Competition is a large-scale competition in predictive analysis in which the final goal is to evaluate the precision of diverse forecasting methodologies and to identify the most effective ones for different kinds of temporal data sequences. This contest incorporated 100,000 time series data from a variety of fields, such as macroeconomic factors, financial sequences, and population data, among others. Contestants had the freedom to employ any forecasting technique they preferred, with the precision of their predictions assessed through a comprehensive set of metrics. The outcomes of this competition have been extensively utilized to inform and shape both the practical application and academic study of forecasting.

One of the most significant aspects that emerged from the competition was the superiority of combined methods, especially statistical models, based on two main evaluation indicators: mean absolute symmetrical percentage error (sMAPE) and mean absolute scalar error (MASE). These error metrics serve as benchmarks to assess the accuracy of the various forecasting models. An interesting feature is that these two metrics were not used independently during the competition. Rather, the overall weighted average (OWA) of both measures was used to evaluate the forecasts submitted by the participants. This method facilitated a more holistic assessment, taking into account both the proportionality of the error (as noted by sMAPE) and the comparative effectiveness of the approach with respect to baseline forecasts (as noted by MASE).

The following information is derived from a table (see Table 2.1) presented in the work of Makridakis et al. (2020) showing the effectiveness of the combinations, which only by using simple arithmetic mean for all predictions of the M4 competition had lower values of sMAPE and MASE than Single, Holt and Damped Exponential Smoothing.

Table 2.1: Accuracy of the Single, Holt and Damped exponential smoothing PFs, as well as those of Comb and ARIMA, across the M4 dataset (Makridakis et al., 2020)

Method	sMAPE	MASE	OWA	Rank
SES	13.087	1.885	0.975	34
Holt	13.775	1.772	0.971	33
Damped	12.661	1.683	0.907	22
Comb	12.555	1.663	0.898	19
ARIMA	12.669	1.666	0.903	20

An additional significant finding in the competition was that a top-performing combination was one that combined one ML model and seven statistical models. The weights for this combination were given on the basis of a ML model that was trained to minimise forecasting errors.

2.1 Forecasting Performance Measures

Hereafter are presented the statistics that have been considered for the evaluation of the fitted models and the combination of forecasts.

Mean squared error (MSE)

$$\text{MSE} = \frac{1}{m}\sum_{t=1}^{m}\left(P_t - \hat{P}_t\right)^2 \tag{2.1}$$

Root mean squared percentage error (RMSPE)

$$\text{RMSPE} = \sqrt{\frac{1}{m}\sum_{t=1}^{m}\left(100 \times \frac{P_t - \hat{P}_t}{P_t}\right)^2} \tag{2.2}$$

Mean absolute percentage error (MAPE)

$$\text{MAPE} = \frac{1}{m}\sum_{t=1}^{m}\left|100 \times \frac{P_{jt} - \hat{P}_{jt}}{P_{jt}}\right| \tag{2.3}$$

Mean absolute scaled error (MASE)

$$\text{MASE} = \frac{m-1}{m}\frac{\sum_{t=2}^{m}\left|P_t - \hat{P}t\right|}{\sum_{t=2}^{m}|P_t - P_{t-1}|} \tag{2.4}$$

Where:

- m**:** This is the total number of observations in the dataset.
- P_t**:** This represents the actual observed value at time t.
- $\hat{P}_t$**:** Represents the predicted value at time t.

In this thesis, forecasting techniques are implemented and final observations are reserved for comparison with predicted values (out-of-sample forecasting), using metrics in equations (2.1-2.4)

2.2 How to choose the optimal weights?

A critical aspect of model combining techniques is the choice of the optimal weights and an assumption that can be made is that the performance of one individual forecast model would be consistent over time.

Again, as pointed out in Wang et al. (2022), the straightforward strategy of taking the arithmetic mean of different forecasts is not only a commonly employed method but also showcases impressive stability as a combination approach. This method, in its simplicity, treats each model equally, irrespective of its individual forecasting accuracy. However, a more reasonable approach would be to assign greater weight to the model exhibiting superior forecast accuracy. It stands to reason that a model demonstrating a higher degree of precision should have a more significant influence on the combined forecast, thereby potentially enhancing the overall predictive performance.

Bates and Granger (Bates & Granger, 1969) express the variance of the combination as follows:

$$\sigma_c^2 = k^2\sigma_1^2 + (1-k)^2\sigma_2^2 + 2\rho k\sigma_1(1-k)\sigma_2 \tag{2.5}$$

where k is the proportionate weight given to the first set of forecasts and ρ is the correlation coefficient between the errors in the first set of forecasts and the second. The final goal is to minimize the variance σ_c^2 and by differentiating with respect to k we obtain:

$$k = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 - \sigma_2^2 - 2\rho\sigma_1\sigma_2} \tag{2.6}$$

Bates and Granger show that if k is determined by equation (2.5), the value of σ_c^2 is no greater than the smaller of the two individual variances. This process is implemented in the ForecastComb R package with the function comb_BG created by authors: Christoph E. Weiss and Gernot R. Roetzer. (Weiss et al., 2018)

2.3 ForecastComb Packages for Choosing Weights

In this section, we are going to illustrate the functions used in the analysis conducted in this thesis to find optimal weights. Functions are part of the R Statistical Software Package. Weiss et al. (2018)

2.3.1 comb_CSR

The function comb_CSR, represents the Complete Subset Regression method utilized for combining forecasts. The function iteratively generates all possible subsets of the ensemble of base-forecasters (i.e., all possible combinations of the columns in the prediction matrix). For each subset, it estimates a linear regression model with the observed vector as the response variable and the selected base-forecasters as the predictors. This is done using the lm function, which performs Ordinary Least Squares (OLS) regression.

For each estimated model, it calculates four Information Criteria (AIC, AICc, BIC, HQ). These criteria are then normalized using the *comp_normalized_weights* function.

The expression for the weights is:

$$w_i = \frac{e^{-0.5*IC_i}}{\sum_{j=1}^{n} e^{-0.5*IC_j}} \tag{2.7}$$

Where:

- w_i is the weight for model i.
- IC_i is the information criterion (e.g., AIC, BIC, etc.) for model i.

2.3.2 comb_LAD

The `comb_LAD()` function is employed to combine forecasts using the Least Absolute Deviation (LAD) approach. Unlike the Ordinary Least Squares method which minimizes the sum of squared differences between observed and predicted values, the LAD method minimizes the sum of absolute differences. This characteristic makes the LAD approach more robust against outliers, as it is less influenced by extreme values compared to the OLS method. Inside the function, linear programming techniques are utilized to solve the optimization problem that seeks to find the optimal weights by minimizing the sum of absolute forecast errors.

The problem for LAD can be expressed as follows:

$$\min_{\beta} \sum_{i=1}^{n} |y_i - x_i'\beta| \tag{2.8}$$

Where:

- y_i is the i-th observed value
- x_i' is the i-th row of the matrix of predictors (including a 1 for the intercept term)
- β is the vector of coefficients (weights in the forecast combination context)

The estimation of the regression coefficients (or weights, $\boldsymbol{\beta}$) in LAD regression is a linear programming problem and it requires numerical methods to solve for the coefficients.

2.3.3 comb_NG

The comb_NG (Newbold & Granger, 1974) function is an implementation of the Newbold and Granger (1974) method for combining forecasts. This method calculates optimal weights for each model's forecast based on the inverse of the sample mean squared prediction error matrix. First, the function computes the error matrix by subtracting the prediction matrix from the observed vector. Then, it calculates the sample mean squared prediction error matrix, which is used to compute the optimal weights for combining forecasts. The weights are calculated by solving the linear system involving the inverse of the sample mean squared prediction error matrix and a vector of ones.

Let y_t be the target variable and $\mathbf{f}_t = (f_{1t}, \ldots, f_{Nt})'$ a set of N predictors, with Σ being the mean squared prediction error matrix of $\mathbf{f}_t$ and $\mathbf{e}$ an $N \times 1$ vector of $(1, \ldots, 1)'$. Their method minimizes the mean squared prediction error, subject to the normalization condition $\mathbf{e}'\mathbf{w} = 1$.

The combination weights derived are:

$$\mathbf{w}^{NG} = \frac{\Sigma^{-1}\mathbf{e}}{\mathbf{e}'\Sigma^{-1}\mathbf{e}} \tag{2.9}$$

The combined forecast is then computed using:

$$\hat{y}_t = \mathbf{f}_t'\mathbf{w}^{NG} \tag{2.10}$$

2.3.4 comb_OLS

In the ForecastComb package, the `comb_OLS()` function is used to perform the Ordinary Least Squares (OLS) combination of forecasts. This method aims to find the best linear fit by minimizing the sum of the squared differences between the observed values and the predicted values. The `comb_OLS()` function assigns weights to each individual forecast based on the OLS method, therefore, the weights are equal to the coefficients estimated by the OLS method. In doing so, the function offers a systematic and statistically sound method for capitalizing on the strengths of multiple forecasts to generate a more accurate combined forecast. Internally, the function employs linear regression techniques to determine the optimal weights that minimize the sum of the squared forecast errors.

This technique estimates the weights, denoted as $\mathbf{w}^{OLS} = (w_1, \ldots, w_N)'$, along with an intercept, b, for consolidating the forecasts.

Given N not perfectly collinear predictors, represented as $\mathbf{f}_t = (f_{1t}, \ldots, f_{Nt})'$, the combination of forecasts for a single data point can be formulated as follows:

$$y_t = b + \sum_{i=1}^{N} w_i f_{it}$$

2.3.5 comb_SA

The comb_SA function utilizes the Simple Average (SA) combination method. In this approach, all forecasting models are given equal weight, and their average is used as the final forecast.

Consider y_t as the target variable, with N distinct, non-collinear predictors denoted by $\mathbf{f}_t = (f_{1t}, \ldots, f_{Nt})'$. A simple averaging approach assigns equal weights to all predictors:

$$\mathbf{w}^{SA} = \frac{1}{N} \tag{2.11}$$

The combined forecast using this strategy is thus given by:

$$\hat{y}_t = \mathbf{f}_t'\mathbf{w}^{SA} \tag{2.12}$$

Chapter 3

Forecasting Models

3.1 ARIMA processes

Autoregressive Integrated Moving Average (ARIMA) processes are commonly estimated in time series forecasting. These models take into account three aspects of the time series: autoregression (AR), differencing (I), and moving average (MA).

An ARIMA model is typically denoted as ARIMA(p, d, q) where:

- p is the order of the Autoregressive part
- d is the degree of first differencing involved
- q is the order of the Moving average part

The general form of the ARIMA model is expressed as follows:

$$(1 - \sum_{i=1}^{p} \phi_i L^i)(1 - L)^d y_t = (1 + \sum_{i=1}^{q} \theta_i L^i)\epsilon_t \tag{3.1}$$

Where:

- y_t is the time series
- ϕ_i are the parameters of the autoregressive part of the model
- d is the order of differencing

- L is the lag operator
- θ_i are the parameters of the moving average part of the model
- ϵ_t is white noise

The ARIMA model is fitted to the time series data for predicting future points in the series.

3.1.1 ARIMA(7,1,0) Model

The ARIMA(7,1,0) model is a particular case of the general ARIMA model. In this model:

- The order of the autoregressive part (p) is 7. This means that the value at a given time t is predicted as a linear combination of the past 7 values.
- The order of differencing (d) is 1. This means that we don't model the time series itself, but the difference between consecutive values of the time series. This is typically used to make the time series stationary.
- The order of the moving average part (q) is 0. This means that the error at a given time is not a linear combination of past errors.

The mathematical representation of the ARIMA(7,1,0) model is:

$$(1 - \sum_{i=1}^{7} \phi_i L^i)(1 - L)y_t = \epsilon_t \tag{3.2}$$

Where:

- y_t is the time series
- ϕ_i are the parameters of the autoregressive part of the model
- L is the lag operator
- ϵ_t is white noise

The ARIMA(7,1,0) model is particularly useful for time series data where the value at a given time is influenced by the values at the previous 7-time points and where differencing is needed to make the time series stationary.

3.1.2 ARIMA(7,0,0) Model

The ARIMA(7,0,0) model is another specific case of the general ARIMA model. In this model:

- The order of the autoregressive part (p) is 7. This implies that the value at a given time t is predicted as a linear combination of the past 7 values.
- The order of differencing (d) is 0. This signifies that we model the time series itself, as opposed to the difference between consecutive values of the time series. This model is appropriate for time series that are already stationary.
- The order of the moving average part (q) is 0. This indicates that the error at a given time is not a linear combination of past errors.

The mathematical representation of the ARIMA(7,0,0) model is:

$$(1 - \sum_{i=1}^{7} \phi_i L^i) y_t = \epsilon_t \tag{3.3}$$

Where:

- y_t is the time series
- ϕ_i are the parameters of the autoregressive part of the model
- L is the lag operator
- ϵ_t is white noise

The ARIMA(7,0,0) model is especially useful for time series data where the value at a given time is significantly influenced by the values at the previous 7-time points and where the series is already stationary, thus not requiring differencing. This model primarily relies on the autoregressive component, capturing the dependencies among the previous observations and disregarding the moving average part, which accounts for the dependencies among the prediction errors.

3.2 Exponential Smoothing

Exponential Smoothing Models are a suite of forecasting methods which use weighted averages of past observations to forecast future data. The weights decay exponentially as the observations get older, hence the name "Exponential Smoothing".

The simplest form of Exponential Smoothing is Single Exponential Smoothing, represented as follows:

$$\hat{y}_{T+1|T} = \sum_{j=0}^{T-1} \alpha(1-\alpha)^j y_{T-j} + (1-\alpha)^T \ell_0 \tag{3.4}$$

Where:

- $\hat{y}_{T+1|T}$ is the smoothed value at time $T+1$
- y_{T-j} is the smoothed value at time $T-j$
- α is the smoothing parameter, $0 \leq \alpha \leq 1$

Which can also be written in another form:

$$\begin{aligned} \text{Forecast equation} \quad & \hat{y}_{t+h|t} = \ell_t \\ \text{Smoothing equation} \quad & \ell_t = \alpha y_t + (1-\alpha)\ell_{t-1} \end{aligned}$$

Exponential smoothing models are particularly useful for time series data where there is a need to give more weight to more recent observations in predicting the future.

3.2.1 Innovation State Space Models

The innovation state space models are described in Hyndman & Athanasopoulos (2018) and are an extension of exponential smoothing models. ETS models provide a general and flexible framework as they extend the exponential smoothing technique to capture three components of time series data: Error, Trend, and Seasonality, hence the abbreviation ETS.

Error (E): Residual (or the difference) between the actual and predicted value. ETS models can handle both additive and multiplicative error structures.

Trend: This captures any consistent upward or downward movement in the data over time. ETS models can handle both additive and multiplicative trends. They can also handle situations where there is no trend.
Seasonality: This captures any repeating patterns or cycles in the data that occur at regular intervals. ETS models can handle both additive and multiplicative seasonality, as well as situations where there is no seasonality.

3.3 Random Forest

Originating from the work of Leo Breiman in 2001 (Breiman, 2001), the Random Forest model has established its place as a key tool in the realm of machine learning. This model excels in tackling tasks such as regression and classification, and exhibits strength in handling issues like dimensionality reduction, outliers, missing values, and deriving feature importance.

In the context of Random Forests, the ensemble is made up of decision trees. A "forest" of decision trees is constructed, each with slight differences, and the final prediction is determined by aggregating the predictions of all the trees - for classification, this is often the mode of the classes, and for regression, it is typically the average of predictions (Breiman, 2001).

Two layers of randomness are employed in the creation of a random forest:

- Each tree is grown on a sample drawn with replacement (a bootstrap sample) from the original dataset, a process known as Bagging.
- When each tree is split, a subset of predictors is selected randomly from all predictors. The best split is found amongst these. A common choice for the number of predictors is $m = \sqrt{p}$, where p is the total number of predictors.

The Random Forest model can be represented mathematically as:

$$RF(y) = \frac{1}{B} \sum_{b=1}^{B} T_b(y) \tag{3.5}$$

where:

- $RF(y)$ denotes the prediction of the Random Forest for an input vector y.
- B represents the number of trees in the forest.
- $T_b(y)$ is the prediction made by the b-th decision tree for the input vector y.

In the code implemented the model looks like this where the train data is a data-frame where the explicatives, which are 7, are the original time series lagged 7 times:

```
model <- randomForest(PUN ~ ., data = train, ntree = 1000)
```

'PUN' is the dependent variable to be predicted, ' .' implies all other variables in the 'train' data frame are the predictors, and 'ntree = 1000' indicates that the model will be built using 1,000 trees.

Upon running this code, the Random Forest model is trained by fitting 1,000 decision trees to various bootstrap samples of the training dataset. To predict a new observation, each tree provides its own prediction. If the task is regression, the final prediction of the model is computed as the average of the 1,000 individual tree predictions. For classification, it's determined by majority voting.

The decision to utilize lagged features in conjunction with the Random Forest model is rooted in the nature of the forecasting problem. By working with time-series data, where temporal dependencies are key - current values often depend on previous values in the series. Random Forests, as they are, do not inherently consider the sequential nature of the data, treating each input independently. To account for the temporal dependencies, lagged features are inserted as the independent variables: features that incorporate information from previous time steps. This is done by creating new features that are the 'lagged' versions of existing ones, effectively shifting the series back by a certain number of steps In this context, seven lagged times series of the 'PUN' feature have been created, meaning that information from the preceding week is being incorporated into the model. With these temporal dependencies in the features, the Random Forest model can make more accurate predictions, as it now can model the effects of previous time steps on the current prediction.

3.4 Spline Modeling

Splines are a statistical tool often used for smoothing, interpolation, and curve fitting. A spline represents a defined polynomial function which is typically designed to be smooth and continuous over the range of the data. The key advantage of using splines is their flexibility in capturing complex non-linear patterns.

In the context of time series analysis, one common challenge is the presence of underlying trends in the data. These trends can often be non-linear and can complicate the task of forecasting. This is where splines can provide significant benefits. By fitting a spline to the time series data, we are effectively estimating the underlying trend in a flexible, non-linear manner.

In this specific scenario, a spline is initially fitted to each of the 24 time series within the dataset. This fitting process is designed to capture the underlying non-linear trend in each series. Subsequently, the fitted spline values are subtracted from the original time series data. This action, often referred to as "de-trending", produces residual series that are hopefully simpler and more suited to further modeling.

In the next step, an ARIMA(7,0,0) model is used to forecast the de-trended series. The ARIMA model is chosen for its effectiveness in capturing autocorrelations in time series data. Following the ARIMA modeling, we then add back the spline values to the forecasts. This re-integrates the original non-linear trend into our forecasts, resulting in final predictions that are consistent with the original series in terms of their underlying trends. This combined approach of spline de-trending and ARIMA modeling effectively handles both the non-linear trend and the autocorrelations in the time series data, resulting in reliable forecasts.

Chapter 4

Case Study. Italian Electricity Prices

This chapter presents an analysis of real data taken from the official website of the Italian electricity market by `https://www.mercatoelettrico.org/en/` . The loaded dataset contains observations from 1 January 2021 to 30 April 2023. However, for the purpose of analysing and validating the model, a subset of the original financial time series was created, consisting of observations from 1 January 2021 to 31 January 2023. This subset was used to fit various models and evaluate their performance. However, the models are, in a rolling window forecasting framework, rescaled at each iteration to forecast at time t+1. So in the case of both the ARIMA models and the ETS class models, what has been done, is to evaluate the model orders on the series until 31 January 2023, and then by keeping the constant orders, at each step forward, the model was being refitted.

The objective of the analysis is to show that by combining forecasts of the Unic National Prices (PUN), the forecasting performance measures used to compare the goodness of models, will indicate that combinations perform generally better than single models.

The last three months of data will be used to evaluate how well the models perform in predicting the PUN values.

The analysis is conducted on the prices for the 24 hours, so, for each day, we will have 24 forecasts.

4.1 Exploratory Data Analysis

4.1.1 Time Series Charts

The following charts illustrate the hourly PUN values for the period under consideration. These charts have been created to provide a visual representation of the data, making it easier to identify any patterns, trends, or anomalies that may be present in the data. The charts have been created using the ggplot2 package in R.

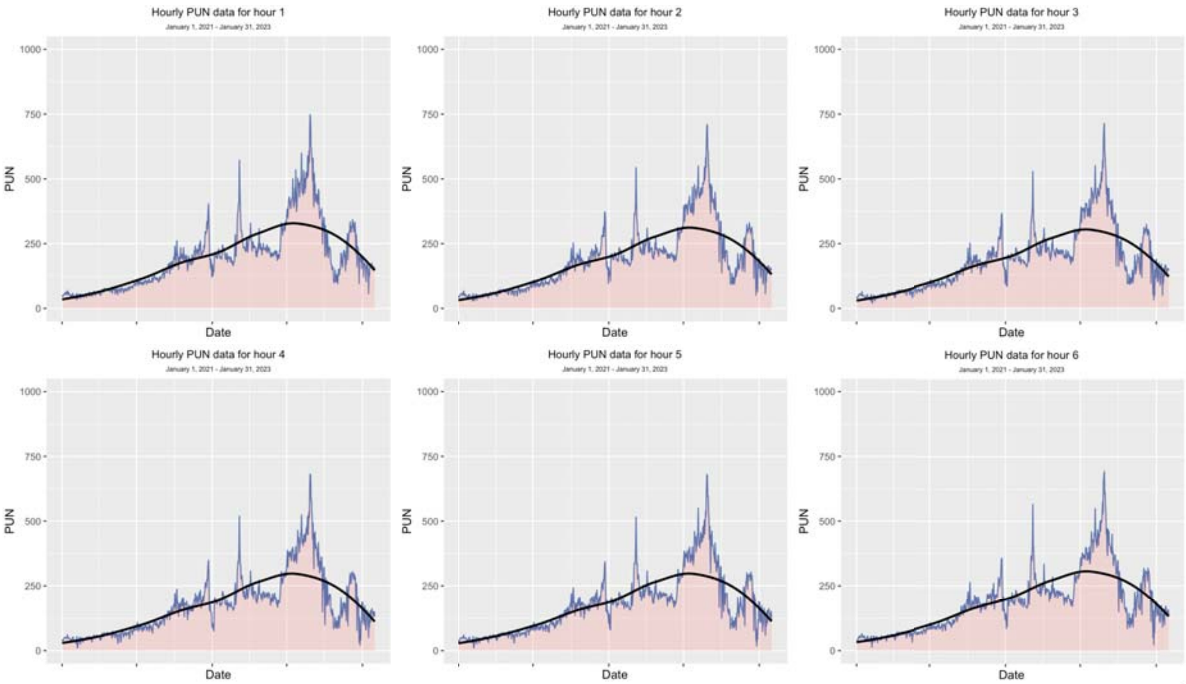

FIGURE 4.1: Hourly prices from 1 to 6. The continuous line is the non-linear estimated by splines.

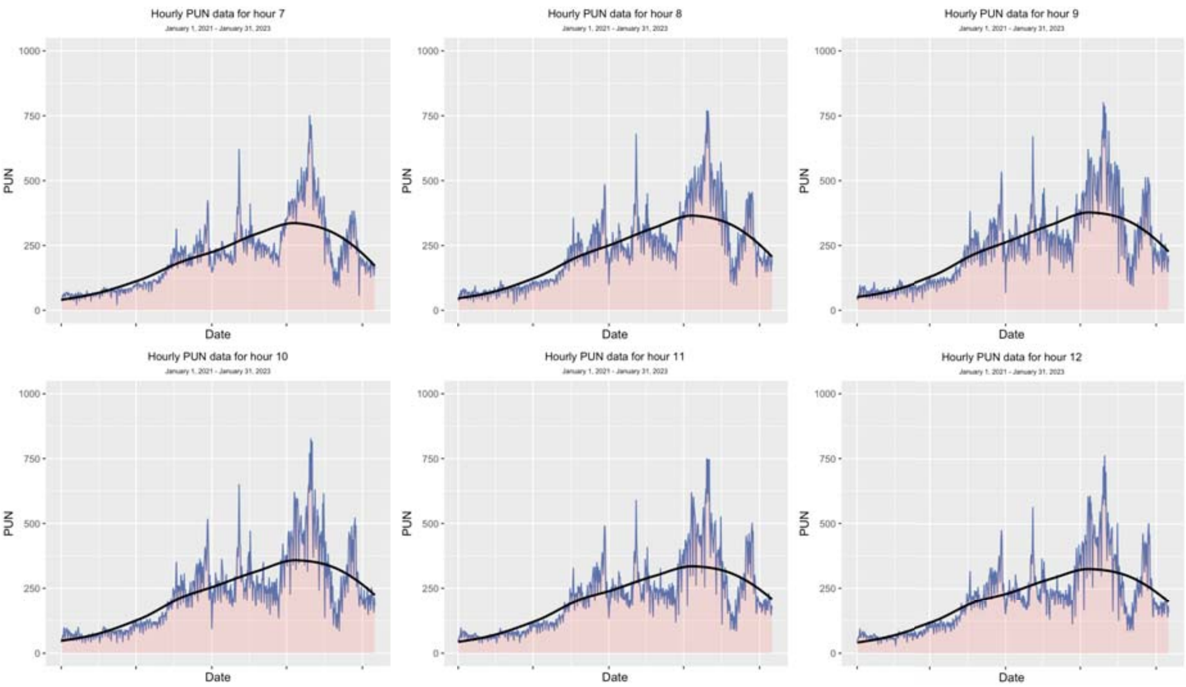

FIGURE 4.2: Hourly prices from 7 to 12. The continuous line is the non-linear estimated by splines.

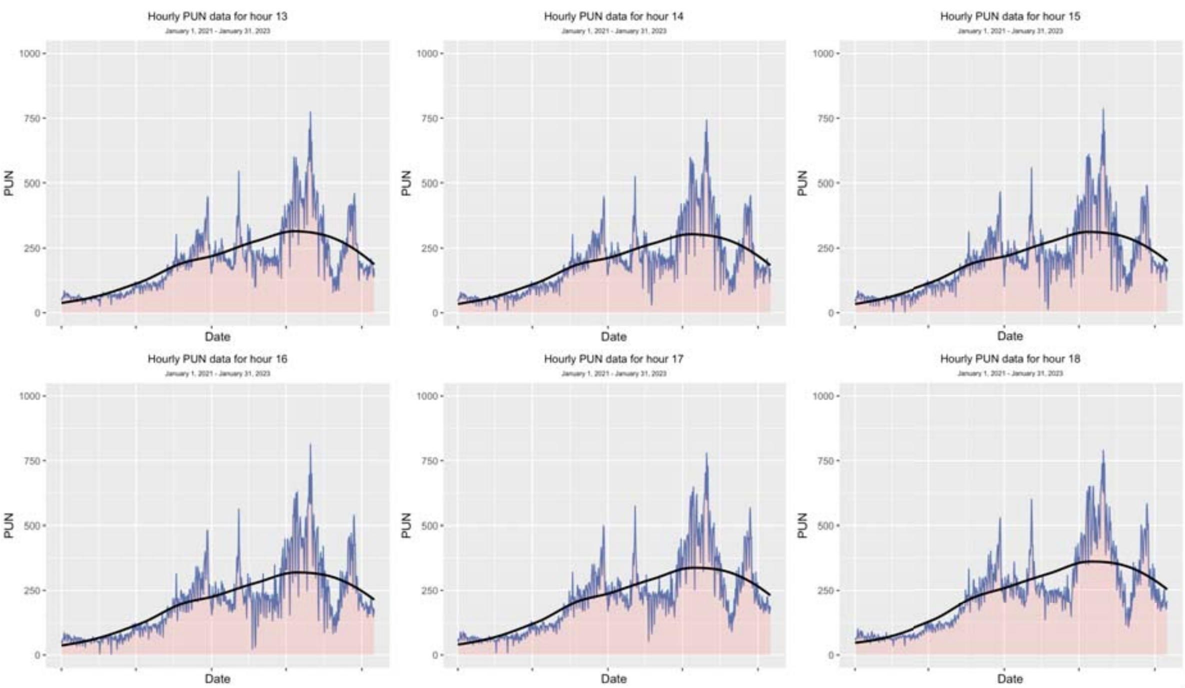

FIGURE 4.3: Hourly prices from 13 to 18. The continuous line is the non-linear estimated by splines.

As can be seen from the charts (4.1-4.4), prices tend to be higher in the morning, from 8 to 10, and in the evening, from 17/18 to 23. This pattern can be attributed to the high demand for electricity during these times when people wake up and get ready

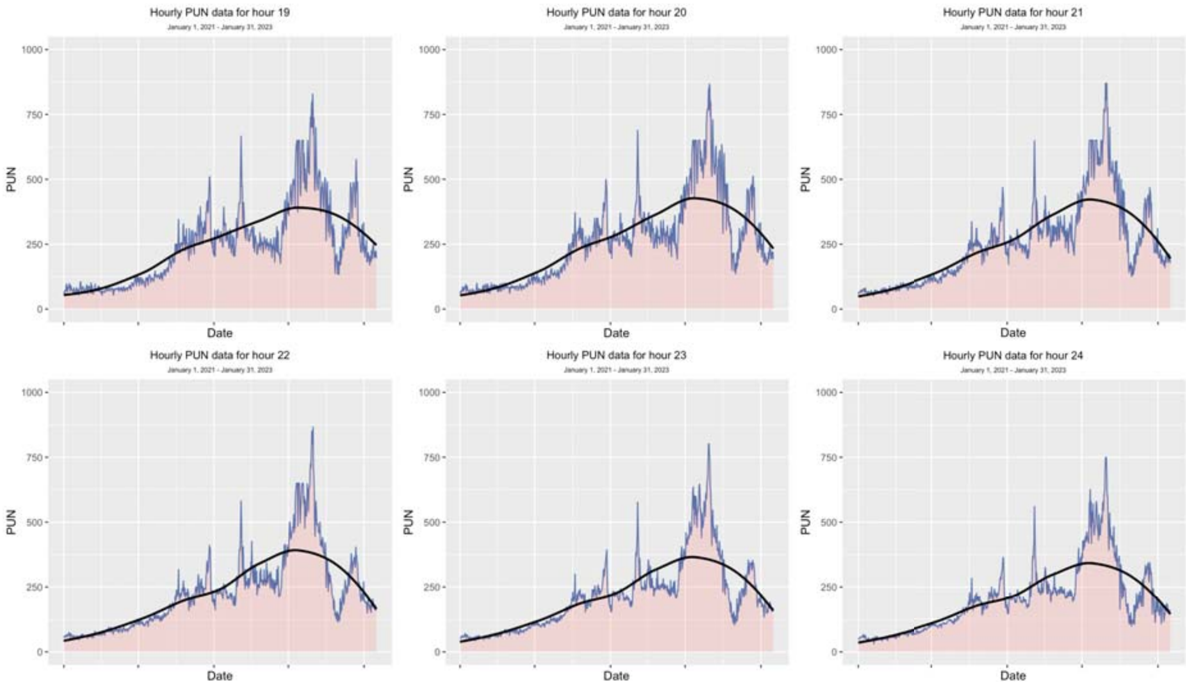

FIGURE 4.4: Hourly prices from 19 to 24. The continuous line is the non-linear estimated by splines.

for work/school and when they return home after work/school. This high demand leads to an increase in the PUN values, which reflects the market price for electricity in Italy.

The charts displayed in figure 4.5-4.6 show the distribution of the hourly PUN values in the Italian electricity market from January 1st, 2021 to January 31st, 2023. The PUN values are displayed in box plot format, with each box representing the interquartile range (IQR) of the data, the line inside the box representing the median, and the whiskers extending to the minimum and maximum values within 1.5 times the IQR. Additionally, a horizontal dashed red line is displayed above the plot to indicate the threshold for identifying potential outliers. These plots provide insight into the distribution of PUN values over the course of the day and allow us to identify any extreme values that may warrant further investigation.

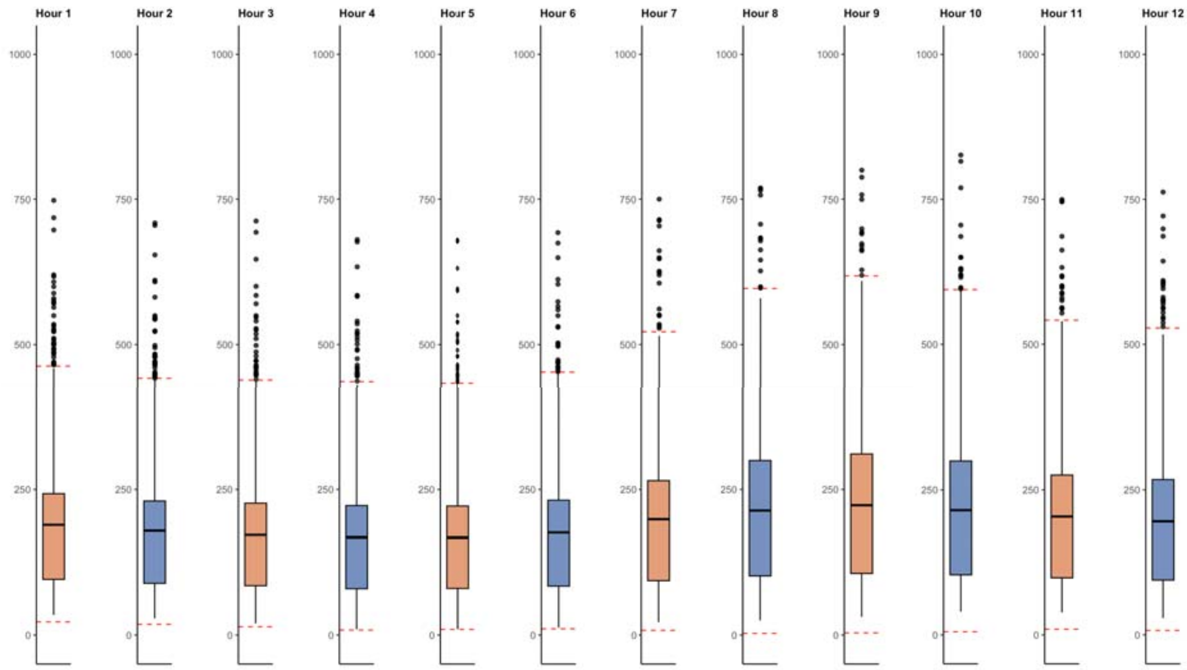

Figure 4.5: Hourly prices from 1 to 12

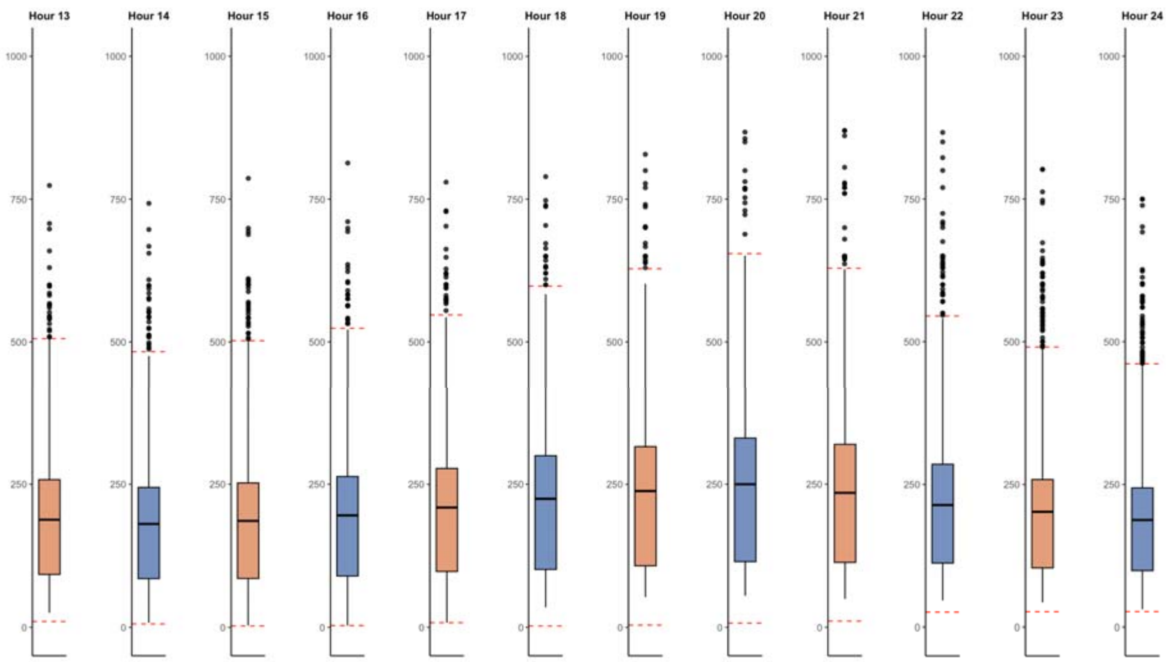

Figure 4.6: Hourly prices from 13 to 24

4.1.2 Stationarity Tests

The Dickey-Fuller (DF) test, proposed by David Dickey and Wayne Fuller (Dickey & Fuller, 1981), is a common statistical procedure designed to test the null hypothesis that a unit root is present in an autoregressive model. In the context of time series analysis, a unit root indicates non-stationarity, suggesting that the statistical properties of the series vary over time.

A simple autoregressive model of order one, AR(1), can be written as $Y_t = \phi Y_{t-1} + \varepsilon_t$, where Y_t is the variable of interest, ϕ is the autoregressive coefficient, and ε_t is a white noise error term. The DF test specifically tests the null hypothesis $H_0 : \phi = 1$ against the alternative hypothesis $H_1 : \phi < 1$. If the null hypothesis is not rejected, it implies the presence of a unit root and thus, the series is non-stationary.

Following are the results of the Dickey-Fuller Test done on the log-prices, and as we could expect there is a non-rejection of the null Hypothesis for all the time series and a rejection of the differenced time series:

Series	p_value	p_value_diff
Time Series 1	0.59	0.01
Time Series 2	0.54	0.01
Time Series 3	0.52	0.01
Time Series 4	0.46	0.01
Time Series 5	0.51	0.01
Time Series 6	0.53	0.01
Time Series 7	0.55	0.01
Time Series 8	0.46	0.01
Time Series 9	0.41	0.01
Time Series 10	0.37	0.01
Time Series 11	0.44	0.01
Time Series 12	0.44	0.01
Time Series 13	0.48	0.01
Time Series 14	0.43	0.01
Time Series 15	0.34	0.01
Time Series 16	0.45	0.01
Time Series 17	0.52	0.01
Time Series 18	0.58	0.01
Time Series 19	0.59	0.01
Time Series 20	0.62	0.01
Time Series 21	0.65	0.01
Time Series 22	0.73	0.01
Time Series 23	0.70	0.01
Time Series 24	0.71	0.01

Table 4.1: Dickey-Fuller Test Results

The Phillips & Perron (1988) test, developed by Peter C.B. Phillips and Pierre Perron, is a statistical methodology employed to investigate the presence of a unit root in a time series, thereby testing for stationarity. Much like the Dickey-Fuller test, the PP test considers an autoregressive model, however, it makes an adjustment for autocorrelated residuals, a feature not contemplated in the original Dickey-Fuller setup.

The PP test improves upon the Dickey-Fuller test by considering the possibility of autocorrelation within the error terms. Autocorrelated residuals can lead to inefficient and biased estimates if not accounted for, making the PP test a valuable tool when this assumption may be violated. As such, both the Dickey-Fuller and Phillips-Perron tests are fundamental in preliminary time series analysis, ensuring the correct identification of the data's underlying properties, and thus, the appropriateness of the subsequent modeling and forecasting.

Following are the results of the test:

Series	p_value_original	p_value_differenced
Time Series 1	0.12	0.01
Time Series 2	0.02	0.01
Time Series 3	0.01	0.01
Time Series 4	0.01	0.01
Time Series 5	0.01	0.01
Time Series 6	0.01	0.01
Time Series 7	0.01	0.01
Time Series 8	0.01	0.01
Time Series 9	0.01	0.01
Time Series 10	0.01	0.01
Time Series 11	0.01	0.01
Time Series 12	0.01	0.01
Time Series 13	0.01	0.01
Time Series 14	0.01	0.01
Time Series 15	0.01	0.01
Time Series 16	0.01	0.01
Time Series 17	0.01	0.01
Time Series 18	0.02	0.01
Time Series 19	0.05	0.01
Time Series 20	0.10	0.01
Time Series 21	0.31	0.01
Time Series 22	0.50	0.01
Time Series 23	0.55	0.01
Time Series 24	0.37	0.01

Table 4.2: Phillips-Perron Test Results

Since the p-value for the time series from 2 to 19 is smaller than the chosen significance level (0.05), we reject the Null Hypothesis, which suggests the presence of a Unit Root. Therefore, a model without differentiating the series is estimated.

4.2 Forecasting Plots

The presented time series plots 4.7-4.10 portray the Actual data, the best ETS (Error, Trend, Seasonality) model, and the LAD (Least Absolute Deviation) method over the period going from February 1st to April 30th.

The 'Actual' lines depict the true data points, which offer a clear picture of the inherent trend, and seasonal changes.

The 'Best ETS' line displays the predictive outcome of the ETS model. Although this model offers good forecasts in general, it may struggle with sudden shocks or anomalies in the data due to its inherent structure.

The 'Least Absolute Deviation' model is the third line, which by combining the forecasts shows to perform better specifically at certain time periods, hours that go from 7 to 20.

This conclusion is strengthened by looking at the comparison between the LAD combination and the ETS model in Table A.8 and in Table 4.3.

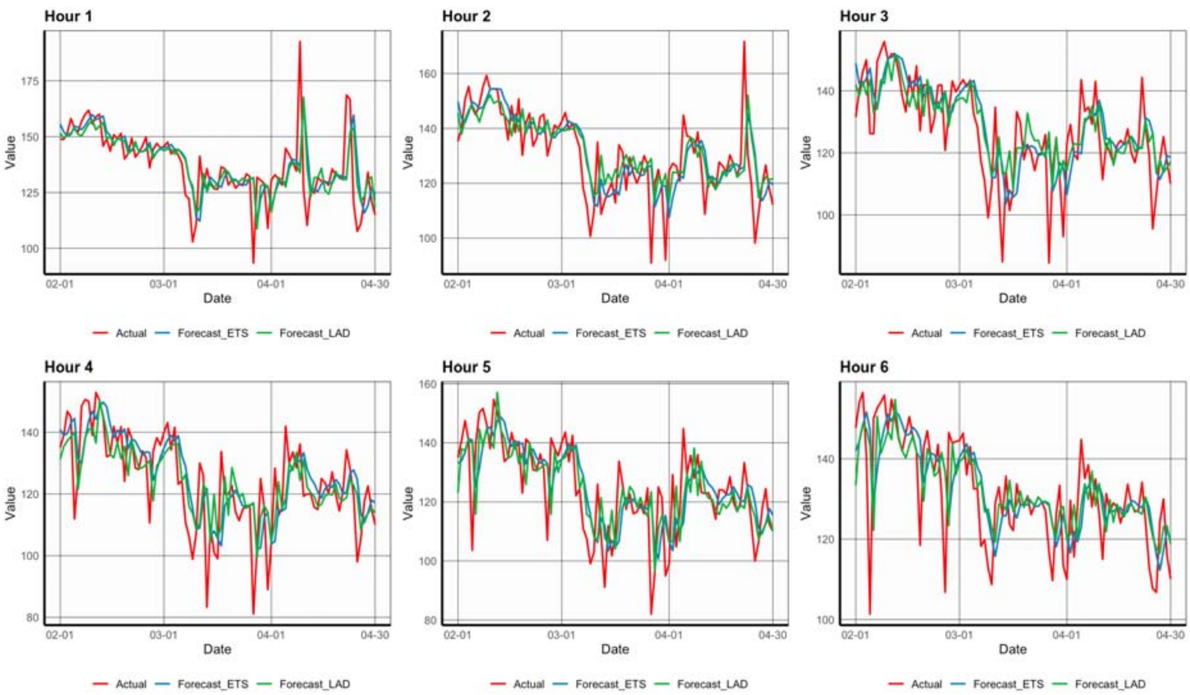

FIGURE 4.7: Hourly Forecasting Plots from 1 to 6

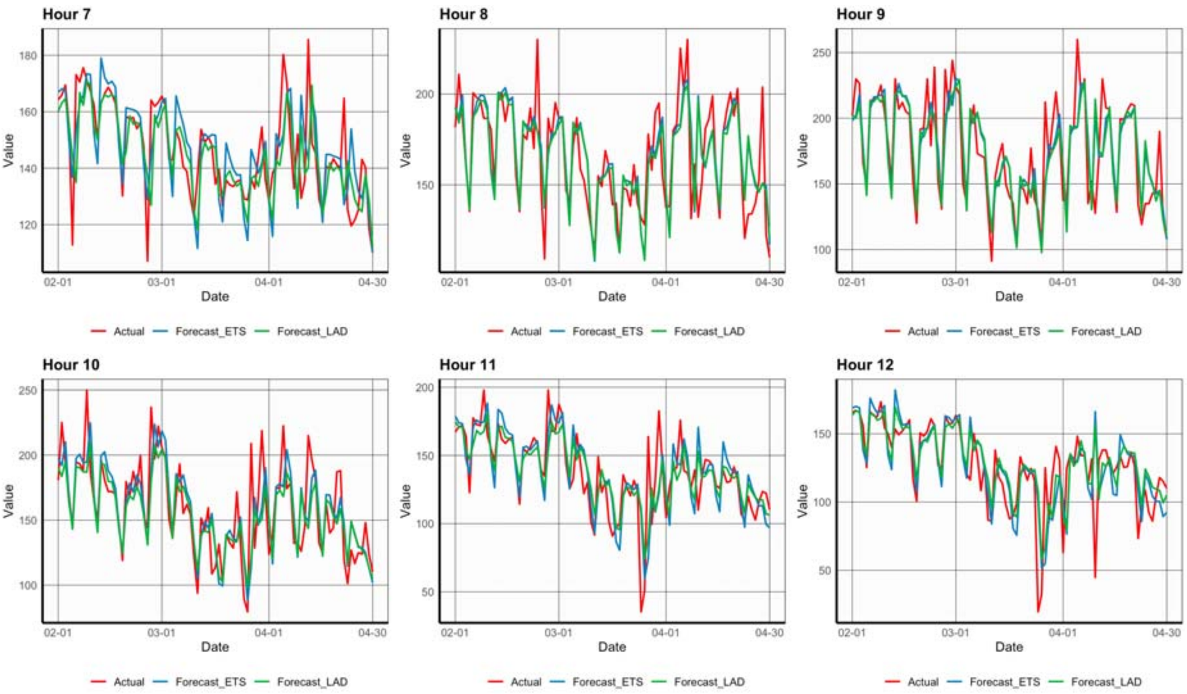

FIGURE 4.8: Hourly Forecasting Plots from 7 to 12

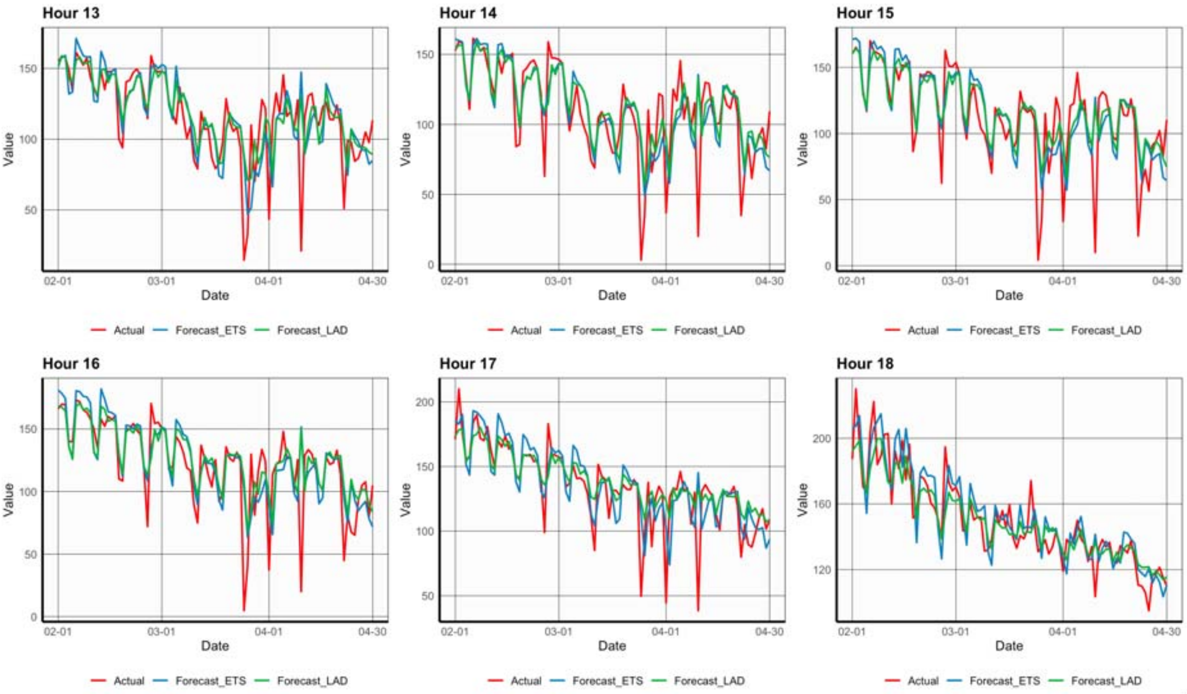

FIGURE 4.9: Hourly Forecasting Plots from 13 to 18

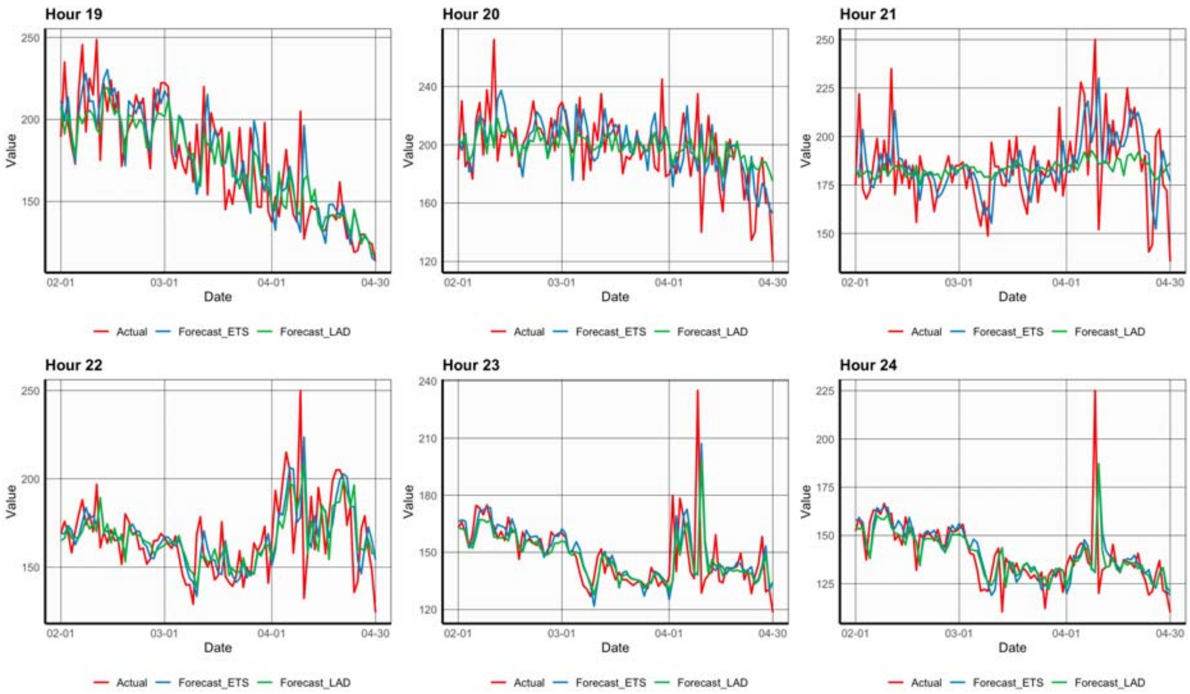

FIGURE 4.10: Hourly Forecasting Plots from 19 to 24

4.3 Forecasting Performance Comparison

TABLE 4.3: Forecasting Performance Table

	MSE	RMSPE	MAPE	MASE
ARIMA	440.6827	50.46526	15.76433	0.8640057
ETS	378.8342	48.73549	14.92897	0.8072066
SPLINE	448.2502	48.74873	15.43225	0.8596179
RandomForest	433.4511	50.55933	16.49392	0.8939981
CSR	335.5321	48.77521	14.62746	0.7748925
LAD	345.9516	50.23261	14.44716	0.7345586
NG	349.4714	47.85949	14.67107	0.7872265
OLS	328.5656	48.10136	14.47071	0.7677305
SA	384.4519	49.07497	15.10721	0.8141691

Table 4.3 presents a comparison of various forecasting models based on four error metrics. The models being compared are ARIMA, ETS, SPLINE, RandomForest, CSR, LAD, NG, OLS, and SA. The last five models (CSR, LAD, NG, OLS, SA) are all combined versions of the models.

Observing the Mean Squared Error (MSE), we notice that combination models like CSR and OLS yield lower values than the non-combination models, indicating they may provide better fits for the dataset. Specifically, CSR yields the smallest MSE.

Considering the Root Mean Squared Percentage Error (RMSPE), the NG model, another combination model, shows the least error, suggesting it could be more accurate in terms of percentage error than the other models.

When we look at the Mean Absolute Percentage Error (MAPE), which measures the absolute percentage deviation, the LAD model performs the best.

Finally, for the Mean Absolute Scaled Error (MASE), the LAD model again performs the best among all models.

In summary, while individual performance varies across metrics, the combination models (CSR, LAD, NG, OLS, SA) generally perform better than the non-combination models. Particularly, the LAD model appears to perform most consistently across different metrics.

Conclusions

In the case study of this thesis, the focus has been on hourly prices in the Italian market.

The study embarked on a journey to investigate the forecasting efficiency of forecast combinations over single forecasts. The culmination of this thesis validates the initial hypothesis by demonstrating the superiority of forecast combination methods over single forecast models.

This study is based on the realization that no single forecast model can enclose all the complexities inherent in a data set. Consequently, it explored the potential benefits of integrating multiple forecasts. The fundamental hypothesis underlying this investigation was the belief that combining different forecasts could mitigate the individual weaknesses of each model, thereby producing more accurate and reliable forecasts. The results of the study confirmed this hypothesis, as the combination of multiple forecasts demonstrated superior statistical performance over the majority of the time intervals tested.

Among the different time series, we noticed a consistent pattern in which forecast combinations, particularly the Least Absolute Deviation (LAD) model, outperformed individual models. They showed greater robustness to data variability, better adaptability to sudden changes, and overall higher forecast accuracy. Thus, the premise that the strength of one model can compensate for the weakness of another in a forecast combination became more than theoretical as it became empirical.

The results of this study invite a more comprehensive exploration. Further investigation could explore the effects of using different methods to find optimal weights, as well as testing additional forecasting methodologies. It would also be interesting to replicate this analysis under different circumstances, such as under different market conditions or using different data sets. This approach could provide broader insights and confirm the robustness of the results under various scenarios. In addition, the incorporation of emerging forecasting techniques and technologies, such as machine learning and artificial intelligence, could provide new insights into forecast combinations.

Appendix A

A.1 The choice of the models

Forecasting performances are calculated following formulas in (2.1).

Holt

TABLE A.1: Forecasting performance for Holt

Series	Model	MSE	RMSPE	MAPE	MASE
1	Holt	194.8086	10.855146	6.302005	0.9408369
2	Holt	148.1555	10.167490	7.008713	0.9234718
3	Holt	157.8804	11.239068	8.080082	0.9376083
4	Holt	160.7750	11.755738	7.702118	0.8900411
5	Holt	169.3648	11.564038	8.085371	0.9110387
6	Holt	131.7862	9.332468	6.718754	0.9941569
7	Holt	254.4639	11.379268	7.816679	0.9914582
8	Holt	871.4160	19.562437	14.824912	0.9838628
9	Holt	1498.7990	24.749134	19.464241	1.0085774
10	Holt	1125.5406	22.105870	17.470053	0.9279978
11	Holt	620.9925	31.976775	15.130060	0.8983710
12	Holt	671.8066	61.083466	21.903252	0.9376296
13	Holt	673.0553	85.063688	27.797333	0.9288163
14	Holt	881.1546	340.651496	61.816760	0.9439775
15	Holt	996.6939	292.429422	62.599573	0.9424394
16	Holt	941.3615	277.915877	53.231427	0.9323206
17	Holt	705.6174	40.319172	18.087995	0.9259500
18	Holt	267.4454	10.349543	8.037564	0.9585507
19	Holt	632.4892	14.394546	10.949379	0.8986839
20	Holt	680.7859	14.366259	10.720625	0.8966905
21	Holt	518.4657	12.815911	9.358799	0.8700798
22	Holt	444.0631	13.364764	8.733549	0.9129820
23	Holt	297.6628	10.811756	5.933634	0.9358160
24	Holt	240.6778	10.715756	5.920313	0.9303616

Holt Winters

Table A.2: Forecasting performance for HW

Series	Model	MSE	RMSPE	MAPE	MASE
1	HW	223.2951	11.103812	7.927016	1.2062210
2	HW	159.1844	10.422394	7.685758	1.0220869
3	HW	176.2076	11.604933	8.641012	1.0123150
4	HW	223.4655	13.380111	9.652779	1.1370332
5	HW	217.1186	12.410038	9.387004	1.0762351
6	HW	143.8335	9.487889	7.207469	1.0713845
7	HW	277.0372	11.466450	9.397788	1.2116823
8	HW	508.7870	14.846123	11.538539	0.7629945
9	HW	637.1899	15.544263	12.006930	0.6422378
10	HW	708.9168	17.113227	14.095837	0.7638317
11	HW	515.5839	27.412640	14.504867	0.8836680
12	HW	585.1271	53.712606	20.100213	0.8942854
13	HW	619.1490	85.533787	26.884384	0.9112153
14	HW	646.4234	278.732006	52.782732	0.7839232
15	HW	678.6356	256.290439	54.906655	0.7577869
16	HW	693.0034	224.849464	44.644742	0.7994695
17	HW	659.9844	38.507084	18.499131	1.0123804
18	HW	421.3075	13.895780	11.383747	1.3206185
19	HW	744.2616	15.514813	11.690438	0.9521283
20	HW	725.3770	14.268561	10.870189	0.9140285
21	HW	609.7793	13.762692	10.510212	0.9764334
22	HW	514.0478	14.437687	9.175783	0.9561884
23	HW	315.7357	11.351057	6.275969	0.9817543
24	HW	290.3343	11.709628	6.384747	1.0107247

Error - Trend - Seasonality

Table A.3: Forecasting performance for ETS

Series	Model	MSE	RMSPE	MAPE	MASE
1	MNN1	184.6831	10.607575	6.237616	0.9302095
2	MNN2	136.6914	9.897417	6.663537	0.8742342
3	MMN3	148.8034	11.102841	7.641282	0.8805286
4	MMN4	150.9118	11.479671	7.592068	0.8752864
5	MNN5	152.6548	11.070017	7.908901	0.8858238
6	MNN6	118.9297	9.161862	6.518719	0.9553698
7	MAM7	178.8186	9.380201	6.902508	0.8818575
8	MAM8	362.1473	12.280675	8.842705	0.5942302
9	MNM9	542.9782	13.744564	9.826246	0.5381555
10	MNM10	607.2737	14.975616	11.642245	0.6458244
11	MNM11	453.5964	26.105807	12.741553	0.7677768
12	MAM12	561.4296	54.814551	19.328247	0.8129077
13	MMM13	567.3150	85.119430	25.689179	0.8169438
14	MMM14	579.0504	290.313727	52.042128	0.7194401
15	MAM15	592.2110	253.560292	53.073903	0.6773723
16	MAM16	631.1892	227.414801	43.991383	0.7175340
17	MNM17	461.1064	37.570493	15.654902	0.7866310
18	MAA18	192.0367	9.454097	7.397545	0.8633879
19	MAM19	535.2290	13.631803	9.974751	0.8060217
20	MNA20	555.9144	12.869354	9.548618	0.7996341
21	MNN21	474.0191	12.256668	9.012410	0.8373547
22	MNN22	412.0743	12.767834	8.534586	0.8917543
23	MMM23	272.9105	10.267774	5.752409	0.9028979
24	MNM24	220.0467	9.804722	5.777878	0.9117822

It can be easily seen from tables A.1-A.3 that the best ETS (Error - Trend - Seasonality) model found per each time series (all possible combinations have been tried and model that presented the best forecasting performance has been selected) performs better both than Holt and Holt-Winters models (cases of RMSPE being better for Holt or HW are rare and the difference is little, thus there is no reason to prefer that model over the ETS class).

ARIMA

Table A.4: Forecasting performance for ARIMA

Series	Model	MSE	RMSPE	MAPE	MASE
1	ARIMA(7,1,0)	176.6261	10.226969	6.025099	0.9034882
2	ARIMA(7,0,0)	134.3910	9.660383	6.716115	0.8854125
3	ARIMA(7,0,0)	160.6189	11.322622	8.095812	0.9402004
4	ARIMA(7,0,0)	159.0919	11.659499	7.904977	0.9162143
5	ARIMA(7,0,0)	169.0728	11.503034	8.152361	0.9198143
6	ARIMA(7,0,0)	125.7102	9.190372	6.798955	1.0041442
7	ARIMA(7,0,0)	183.5310	9.806684	6.771965	0.8521611
8	ARIMA(7,0,0)	513.4157	14.412356	11.416072	0.7704144
9	ARIMA(7,0,0)	821.9761	16.835542	13.692489	0.7389685
10	ARIMA(7,0,0)	781.5655	17.677961	14.244470	0.7685053
11	ARIMA(7,0,0)	502.7964	27.874463	13.717404	0.8290781
12	ARIMA(7,0,0)	582.2985	54.246475	19.175474	0.8494999
13	ARIMA(7,0,0)	648.2412	80.609722	26.333537	0.9286192
14	ARIMA(7,0,0)	728.2519	305.507780	55.071638	0.8319788
15	ARIMA(7,0,0)	836.2139	263.150720	55.788957	0.8284879
16	ARIMA(7,0,0)	752.6635	239.478251	45.091894	0.7919211
17	ARIMA(7,0,0)	543.5687	36.541867	16.249653	0.8307360
18	ARIMA(7,0,0)	204.0451	9.241608	7.130843	0.8454128
19	ARIMA(7,0,0)	549.0851	13.267790	10.259716	0.8435085
20	ARIMA(7,1,0)	643.3754	14.041045	10.310906	0.8597128
21	ARIMA(7,1,0)	494.5999	12.543323	9.161942	0.8523164
22	ARIMA(7,1,0)	399.2890	12.669806	8.375965	0.8755426
23	ARIMA(7,1,0)	246.0518	9.623547	5.901374	0.9301790
24	ARIMA(7,1,0)	219.9065	10.074421	5.956210	0.9398217

Random Forest

Table A.5: Forecasting performance for Random Forest

Series	Model	MSE	RMSPE	MAPE	MASE
1	Random Forest	195.3439	10.927763	6.906471	1.0238869
2	Random Forest	141.5748	10.093316	6.860839	0.8953425
3	Random Forest	145.8494	11.187560	7.784235	0.8954344
4	Random Forest	158.0006	11.822360	7.957520	0.9158449
5	Random Forest	164.6214	11.608516	8.228056	0.9154591
6	Random Forest	115.7709	9.177513	6.404533	0.9318559
7	Random Forest	203.3077	10.850381	7.495310	0.9231312
8	Random Forest	566.3334	16.295990	11.877128	0.7684845
9	Random Forest	806.2015	19.082347	14.529780	0.7353361
10	Random Forest	846.0731	21.153088	16.378293	0.8320517
11	Random Forest	504.9666	30.145489	15.151382	0.8674940
12	Random Forest	598.9582	58.408337	22.372208	0.9366856
13	Random Forest	591.1828	80.558665	27.437350	0.8997803
14	Random Forest	681.1398	294.249731	55.090274	0.8322420
15	Random Forest	718.9896	257.863625	55.504128	0.7829670
16	Random Forest	676.6943	239.361460	46.523632	0.8125187
17	Random Forest	524.8039	35.748602	16.608465	0.8500867
18	Random Forest	284.3108	11.487639	8.992773	1.0313544
19	Random Forest	537.1972	13.963308	11.635903	0.9144822
20	Random Forest	579.2392	13.869078	10.322038	0.8396361
21	Random Forest	487.2534	12.871046	9.322590	0.8522893
22	Random Forest	417.7612	13.049046	9.341134	0.9626416
23	Random Forest	242.8352	9.803799	6.622250	1.0228310
24	Random Forest	214.4184	9.845303	6.507670	1.0141175

SPLINE

Table A.6: Forecasting performance for SPLINE

Series	Model	MSE	RMSPE	MAPE	MASE
1	SPLINE - ARIMA (7,0,0)	186.6574	10.301139	6.098368	0.9249691
2	SPLINE - ARIMA (7,0,0)	142.8329	9.735062	6.807050	0.9060594
3	SPLINE - ARIMA (7,0,0)	157.3336	10.999902	7.998411	0.9361670
4	SPLINE - ARIMA (7,0,0)	150.9003	11.184962	7.535804	0.8776735
5	SPLINE - ARIMA (7,0,0)	161.9937	11.148226	7.702580	0.8723717
6	SPLINE - ARIMA (7,0,0)	125.4922	8.998134	6.410208	0.9568498
7	SPLINE - ARIMA (7,0,0)	215.7137	10.312699	6.865994	0.8782949
8	SPLINE - ARIMA (7,0,0)	639.8368	15.766229	12.139912	0.8320489
9	SPLINE - ARIMA (7,0,0)	939.8511	17.637440	14.397296	0.7856129
10	SPLINE - ARIMA (7,0,0)	816.8854	17.721509	14.135198	0.7735174
11	SPLINE - ARIMA (7,0,0)	495.4144	27.536790	13.398690	0.8153052
12	SPLINE - ARIMA (7,0,0)	537.4199	53.285797	19.172080	0.8335872
13	SPLINE - ARIMA (7,0,0)	570.0708	77.058191	25.300666	0.8632863
14	SPLINE - ARIMA (7,0,0)	655.3167	287.946243	52.161283	0.7944469
15	SPLINE - ARIMA (7,0,0)	705.4506	248.541601	52.673262	0.7773521
16	SPLINE - ARIMA (7,0,0)	706.2945	231.888963	44.178165	0.7844451
17	SPLINE - ARIMA (7,0,0)	590.1085	36.612533	16.642432	0.8657552
18	SPLINE - ARIMA (7,0,0)	218.0134	9.197366	6.993712	0.8470011
19	SPLINE - ARIMA (7,0,0)	622.6959	13.928481	10.375852	0.8690786
20	SPLINE - ARIMA (7,0,0)	657.3000	13.841010	10.280060	0.8679845
21	SPLINE - ARIMA (7,0,0)	515.9001	12.556743	9.257284	0.8697232
22	SPLINE - ARIMA (7,0,0)	428.0644	12.952249	8.536109	0.8994696
23	SPLINE - ARIMA (7,0,0)	281.0446	10.345627	5.700367	0.9069098
24	SPLINE - ARIMA (7,0,0)	237.4147	10.472701	5.613324	0.8929204

According to the results, it's evident that the model performance varies significantly across different series. Across all models, the Mean Squared Error (MSE), Root Mean Squared Percentage Error (RMSPE), Mean Absolute Percentage Error (MAPE), and Mean Absolute Scaled Error (MASE) were utilized to evaluate the model's predictive capabilities.

Looking at the ARIMA model results in Table A.4, it can be seen that this model performed well on certain series while performing worse on others. The same can be said for the Random Forest model, as shown in Table A.5 and for the SPLINE model, as shown in Table A.6. This indicates that the performance of the model depends highly on the characteristics of the individual series.

A.2 Comparison with Diebold Mariano test

Following are the results of the Diebold Mariano test. To show the statistical gain of Forecast combinations, the best model in terms of Combinations is compared with the single forecasting models, and even against the best single forecasting model according to MSE, RMSPE, MAPE and MASE, the combinations show a statistical improvement.

ARIMA

Table A.7: LAD vs Arima

	Hour	Combination	Individual	DM_Stat	p_value
DM...1	1	LAD	Arima	-1.14	0.26
DM...2	2	LAD	Arima	-1.40	0.17
DM...3	3	LAD	Arima	-2.03	0.05
DM...4	4	LAD	Arima	-2.05	0.04
DM...5	5	LAD	Arima	-1.93	0.06
DM...6	6	LAD	Arima	-2.05	0.04
DM...7	7	LAD	Arima	-2.19	0.03
DM...8	8	LAD	Arima	-3.33	0.00
DM...9	9	LAD	Arima	-4.54	0.00
DM...10	10	LAD	Arima	-3.69	0.00
DM...11	11	LAD	Arima	-3.20	0.00
DM...12	12	LAD	Arima	-1.82	0.07
DM...13	13	LAD	Arima	-2.60	0.01
DM...14	14	LAD	Arima	-2.45	0.02
DM...15	15	LAD	Arima	-2.94	0.00
DM...16	16	LAD	Arima	-2.23	0.03
DM...17	17	LAD	Arima	-3.57	0.00
DM...18	18	LAD	Arima	-2.62	0.01
DM...19	19	LAD	Arima	-1.55	0.13
DM...20	20	LAD	Arima	-2.89	0.00
DM...21	21	LAD	Arima	-1.71	0.09
DM...22	22	LAD	Arima	-0.90	0.37
DM...23	23	LAD	Arima	-1.93	0.06
DM...24	24	LAD	Arima	-2.21	0.03

Upon inspecting the data presented in Table A.7, several points can be discussed about the performance of the Least Absolute Deviation (LAD) forecast combination method compared to the individual ARIMA model across 24 hours.

The Diebold-Mariano (DM) Statistic is used to compare the predictive accuracy of these two models, and the associated p-value provides a statistical significance level for this comparison. A negative DM Statistic indicates the LAD method's superior forecasting performance over the ARIMA model, while a p-value less than 0.05 implies statistical significance at the 95% confidence level.

We observe that the DM Statistic is negative for all 24 hours, indicating that the LAD combination forecast consistently outperforms the ARIMA forecast. Moreover, the DM Statistic becomes increasingly negative from hour 8 onwards, indicating that the performance of the LAD model becomes increasingly superior to the ARIMA model during these hours.

In this case, the p-value is less than 0.05 for hours 4, 6-11,13-18, as well as at hour 20 and 24. This suggests that for these hours, the superior performance of the LAD forecast over the ARIMA forecast is statistically significant.

However, it is important to note that during the other hours the p-value is above the 0.05 threshold, suggesting that we cannot reject the null hypothesis that the two methods perform equally well.

ETS

Table A.8: LAD vs ETS

	Hour	Combination	Individual	DM_Stat	p_value
DM...1	1	LAD	ETS	-1.37	0.17
DM...2	2	LAD	ETS	-1.18	0.24
DM...3	3	LAD	ETS	-1.36	0.18
DM...4	4	LAD	ETS	-1.27	0.21
DM...5	5	LAD	ETS	-1.40	0.16
DM...6	6	LAD	ETS	-1.37	0.18
DM...7	7	LAD	ETS	-2.24	0.03
DM...8	8	LAD	ETS	-1.24	0.22
DM...9	9	LAD	ETS	-1.99	0.05
DM...10	10	LAD	ETS	-2.77	0.01
DM...11	11	LAD	ETS	-1.84	0.07
DM...12	12	LAD	ETS	-2.09	0.04
DM...13	13	LAD	ETS	-1.53	0.13
DM...14	14	LAD	ETS	-1.34	0.18
DM...15	15	LAD	ETS	-1.90	0.06
DM...16	16	LAD	ETS	-2.46	0.02
DM...17	17	LAD	ETS	-3.18	0.00
DM...18	18	LAD	ETS	-2.73	0.01
DM...19	19	LAD	ETS	-1.19	0.24
DM...20	20	LAD	ETS	-2.20	0.03
DM...21	21	LAD	ETS	-1.56	0.12
DM...22	22	LAD	ETS	-1.07	0.29
DM...23	23	LAD	ETS	-1.64	0.10
DM...24	24	LAD	ETS	-2.21	0.03

Table A.8 shows the comparison of forecast performance between the LAD forecast combination method and the individual ETS model. Here, we see the DM Statistic is negative for every hour, suggesting that the LAD consistently outperforms ETS over 24 hours.

Moreover, the p-value is below the 0.05 threshold (indicating significant outperformance by the LAD model over the ETS model) for hours 7, 10, 12, 16-18, 20, and 24.

However, there are hours (1-6, 8, 9, 11, 13-15, 19, and 21-23) where the p-value is above the 0.05 level. This indicates that, while the LAD method shows superior performance in these periods, the results are not statistically significant. This led, during the analysis, to consider the ETS as a whole as the best model with which to compare the LAD, in the 4.2 section.

RandomForest

Table A.9: LAD vs Random Forest

	Hour	Combination	Individual	DM_Stat	p_value
DM...1	1	LAD	RandomForest	-3.13	0.00
DM...2	2	LAD	RandomForest	-1.91	0.06
DM...3	3	LAD	RandomForest	-1.86	0.07
DM...4	4	LAD	RandomForest	-2.00	0.05
DM...5	5	LAD	RandomForest	-1.74	0.08
DM...6	6	LAD	RandomForest	-0.95	0.35
DM...7	7	LAD	RandomForest	-2.23	0.03
DM...8	8	LAD	RandomForest	-3.07	0.00
DM...9	9	LAD	RandomForest	-4.51	0.00
DM...10	10	LAD	RandomForest	-4.54	0.00
DM...11	11	LAD	RandomForest	-3.58	0.00
DM...12	12	LAD	RandomForest	-3.06	0.00
DM...13	13	LAD	RandomForest	-2.95	0.00
DM...14	14	LAD	RandomForest	-2.86	0.01
DM...15	15	LAD	RandomForest	-3.28	0.00
DM...16	16	LAD	RandomForest	-2.85	0.01
DM...17	17	LAD	RandomForest	-4.32	0.00
DM...18	18	LAD	RandomForest	-3.78	0.00
DM...19	19	LAD	RandomForest	-2.60	0.01
DM...20	20	LAD	RandomForest	-3.45	0.00
DM...21	21	LAD	RandomForest	-2.00	0.05
DM...22	22	LAD	RandomForest	-2.42	0.02
DM...23	23	LAD	RandomForest	-2.46	0.02
DM...24	24	LAD	RandomForest	-2.58	0.01

Table A.9 represents the comparative forecast performance of the LAD forecast combination method and the individual Random Forest model. The DM Statistic is negative across all hours, suggesting the LAD model's consistent outperformance over the Random Forest.

It's noteworthy that p-values are below the significance level (0.05) for hours 1, 4, 7-20 and 22-24. This indicates a statistically significant superior performance of the LAD model over the Random Forest model during most of the hours of the day.

However, during hours 2, 3, 5, 6 and 21 we cannot reject the null hypothesis that the two methods perform equally well. Despite these hours, the results reinforce the overall finding that the LAD combination forecast method generally yields better results than individual models.

SPLINE

TABLE A.10: LAD vs SPLINE

	Hour	Combination	Individual	DM_Stat	p_value
DM...1	1	LAD	SPLINE	-1.35	0.18
DM...2	2	LAD	SPLINE	-1.61	0.11
DM...3	3	LAD	SPLINE	-1.96	0.05
DM...4	4	LAD	SPLINE	-1.49	0.14
DM...5	5	LAD	SPLINE	-1.45	0.15
DM...6	6	LAD	SPLINE	-1.23	0.22
DM...7	7	LAD	SPLINE	-2.87	0.01
DM...8	8	LAD	SPLINE	-3.79	0.00
DM...9	9	LAD	SPLINE	-4.79	0.00
DM...10	10	LAD	SPLINE	-3.74	0.00
DM...11	11	LAD	SPLINE	-2.89	0.00
DM...12	12	LAD	SPLINE	-1.63	0.11
DM...13	13	LAD	SPLINE	-2.18	0.03
DM...14	14	LAD	SPLINE	-2.12	0.04
DM...15	15	LAD	SPLINE	-2.81	0.01
DM...16	16	LAD	SPLINE	-2.38	0.02
DM...17	17	LAD	SPLINE	-3.59	0.00
DM...18	18	LAD	SPLINE	-2.40	0.02
DM...19	19	LAD	SPLINE	-1.71	0.09
DM...20	20	LAD	SPLINE	-3.05	0.00
DM...21	21	LAD	SPLINE	-1.87	0.06
DM...22	22	LAD	SPLINE	-1.08	0.28
DM...23	23	LAD	SPLINE	-1.53	0.13
DM...24	24	LAD	SPLINE	-1.36	0.18

Table A.10 contrasts the forecast performance of the LAD combination method and the individual SPLINE model.

The DM Statistic is negative across all hours, implying the LAD model generally provides better forecasts. This difference is statistically significant (p-value < 0.05) for hours 7-11, 13-18, and 20.

However, for the remaining hours even though LAD outperforms SPLINE we cannot reject the null hypothesis that the two methods perform equally well.

As can be seen from the previous tables A.7-A.10 the DM statistic is always negative and in most cases statistically significant.

Bibliography

Bates, J. M. & Granger, C. W. (1969). The combination of forecasts. *Journal of the operational research society* **20**, 451–468.

Breiman, L. (2001). Random forests. *Machine learning* **45**, 5–32.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International journal of forecasting* **5**, 559–583.

Dickey, D. A. & Fuller, W. A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica* **49**, 1057–1072.

Giulietti, M. & Sicca, R. (1999). The liberalisation of the internal market for electricity: what choices for italy? *Utilities Policy* **8**, 173–182.

Hyndman, R. J. & Athanasopoulos, G. (2018). *Forecasting: principles and practice.* OTexts.

Makridakis, S., Spiliotis, E. & Assimakopoulos, V. (2020). The m4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting* **36**, 54–74.

Newbold, P. & Granger, C. W. J. (1974). Experience with forecasting univariate time series and the combination of forecasts. *Journal of the Royal Statistical Society. Series A (General)* **137**, 131–165.

Phillips, P. C. & Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika* **75**, 335–346.

Wang, X., Hyndman, R. J., Li, F. & Kang, Y. (2022). Forecast combinations: an over 50-year review. *International Journal of Forecasting* .

Weiss, C. E., Roetzer, G. R. & Raviv, E. (2018). *ForecastComb: Forecast Combination Methods.* R package version 1.3.1.