Master Thesis Econometrics and Operations Research

Advanced Forecasting with the Theta-MLP Model: Insights from the M4 Competition

QAV
2676947

September 10, 2024

Supervisor: MM
Second Assessor: LH

Acknowledgement

I would like to thank my supervisor, Professor MM, for his guidance and support throughout this thesis. His expertise and feedback were invaluable in addressing challenges and refining my work. I am grateful for his encouragement and the freedom he gave me to develop this thesis independently. I would also like to thank my second assessor, LL, for his role in evaluating my work and contributing to the completion of this thesis.

Abstract

This thesis evaluates the performance of various statistical forecasting methods on high-frequency time series data and introduces a novel hybrid approach—the Theta-MLP method—which combines extensions of the Classical Theta model with a Multilayer Perceptron (MLP) to capture non-linear residual patterns. The study is structured into three parts: Part I outlines the theoretical framework, Part II conducts simulation experiments using synthetically generated hourly time series characterized by diverse trends and seasonalities, and Part III applies the models to real-world hourly and daily time series from the M4 Competition.

The results demonstrate that the Theta-MLP method consistently outperforms simpler statistical benchmarks across both simulated and real-world datasets, achieving substantial accuracy improvements while highlighting the benefits of integrating statistical and machine learning approaches. The findings support the hypothesis that complex methods achieve superior forecasting accuracy. However, the study provides mixed evidence regarding the effectiveness of combining simpler statistical methods, with their performance highly dependent on the characteristics of the time series, such as trend strength and seasonality. Furthermore, the results underscore the importance of trend damping in improving forecasting accuracy for moderate trends, while cautioning against its application in strongly trending series due to the risk of under-forecasting. These insights advance the understanding of hybrid forecasting models and offer a foundation for the development of more sophisticated forecasting methods.

Contents

Introduction

Forecasting competitions have long been a cornerstone for evaluating the effectiveness of forecasting methods, providing empirical benchmarks akin to laboratory experiments in the physical sciences (Hyndman, 2020). These competitions have driven advancements in both theoretical understanding and practical applications, promoting the adoption of forecasting techniques by decision-makers and policymakers (Fildes & Ord, 2004). Specifically, they aim to identify factors that enhance forecasting accuracy by empirically evaluating various forecasting methods and determining the most effective approaches for different data types and forecasting horizons. Among these, the Makridakis competitions, commonly referred to as the M Competitions, have been particularly influential in shaping the field of forecasting (Hyndman, 2020).

The M3 Competition, held in 1998, featured 3,003 time series and demonstrated that combining methods, such as the Comb S-H-D raw ensemble of Single, Holt, and Damped exponential smoothing (ES), often outperformed individual methods on average in terms of symmetric MAPE (sMAPE), although the difference from Damped ES was small (Makridakis & Hibon, 2000). Damped ES, a simple forecasting technique designed to dampen linear trends (Gardner & McKenzie, 2010) and a benchmark in the M3 Competition, along with other simple models, outperformed more complex methods on average in terms of symmetric MAPE. This supported the conclusion that "statistically sophisticated or complex methods do not necessarily produce more accurate forecasts than simpler ones" (Makridakis & Hibon, 2000).

Makridakis & Hibon (2000) attributed this to the fact that many time series behave like random walks, aligning their reasoning with the finding that forecasting accuracy can be improved by recognizing that many trends follow a random walk and that established trends in the data can and do change. The Damped ES method exemplifies this approach (Fildes & Makridakis, 1995; Makridakis & Hibon, 2000). Indeed, Gardner & McKenzie (2010) compared linear and damped trend versions of ES forecasts for the 3,003 time series originally collected for the M3 Competition. Their results revealed that incorporating trend damping enhanced forecast accuracy, particularly for extended horizons. Nonetheless, Gardner (2015) acknowledged that the damped trend method is not universally superior and may underperform in certain cases, particularly when applied to strongly trending series. Furthermore, the main conclusions of the M3 Competition were derived from descriptive statistical analyses with no formal statistical testing. Subsequent analysis using the Multiple Comparisons with the Best (MCB) procedure revealed that the claim of a combination of methods outperforming the methods being combined was not statistically significant (Fildes *et al.*, 2005).

The M4 Competition, held in 2018, significantly expanded its scope by including 100,000 time series across diverse frequencies, offering greater opportunities for complex methods that require substantial amounts of data for effective training. While the M3 Competition and related studies (Crone *et al.*, 2011; Nikolopoulos & Petropoulos, 2018) found that statistically sophisticated or complex methods do not necessarily outperform simpler ones, this trend was reversed in the M4 Competition. On average, the accuracy of point forecasts improved as computational time increased, indicating that greater processing power was effectively leveraged to achieve higher accuracy (Assimakopoulos *et al.*, 2020). Another important finding of the M4 Competition was that all of the top-performing methods, in terms of both point forecasts and prediction intervals, were combinations of mostly statistical methods, with such combinations being more accurate numerically than either pure statistical or pure machine learning methods. The results also highlighted that upward/downward trend extrapolation is predicted more accurately when the trend is damped for longer horizons (Assimakopoulos *et al.*, 2020; Assimakopoulos *et al.*, 2018).

This divergence in findings between the M3 and M4 Competitions reflects ongoing debates in forecasting research regarding the roles of model complexity, trend damping, and combining methods. As noted earlier, simple methods were shown to outperform more complex methods in earlier studies, such as the M3 Competition, while this trend was reversed in the M4 Competition. Similarly,

combining methods were shown to be effective in the M3 Competition; however, only one combination method (Comb) was considered, and its performance was not always statistically significant (Franses *et al.*, 2005). McKenzie (2010) further noted that, while the motivation for damping trends is intuitively appealing, its effectiveness is not always well understood, and little is known about why or when it provides optimal forecasts. Fry & Brundage (2020) also only partially agreed with the improved accuracy of combining methods in the M4 Competition, emphasizing the importance of differentiating between raw ensembling, such as the Comb, and smart combining, such as Smyl's hybrid algorithm, the winning method of the M4 Competition. The latter refers to hybrid models that integrate multiple forecasting methods in a directed and problem-specific manner. Notably, Smyl's hybrid algorithm (ES-RNN) was the only true hybrid algorithm run in the M4 Competition at the time. Since the winning methods are better classified as hybrid methods or examples of smart combining rather than mere raw ensembling, the role of combining might be overstated, aligning with the findings of Franses *et al.* (2005).

This thesis comprehensively evaluates forecasting methods using both simulated and real-world data from the M4 Competition to identify key factors that enhance forecasting accuracy, focusing on forecast combination, trend damping, and computational complexity. Central to this research is the development of the Theta-MLP method, a hybrid forecasting algorithm that extends the Classical Theta method through: (i) a Box-Cox transformation, (ii) linear and non-linear trend components, (iii) additive and multiplicative Theta lines, (iv) additive and multiplicative seasonality, and (v) optimization of θ using the Mean Absolute Error (MAE) instead of predefined values, which allows the slope of the trend to either be damped or expanded when needed. These enhancements yield a topology of eight candidate models, from which the best-performing model is selected for forecasting using the in-sample MAE. Furthermore, the Theta-MLP method incorporates a Multi-Layer Perceptron (MLP) component, regulated by the parameter ϕ, to introduce an error correction term that captures residual non-linearities from the selected Theta model.

The guiding research question is: *How do model complexity, trend damping, and combining methods influence forecasting accuracy across different time series data?* This question builds on unresolved debates in prior studies, particularly those from the M3 and M4 Competitions, concerning the effectiveness of simple versus complex forecasting methods (Makridakis & Hibon, 2000; Crone *et al.*, 2011; Assimakopoulos *et al.*, 2020), the benefits of combining methods (Franses *et al.*, 2005; Assimakopoulos *et al.*, 2020), and the role of trend damping in improving forecast accuracy (Gardner & McKenzie, 2010; Gardner, 2015). By systematically addressing these issues, the thesis advances forecasting research through simulation studies and empirical analyses, demonstrating that accuracy gains achieved via machine learning in hybrid approaches justify the cost of additional complexity.

The study has three primary objectives: first, to test whether combining statistical methods consistently outperforms individual methods (**H1**); second, to examine whether more complex methods, such as Theta-MLP, lead to higher forecasting accuracy compared to simpler models (**H2**); and third, to investigate whether trend damping improves overall accuracy on average, while potentially causing under-forecasting in strongly trending series (**H3**). These objectives aim to provide insights into how model complexity, trend damping, and method combination influence forecasting performance, offering valuable implications for both theory and practical applications.

To test these hypotheses, the research uses simulated hourly data-generating processes (DGPs) designed to reflect varying levels of trend complexity—linear, damped, and exponential—while incorporating additive and multiplicative seasonality. Each DGP is simulated $T = 100$ times to ensure robust results, with each simulation consisting of 505 observations, corresponding to just over 21 days, or slightly more than three full cycles of weekly seasonality. In addition, real-world data from the M4 Competition are analyzed, comprising 414 hourly series with moderate trends and strong seasonality, and 4,227 daily series characterized by very strong trends and minimal seasonality, with varying series lengths. The Overall Weighted Average (OWA), which combines sMAPE and MASE, is the primary metric used to evaluate forecasting accuracy. The Diebold-Mariano (DM) Test assesses the statistical significance of the observed differences in forecasting performance, based on mean squared error (MSE), at a 95% confidence level.

The findings confirm that the Theta-MLP hybrid algorithm consistently outperforms simpler models in both simulated and real-world datasets, validating that computational resources can be effectively utilized to achieve superior forecasting accuracy. In the simulation study, Theta-MLP achieved the lowest OWA across all three hourly DGPs, demonstrating substantial improvements over simpler benchmarks such as the Naïve method and the Classical Theta method upon which it is based. The DM test confirmed that the observed differences in forecasting performance are statistically significant at a 95% confidence level. Similarly, in real-world data from the hourly and daily series of the M4 Competition, the Theta-MLP method consistently secured the second-highest OWA rank. Notably, the DM test did not reject the null hypothesis of equal forecasting accuracy between the Theta-MLP method and the second-most accurate models in the hourly and daily series.

Furthermore, the results highlight the potential of trend-damping mechanisms in mitigating over-forecasting for moderate trends, as demonstrated by the hourly series of the M4 Competition and DGP 2 in the simulation study. However, their limitations become apparent in strongly trending series, such as the daily series, the exponential trend of DGP 3, or the linear trend of DGP 1. This underscores the need for adaptive mechanisms capable of dynamically responding to varying trend strengths.

At the same time, the findings emphasize the limitations of simple ensembling methods like Comb, which underperformed individual methods in most cases, both for the simulated DGPs and the hourly series of the M4 Competition, in terms of OWA. The DM test confirmed that the observed differences in performance were statistically significant, rejecting the null hypothesis of equal forecasting accuracy between Comb and the best-performing individual ES methods. However, for the daily series, Comb achieved the best OWA, although the difference compared to Theta-MLP was not statistically significant. This suggests that the success of combining methods may depend on specific data characteristics, such as the dominance of strong trends or seasonality. These findings underscore the importance of future research into adaptive and weighted combination strategies.

This thesis is structured into three parts: Part I provides a detailed overview of eight statistical benchmarks from the M4 Competition and introduces the Theta-MLP method as a novel hybrid approach; Part II presents extensive simulations using various hourly DGPs to test hypotheses related to model complexity, trend damping, and forecast combination; and Part III evaluates these hypotheses using real-world hourly and daily series from the M4 Competition, offering key insights into their practical effectiveness. Overall, this thesis contributes to forecasting theory by providing insights into how specific methodological choices influence accuracy across different scenarios and by introducing the Theta-MLP hybrid algorithm as an effective tool for forecasting time series.

1 Part I: Methodical Framework

Part I outlines the methodological framework underpinning this thesis, presenting an in-depth analysis of the statistical benchmarks used in the M4 Competition and introducing the novel hybrid Theta-MLP method. This hybrid model combines the robust trend and seasonality decomposition capabilities of the extensions of the Classical Theta method with the flexibility of a Multilayer Perceptron (MLP) to address non-linear patterns in time series data.

The chapter begins by discussing the foundational Naïve forecasting models—Naïve 1, Naïve 2, and Seasonal Naïve—which provide straightforward benchmarks. These are followed by the Exponential Smoothing (ES) family, a more advanced class of methods that includes Simple Exponential Smoothing (SES), Holt's Linear Trend method, and the Damped Additive Trend method. The Comb method, a simple arithmetic average of SES, Holt, and Damped Exponential Smoothing,

is introduced as a composite benchmark frequently used in forecasting competitions. The Classical Theta method is presented next, highlighting its innovative decomposition of time series into multiple θ-lines to enhance accuracy. Despite its success, particularly in the M3 Competition where it was identified as the most accurate method, the Classical Theta method has limitations in capturing non-linear trends and additive seasonality.

Building on these limitations, the Theta-MLP method advances the Classical Theta approach by extending it and integrating machine learning components. This hybrid model retains the strengths of the Classical Theta method, such as its ability to capture linear trends and handle multiplicative seasonality, while introducing several key innovations. These innovations include the optimization of θ using in-sample Mean Absolute Error (MAE), which allows the trend to be either damped or expanded as needed, the application of a Box-Cox transformation to stabilize variance, the inclusion of both additive and multiplicative seasonality and Theta lines, and a rigorous validation framework that selects the most appropriate model from a set of eight candidate Theta models based on in-sample MAE. Central to the Theta-MLP method is the inclusion of an MLP correction term, regulated by the parameter ϕ, which learns from residual patterns that the selected Theta model does not capture, significantly enhancing the model's ability to handle non-linearities and improve forecasting accuracy.

Part I concludes with an overview of the performance metrics—sMAPE, MASE, and OWA—used to evaluate forecasting accuracy across both simulated and real-world datasets, along with the Diebold-Mariano (DM) Test, which is applied to assess the statistical significance of observed differences in forecasting performance between competing methods, based on mean squared error (MSE), at a 95% confidence level.

Statistical Benchmarks in the M4 Competition: Methodological Framework and Model Characteristics

Since the inception of the Makridakis competitions in the 1980s, the forecasting field has evolved significantly. The initial conclusion that complex methods do not necessarily outperform simpler ones has been challenged by newly proposed methods (Assimakopoulos *et al.*, 2020). The M4 Competition reflects this evolution by including ten benchmarks, aiming to assess how modern approaches compare to traditional ones and to identify the specific improvements made over time. For the M4 Competition, eight statistical benchmarks have been selected. These benchmarks, along with two standards of comparison, were estimated using version 8.2 of the *forecast* package for R (Hyndman, 2017). The choice of these benchmarks allows for a comprehensive evaluation of current forecasting techniques against well-established methods, highlighting the progress and effectiveness of newer approaches. The statistical benchmarks encompass a variety of methods, including simple Naïve approaches, linear and damped trend models such as Holt and Damped ES, and a combination method, the Comb. Evaluating the performance of these benchmarks alongside the more complex Theta-MLP method enables the testing of the hypotheses outlined in the introduction.

1.1 Naïve Methods

In probability theory, Naïve methods, also known as random walk models, describe stochastic processes that result from the continuous summation of independent, identically distributed (i.i.d.) random variables (Spitzer, 2001). In simple terms, in a random walk, future movements or directions cannot be predicted based on past history (Lawler & Limic, 2010). Naïve methods accurately model a variety of natural phenomena. A traditional example comes from Louis Jean-Baptiste Alphonse Bachelier, a French mathematician who was the first to model Brownian motion and wrote the first paper applying advanced mathematics to finance. Bachelier's example illustrates a social vice, specifically describing the position of a drunk individual staggering left and right with equal probability while progressing forward (Spitzer, 2001). Numerous other natural phenomena are effectively modeled by random walks. For example, applying the Naïve method to the stock

market implies that short-term fluctuations in stock prices are inherently unpredictable (Lawler & Limic, 2010).

In the M4 Competition benchmarks, three variations of the Naïve model are utilized: the Naïve 1 model, representing the most basic version, and the Naïve 2 and Seasonal Naïve models, both of which are designed to address seasonality in time series data, albeit through distinct methodologies. The Seasonal Naïve model is particularly well-suited for datasets with well-defined seasonal patterns, whereas the Naïve 2 model builds upon the Naïve 1 model by incorporating seasonal adjustments when applicable (Assimakopoulos *et al.*, 2020). The inclusion of Naïve methods provides benchmarks to test **H2** by evaluating whether more complex forecasting methods, such as Theta-MLP, achieve higher accuracy compared to simpler approaches.

1.1.1 Naïve 1

The Naïve 1 model, also known as the simple random walk, is one of the simplest yet most important models in time series forecasting (Van Horne & Parker, 1967). It operates under the assumption that, in each period the variable takes a random step away from its previous value, and the steps are i.i.d in size (Nau, 2014). The Naïve 1 model is given by the following formula, where y_t and f_t are the actual and forecasted values at time t:

$$f_t = y_{t-1} \tag{1}$$

One property of the Naïve 1 model is its simplicity in assuming that all future values will equal the last observed value, making it easy to implement and understand, and not requiring complex computations or parameter estimations (Hyndman & Athanasopoulos, 2018). Moreover, given its simplicity, the Naïve 1 model also requires minimal computational resources, allowing for quick, preliminary forecasts (Hyndman *et al.*, 1998). Following this, the Naïve 1 model was incorporated as a benchmark in the M4 Competition, helping in the evaluation of more complex models.

While the Naïve 1 model is effective when data lacks significant trends or seasonal patterns, its simplicity often results in low accuracy, particularly for time series with complex patterns (Spitzer, 2001). The Naïve 1 model does not account for trends, seasonal variations, or cyclic behaviors in the data (Van Horne & Parker, 1967). Therefore, although useful as a baseline, it is generally unsuitable for producing accurate forecasts in real-world scenarios where data exhibits more intricate structures (Lawler & Limic, 2010).

The M4 Competition introduced both high-frequency and low-frequency data. The organizers considered different data frequencies: 12 for monthly, 4 for quarterly, 24 for hourly, and 1 for yearly, weekly, and daily data. There is clear seasonality in quarterly, monthly, and hourly frequencies, whereas seasonality is much more ambiguous in yearly, weekly, and daily data (Assimakopoulos *et al.*, 2020). A year, depending on leap years, may not consist of exactly 52 weeks, and a week, depending on business days, may consist of five, six, or seven days. For hourly data, double seasonality may occur with a cycle of 7 days $\times$ 24 hours, or it may display triple seasonality with a cycle of 7 days$\times$24 hours$\times$12 months. The seasonality exhibited in some of the data, combined with the Naïve 1 model's inability to capture seasonality, is the primary reason the M4 Competition's organizers chose not to use the Naïve 1 model as the benchmark (denominator) for scaling the absolute error of the examined method (numerator) in the mean absolute scaled error (MASE; Hyndman & Koehler, 2006), one of the accuracy measures used to evaluate point forecasts. This decision provides a more representative benchmark for seasonal series and a more accurate measure of performance.

1.1.2 Naïve 2

The Naïve 2 model is similar to the Naïve 1 model but is applied to seasonally adjusted data (Assimakopoulos *et al.*, 2020). Seasonality in data refers to periodic fluctuations that recur at regular intervals due to factors such as weather, holidays, or cultural events. These patterns repeat at a consistent frequency, which defines how often the fluctuations occur within a specific time frame. To determine whether a series requires seasonal adjustment, a seasonality test is conducted by examining significant autocorrelation in the m_{th} term of the Auto Correlation Function (ACF) (Assimakopoulos *et al.*, 2020).

Given that a time series contains at least $n \geq 3m$ observations, where $m > 1$ represents the frequency, and a 90% confidence level is stipulated, the decision to apply seasonal adjustments depends on meeting the following criterion (Xia *et al.*, 2020):

$$f_t = y_{t-1} \tag{2}$$

$$|ACF_m| > 1.645\sqrt{\frac{1 + 2(ACF_1 + \sum_{i=2}^{m-1} ACF_i^2)}{n}} \tag{3}$$

Series with a frequency of one ($m = 1$) and datasets containing fewer observations than three full seasonal cycles are exempt from testing and are assumed to lack seasonality (Assimakopoulos *et al.*, 2020). Once a time series is recognized as seasonal, classical multiplicative decomposition techniques are applied to isolate its seasonal component. This process involves decomposing the series into trend and seasonal components, using methods such as moving average smoothing for trend identification and calculating seasonal indices by averaging standardized deviations across all time units. The seasonally adjusted series is then computed by dividing the original series by the seasonal indices corresponding to each respective period (Assimakopoulos *et al.*, 2020). After extrapolating the adjusted series, non-seasonal forecasts can be re-seasonalized by multiplying their values by the corresponding indices (Xia *et al.*, 2020).

Similar to the Naïve 1 method, the structural advantages of the Naïve 2 method are its simplicity and computational efficiency, providing a straightforward baseline for comparing the performance of more complex forecasting models and enabling quick computations (Xia *et al.*, 2020). As a result, the Naïve 2 model has been extensively utilized as a benchmark in historical forecasting studies. It has been incorporated in all previous M Competitions, thus facilitating straightforward comparisons (Hyndman *et al.*, 1998). For example, the Honeywell data series from the M2 Competition (1993) demonstrated a high degree of randomness and exhibited characteristics of a random walk after seasonality was identified and incorporated into the forecasts (Hyndman *et al.*, 1998). The greatest improvements over the Naïve 2 method were observed in the car and macroeconomic data, which exhibited high levels of randomness. Conversely, with the Honeywell data, alternative methods and forecasters rarely outperformed the predictions of the Naïve 2 method (Chatfield *et al.*, 1993). Additionally, a notable structural advantage of the Naïve 2 method is its optimal performance in the presence of both cyclical and seasonal patterns, making it the preferred method under such conditions (Petropoulos *et al.*, 2014).

On the other hand, since the Naïve 2 model is limited to capturing seasonality, a structural drawback is the absence of a unique method for identifying which time series are seasonal or for estimating the seasonal indices (Xia *et al.*, 2020). This limitation is also why the Naïve 2 model was not considered as the benchmark for MASE, as using it would significantly complicate the replication process (Assimakopoulos *et al.*, 2020). Despite its more complex computation, the Naïve 2 model was chosen over the Seasonal Naïve model for estimating the OWA, as the Naïve 2 model is generally more accurate than the Seasonal Naïve model.

1.1.3 Seasonal Naïve

Analogous to the Naïve 2 model, the Seasonal Naïve model is a random walk methodology particularly suitable for data exhibiting seasonal patterns (Van Horne & Parker, 1967). In this approach, forecasts are determined by the most recent observation from the corresponding period in the previous season. The long-term trend in forecasts equals the average long-term trend observed in historical data (Van Horne & Parker, 1967). The Seasonal Naïve method, while inherently an additive model, can be modified using logging and/or deflating techniques to accommodate multiplicative seasonal patterns. Mathematically, the Seasonal Naïve model is represented by the following equation (Lawler & Limic, 2010):

$$f_t = y_{t-m} \tag{4}$$

Here, m denotes the seasonal period, defined as 12 for monthly data, 4 for quarterly data, 24 for hourly data, and 1 for yearly, weekly, and daily data (Assimakopoulos *et al.*, 2020). The intuition behind the Seasonal Naïve model is that seasonal patterns are recurrent (Van Horne & Parker, 1967). Consequently, the optimal forecast for a given period is the value from the corresponding period in the preceding season. For example, in the case of yearly seasonality ($m = 1$), the most accurate forecast for January 2024 would be the value recorded in January 2023.

Like the Naïve 1 and 2 models, the Seasonal Naïve model is renowned for its simplicity and computational efficiency, requiring no additional assumptions or information, although it is more complex in computation than the Naïve 1 and 2 models (Assimakopoulos *et al.*, 2020). It achieves reasonable accuracy, especially when the data exhibits seasonality, but is typically less accurate than the Naïve 2 model (Assimakopoulos *et al.*, 2020). This influenced the M4 organizers' decision to use the Naïve 2 model for estimating the OWA instead of the Seasonal Naïve. In contrast, the Seasonal Naïve model does not require the calculation of seasonal indices, which is why it was used over the Naïve 2 model in scaling the absolute error of the examined method in the MASE, as it facilitates straightforward replication. Additionally, the Seasonal Naïve model is slow to respond to cyclical upturns or downturns, as it always relies on past observations from m periods ago, assuming that the seasonal pattern remains consistent over time. It is useful only for data with well-defined, regular seasonal patterns (Van Horne & Parker, 1967).

1.2 Exponential Smoothing Family

Exponential Smoothing (ES) techniques are extensively utilized in the field of forecasting due to their simplicity, robustness, and high accuracy, making them highly efficient for automatic forecasting procedures (Hyndman & Athanasopoulos, 2014). The concept of Exponential Smoothing (ES) was first proposed in the late 1950s, based on the original work of Brown (1959) and Holt (1957), who were developing forecasting models for inventory control systems (Holt, 1957; Brown, 1959; Winters, 1960). Known for their intuitive simplicity, ES methods employ an unequal weighting mechanism, where more recent observations are weighted more heavily than older ones in the forecasting process (Holt, 1957). This unequal weighting is facilitated by one or more smoothing parameters that control the weight given to each observation. ES methods smooth the original series in a manner similar to the moving average, and the resulting smoothed series is then used for forecasting, with recent data points having a greater impact on the forecast (Brown, 1959). The popularity of ES models in time-series analysis is due not only to their simplicity but also to their computational efficiency, ease of adjusting responsiveness to process dynamics, and reasonable accuracy (Taylor, 2003). Overall, ES methods are seen as cost-effective approaches for providing accurate forecasts in a wide array of real-world applications.

In ES models, it is assumed that time-series data can have up to three underlying components: level, trend, and seasonality. The forecasts are then formulated using the final values of these

components. Moreover, ES models are characterized by one of five trends: none, additive, damped additive, multiplicative, or damped multiplicative, and one of three seasonal components: none, additive, or multiplicative. The taxonomy of ES methods, initially proposed by Pegels (1969), was later extended and refined by Gardner Jr. and McKenzie (1985), Hyndman *et al.* (2002), Taylor (2003), and Hyndman and Athanasopoulos (2014). As a result, there are now 15 distinct ES models, leading to the notion of ES as a family of models rather than a single model.

ES models, as outlined in the introduction, play a critical role in testing **H1** and **H3**. By comparing the performance of the Comb method—defined as the simple arithmetic average of Single, Holt, and Damped ES—with the individual ES models on both simulated data and real-world data from the M4 Competition, the study evaluates the advantages of combining methods, addressing **H1**. Furthermore, the comparison between Damped ES, which incorporates a damped trend, and Holt ES, which assumes a linear trend, offers a framework for assessing the effects of trend damping in different scenarios, thereby addressing **H3**.

1.2.1 Simple Exponential Smoothing (SES)

Among the family of exponential smoothing methods, the Single Exponential Smoothing (SES) model—characterized by the absence of trend and seasonality—is the simplest and one of the most widely used approaches in all forecasting techniques (Taylor, 2003). This method is most suitable for forecasting data where no apparent trend or seasonal pattern is present (i.e., the data pattern is approximately horizontal, implying that there is no cyclic variation or pronounced trend in the historical data) (Hyndman & Athanasopoulos, 2014). Given a sequence of observed time series, $y_1, y_2, \ldots, y_n$, the equation for SES is formally expressed as follows (Brown, 1959):

$$\hat{y}_{i+1} = \alpha y_i + (1 - \alpha)\hat{y}_i \tag{5}$$

where y_i is the observed value of the series at time period i, $\hat{y}_i$ represents the forecasted value of y for time period i, $\hat{y}_{i+1}$ denotes the forecast for time period $i+1$, and α is the smoothing coefficient (Brown, 1959). Invented in the 1950s by Robert G. Brown, the underlying intuition of SES is that recent observations are more indicative of future values than older ones. $\hat{y}_{i+1}$ is derived by assigning a weight of α to the most recent observation y_i and a weight of $1 - \alpha$ to the most recent forecast $\hat{y}_i$.

The smoothing parameter, α, which lies between zero and one ($0 < \alpha < 1$), determines the degree of smoothing applied to the series (Winters, 1960). When $\alpha = 1$, the model assigns full weight to the most recent observation, resulting in the original and smoothed series being identical. Consequently, no smoothing is applied. On the other hand, when $\alpha = 0$, the model assigns no weight to the most recent observation, leading to a completely flat smoothed series, representing the extreme form of smoothing (Holt, 1957). To derive the forecast $\hat{y}_{i+1}$, the sequence of weights is given by $\alpha, \alpha(1-\alpha), \alpha(1-\alpha)^2, \ldots$. These weights reduce exponentially until they approach zero. As a result, the further back we go in the series, the smaller the weight of each value, thereby diminishing its impact on the forecast (Brown, 1959). Ultimately, the weights diminish towards zero.

The simplicity and effectiveness of the SES model are attributed to its method of updating forecasts based on past errors. The fundamental equation for SES is given by Brown (1959) as follows:

$$\hat{y}_{i+1} - \hat{y}_i = \alpha(y_i - \hat{y}_i) \tag{6}$$

The equation indicates that the change in the forecast value ($\hat{y}_{i+1} - \hat{y}_i$) is proportional to the forecast error ($y_i - \hat{y}_i$), with α serving as the proportionality constant. This demonstrates a key

principle of SES: the forecast is adjusted based on the size of the previous error, moderated by α (Winters, 1960). Reformulating the equation yields the following expression:

$$\hat{y}_{i+1} = \hat{y}_i + \alpha\epsilon_i \tag{7}$$

where the residual $\epsilon_t = y_i - \hat{y}_i$ denotes the forecast error for the time period i. Hence, the exponential smoothing forecast is computed as the previous forecast adjusted by the error encountered in the latest forecast. By iteratively substituting prior forecast values back into the model, the following expression is derived (Brown, 1959):

$$\begin{aligned}
\hat{y}_{i+1} &= \alpha y_i + (1-\alpha)[\alpha y_{i-1} + (1-\alpha)\hat{y}_{i-1}] \\
&= \alpha y_i + \alpha(1-\alpha)y_{i-1} + (1-\alpha)^2\hat{y}_{i-1}, \\
\hat{y}_{i+1} &= \alpha y_i + \alpha(1-\alpha)y_{i-1} + \alpha(1-\alpha)^2 y_{i-2} + (1-\alpha)^3\hat{y}_{i-2}, \\
\hat{y}_{i+1} &= \alpha y_i + \alpha(1-\alpha)y_{i-1} + \alpha(1-\alpha)^2 y_{i-2} + \alpha(1-\alpha)^3 y_{i-3} + (1-\alpha)^4\hat{y}_{i-3}, \\
&\vdots
\end{aligned} \tag{8}$$

The general expression for the forecast equation is (Brown, 1959):

$$\begin{aligned}
\hat{y}_{i+1} &= \alpha y_i + \alpha(1-\alpha)y_{i-1} + \alpha(1-\alpha)^2 y_{i-2} + \cdots + \alpha(1-\alpha)^{i-2} y_2 + \alpha(1-\alpha)^{i-1} y_1 \\
&= \alpha \sum_{k=0}^{i-1} (1-\alpha)^k y_{i-k}
\end{aligned} \tag{9}$$

$\hat{y}_{i+1}$ reflects the forecast value of the variable y at time period $i+1$, constructed from the actual series values $y_i, y_{i-1}, y_{i-2}, \ldots$ back to the earliest known value of the time series, y_1. As a result, $\hat{y}_{i+1}$ represents the weighted moving average of all previous observations.

Although the SES model is widely utilized and has proven successful across various fields, it possesses certain limitations that may impact its forecasting accuracy. For instance, there is no consensus regarding the selection of initial values and the determination of the optimal smoothing parameter, α (Hyndman & Athanasopoulos, 2014). Given that $\hat{y}_1$ is unknown, Hyndman and Athanasopoulos (2014) proposed two alternatives for the initialization of the SES model. One approach involves utilizing the initial values and optimized smoothing parameter with ETS, commonly referred to as the 'optimal' method. The alternative approach entails deriving the initial values from the first few observations, also known as the 'simple' method.

Overall, the SES model is esteemed for its simplicity, computational efficiency, reasonable accuracy, and robustness in capturing the level (Taylor, 2003). However, factors such as the lack of consensus on the optimal value of the smoothing parameter and initialization, as well as its inability to handle data with trends or seasonality, can significantly impact its forecasting efficiency.

1.2.2 Holt's Linear Trend Method

As mentioned in the introduction to the Exponential Smoothing family, the trend in an ES model can be represented by one of five forms: none, additive, damped additive, multiplicative, or damped multiplicative (Hyndman & Athanasopoulos, 2014). In 1957, Holt expanded the SES model to enable the forecasting of data assuming a linear trend, also known as "Holt's linear trend method." This technique incorporates a forecasting equation as well as two smoothing equations, one for the level and another for the trend (Holt, 1957):

Forecast equation:

$$\hat{y}_{t+h|t} = \ell_t + hb_t \tag{10}$$

Level equation:

$$\ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1}) \tag{11}$$

Trend equation:

$$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1} \tag{12}$$

Within this framework, ℓ_t denotes an estimate of the level of the series at time t, and b_t denotes an estimate of the trend (slope) of the series at time t. The level smoothing parameter α ranges from 0 to 1 ($0 \leq \alpha \leq 1$), and the trend smoothing parameter β^* also ranges from 0 to 1 ($0 \leq \beta^* \leq 1$) (Holt, 1957). Analogous to the SES model, the level equation indicates that ℓ_t is the weighted average of the present observation y_t and the one-step-ahead forecast at time t, denoted as $\ell_{t-1} + b_{t-1}$ (Holt, 1957). The trend equation specifies that b_t is calculated as a weighted average of the trend estimate at time t, based on $\ell_t - \ell_{t-1}$ and the preceding trend estimate b_{t-1}. As a result, the forecasting function has transitioned from being flat to exhibiting a trend (Holt, 1957). Consequently, the h-step-ahead forecast is determined by adding the last estimated level to the product of h and the last estimated trend value. Therefore, these forecasts are linearly dependent on h.

Holt's method is extensively employed in time series forecasting due to its simplicity, computational efficiency, robust performance, relatively high accuracy, and its capability to effectively model linear trends (Gardner & McKenzie, 1985). However, since the Holt ES method assumes a linear trend, it is not well-suited for data with complex, non-linear trends (Gardner & McKenzie, 1985). Additionally, similar to the SES model, there is no consensus on the initial values and the optimal smoothing parameters (Hyndman & Athanasopoulos, 2014). These factors significantly impact the forecasting accuracy of the Holt Exponential Smoothing model.

1.2.3 Damped Additive Trend Method

Forecasts generated by Holt's linear method exhibit a constant trend, whether increasing or decreasing, indefinitely into the future. In their 1982 study, Makridakis *et al.* evaluated the post-sample accuracy of 21 automatic forecasting methods on 1,001 time series and found that linear trend methods frequently over-forecast, especially over longer forecasting horizons. Motivated by this finding, Gardner and McKenzie, in their series of three papers (1985, 1988, 1989), developed advanced versions of the Holt-Winters methods (Holt, 2004; Winters, 1960) of exponential smoothing, designed to dampen the trend as the forecast horizon extends. The Damped ES method has two smoothing parameters, α and β, along with a damping parameter ϕ, with all values between 0 and 1, and is given by the following equations:

$$
\begin{aligned}
\text{Forecast equation:} \quad & \hat{y}_{t+h|t} = \ell_t + (\phi + \phi^2 + \cdots + \phi^h) b_t \\
\text{Observation equation:} \quad & y_t = \ell_{t-1} + \phi b_{t-1} + \epsilon_t \\
\text{State equations:} \quad & \ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha \epsilon_t \\
& b_t = \phi b_{t-1} + \beta \epsilon_t
\end{aligned} \tag{13}
$$

In the case where $\phi = 1$, this method becomes equivalent to Holt's linear method. When $0 < \phi < 1$, the damping parameter ϕ mitigates the trend, causing it to converge to a constant value over time. It is observed that the forecasts approach $\ell_T + \frac{\phi b_T}{1-\phi}$ as $h \to \infty$ for all $0 < \phi < 1$. This implies that short-term forecasts are trended, whereas long-term forecasts stabilize at a constant value. The Naïve 2 method is utilized for making seasonal adjustments.

Ever since the publication of the papers by Gardner & McKenzie, numerous empirical studies have shown that damped ES yields favorable results (Gardner, 2006). For instance, in his review of evidence-based forecasting, Armstrong (2006) endorsed the damped ES as a well-established method that enhances accuracy in real-world applications. Moreover, Fildes, Nikolopoulos, Crone, and Syntetos (2008), in a review of forecasting in operational research, asserted that the damped ES could be considered "a benchmark forecasting method for all others to beat." In addition, Hyndman, Koehler, Ord, and Snyder (2008) present evidence from the M3 competition data (Makridakis & Hibon, 2000), demonstrating that the damped ES method performs favorably based on selection criteria. Just like the SES and Holt ES models, the Damped ES is known for its simplicity, computational efficiency, and increased flexibility in modeling trends compared to the Holt ES model.

Despite its proven track record in forecasting accuracy, a strong rationale for the damped ES method has yet to be presented (McKenzie, 2010). Moreover, like the SES and Holt ES models, the Damped ES model is highly sensitive to initial values, as well as to the smoothing and damping parameters. The impact of these factors will be evaluated in the simulation study in Part II and in the analysis of real-world data from the M4 Competition in Part III.

1.2.4 Comb

In the M4 Competition, the Comb method—defined as the simple arithmetic average of Single, Holt, and Damped exponential smoothing methods, as described in the preceding subsections—served as the primary benchmark (Assimakopoulos *et al.,* 2020). It was the most accurate method in the M2 Competition, hosted in 1993 by Makridakis *et al.* (1993). The M2 Competition utilized a limited dataset of 29 series, yet it integrated significantly richer contextual information and was conducted in real-time. Due to the constrained sample size and the incorporation of supplementary data, it was challenging to draw broad generalizations regarding time series forecasting methodologies. However, the key finding indicated that combining various forecasting approaches, including both ML and statistical methods, often led to enhanced accuracy compared to relying on a single technique (Hyndman, 2019; Makridakis *et al.*, 1993).

The Comb method once again demonstrated the efficacy of combining approaches in the more advanced M4 Competition, which utilized 100,000 time series. In this context, the sMAPE, MASE, and OWA of the Comb method were consistently lower than those of the individual Single, Holt, and Damped exponential smoothing methods it combined (Assimakopoulos *et al.,* 2020). Indeed, the M4 Competition, in conjunction with the M2 and M3 Competitions, once more demonstrated the enhanced accuracy achieved through the combination of statistical and/or machine learning methods (Assimakopoulos *et al.,* 2020).

$$
y_t^{\text{Comb}} = \frac{1}{3}\left(y_t^{\text{SES}} + y_t^{\text{Holt}} + y_t^{\text{Damped}}\right) \tag{14}
$$

Where:

- y_t^{Comb} represents the combined forecast value at time t.
- y_t^{SES} represents the forecast value at time t using Single Exponential Smoothing.
- y_t^{Holt} represents the forecast value at time t using Holt's linear method.
- y_t^{Damped} represents the forecast value at time t using Damped Exponential Smoothing.

1.3 Theta Methods

Theta methods are widely recognized as robust and versatile forecasting approaches, known for their simplicity and competitive performance in large-scale forecasting competitions like the M3 and M4 (Dudek, 2019). Introduced by Assimakopoulos and Nikolopoulos (2000), Theta methods are univariate forecasting models designed to improve accuracy by decomposing the original time series into multiple lines, referred to as Theta lines and denoted by $Z_t(\theta)$ (Assimakopoulos *et al.*, 2020). Each Theta line is individually extrapolated using a forecasting model of choice, and the resulting predictions are merged to form the final forecast (Dudek, 2019).

The estimation of Theta lines relies on modifying the second differences of the time series using a Theta coefficient (θ), where $\theta \in \mathbb{R}$. This transformation preserves the mean and slope of the original series while adjusting its curvature and local variability (Fiorucci *et al.*, 2016; Assimakopoulos *et al.*, 2020). Depending on the value of θ, the curvature is either smoothed or enhanced. For example, $\theta < 1$ reduces local curvature, emphasizing long-term trends, while $\theta = 0$ results in a straight line, and $\theta > 1$ amplifies short-term dynamics (Assimakopoulos & Nikolopoulos, 2000).

Over the years, several extensions of the Classical Theta method have emerged. For instance, Nikolopoulos and Assimakopoulos (2005) and Petropoulos and Nikolopoulos (2013) proposed expanding the range of θ values to $\{-1, 0, 1, 2, 3\}$, which could improve forecasting accuracy by extracting additional information from the time series. Other advancements include the use of unequal weights in combining forecasts (Constantinidou *et al.*, 2012; Petropoulos & Nikolopoulos, 2013), non-linear trend modeling (Fiorucci, 2016), and a multiplicative formulation for combining level, trend, and seasonality components (Assimakopoulos *et al.*, 2020). Additionally, Legaki and Koutsouri (2020) enhanced the method's robustness by integrating the Box-Cox transformation to handle non-linearity.

Despite these developments, literature includes only a limited number of generalizations of the Classical Theta method (Fiorucci, 2016). This thesis aims to address this gap by introducing the Theta-MLP method, which extends the Classical Theta method and leverages machine learning to further enhance forecasting accuracy. By comparing the performance of the complex Theta-MLP method against the simpler Classical Theta method, the study aims to assess the impact of model complexity on forecasting accuracy, addressing **H2**.

1.3.1 Classical Theta Method

In its original and simplest form, known as the Classical Theta method, the model decomposes the original time series into two Theta lines with ad hoc values for the θ parameters: $\theta_1 = 0$ and $\theta_2 = 2$ (Assimakopoulos *et al.*, 2020). These lines are extrapolated using simple linear regression and SES and then combined with equal weights (50%-50%) (Assimakopoulos *et al.*, 2020). Thanks to its simplicity, efficiency, and ease of parameterization, the Classical Theta method employs an additive double-lined model, as expressed in the following equation (Hyndman & Billah, 2003; Assimakopoulos *et al.*, 2020):

$$y_t = \omega Z_t(\theta_1) + (1 - \omega) Z_t(\theta_2); \tag{15}$$

$$\omega = \frac{\theta_2 - 1}{\theta_2 - \theta_1}, \tag{16}$$

Here, w and $(1 - w)$ denote the weights of the two Theta lines, which sum to one, subject to the restrictions $\theta_1 < 1$ and $\theta_2 \geq 1$. In double-lined Theta models, the first line ($\theta_1 = 0$) is fixed and predefined as the trend. As a result, θ_2 is simply referred to as the θ coefficient, and the additive expression simplifies as follows (Hyndman & Billah, 2003; Assimakopoulos & Nikolopoulos, 2000):

$$y_t = \frac{\theta - 1}{\theta} Z_t(0) + \frac{1}{\theta} Z_t(\theta), \tag{17}$$

The Theta line of coefficient θ, $Z_t(\theta)$, is obtained as the solution to the equation outlined by Assimakopoulos and Nikolopoulos (2000) and further discussed by Spiliotis, Assimakopoulos, and Makridakis (2020):

$$\nabla^2 Z_t(\theta) = \theta \nabla^2 y_t = \theta \left(y_t - 2y_{t-1} + y_{t+2} \right), \quad \text{at time } t = 3, 4, \ldots, n \tag{18}$$

Where $y_1, \ldots, y_n$ denote the original time series. All Theta methods are applied to non-seasonal or deseasonalized time series, if applicable. In the case of the Classical Theta method, deseasonalization is performed using multiplicative classical decomposition by moving averages (Fiorucci, 2016). This implies that, for the Classical Theta method, the seasonal component—if present—is subsequently multiplied with the trend and residual components, indicating a multiplicative relationship between the seasonal component and the trend and level components (Fiorucci, 2016). Furthermore, $\nabla^2 y_t$ represents the second difference of y_t (i.e., $\nabla^2 y_t = y_t - 2y_{t-1} + y_{t-2}$). Hyndman & Billah (2003) provide the analytical solution to the additive Theta line $Z_t(\theta)$, expressed by the following equation:

$$Z_t(\theta) = \theta y_t + (1 - \theta) Z_t(0), \quad t = 1, \ldots, n \tag{19}$$

In the case of the Classical Theta method, $Z_t(0)$ is defined as a linear trend, represented by the following expression (Dudek, 2019):

$$Z_t(0) = A_n + B_n t; \tag{20}$$

Where the constants A_n and B_n are optimized through the minimization process, which involves reducing the sum of the squared differences, formulated as $\sum_{i=1}^{n} [y_i - Z_i(\theta)]^2 = \sum_{i=1}^{n} [(1 - \theta) y_i - A_n - B_n t]^2$ (Dudek, 2019):

$$A_n = \frac{1}{n} \sum_{t=1}^{n} y_t - \frac{n+1}{2} B_n; \tag{21}$$

$$B_n = \frac{6}{n^2 - 1} \left(\frac{2}{n} \sum_{t=1}^{n} t y_t - \frac{1 + n}{n} \sum_{t=1}^{n} y_t \right). \tag{22}$$

It should be emphasized that the parameters A_n and B_n are solely dependent on the initial data. The Theta line, as described in the analytical solution, is essentially a linear regression model implemented directly on this dataset (Assimakopoulos & Nikolopoulos, 2000). Filling in $\theta = 2$, the expression simplifies to:

$$y_t = 0.5Z_t(0) + 0.5Z_t(2) \tag{23}$$

Where, as mentioned earlier, $Z_t(0)$ is extrapolated using a linear regression (LR) model, and $Z_t(2)$ is estimated using simple exponential smoothing (SES). The final h-step-ahead forecast is then the simple combination of the $Z_t(0)$ and $Z_t(2)$ forecasts (Fiorucci, 2016):

$$\hat{y}_{t+h} = 0.5\hat{Z}_{t+h}(0) + 0.5\hat{Z}_{t+h}(2) \tag{24}$$

The forecasting procedure for the Classical Theta method, as delineated by Assimakopoulos & Nikolopoulos (2000), encompasses the following steps:

Step 0. Seasonality Test: The initial step entails testing the time series for statistically significant seasonal behavior using an autocorrelation function (Fiorucci *et al.*, 2016). A time series exhibits seasonality if:

$$|r_m| > q_{1-a/2}\sqrt{\frac{1 + 2\sum_{i=1}^{m-1} r_i^2}{n}}, \tag{25}$$

Where r_k denotes the autocorrelation function at lag k, while m represents the number of periods within a seasonal cycle, such as 12 for monthly data. The variable n refers to the sample size, and q is the quantile function of the standard normal distribution (Fiorucci *et al.*, 2015). The expression $(1 - \alpha)\%$ indicates the confidence level (Assimakopoulos & Nikolopoulos, 2000). The criterion utilized was the t-test statistic for the autocorrelation function with a one-year lag, which translates to 12 observations for monthly time series and 4 observations for quarterly time series. Assimakopoulos and Nikolopoulos (2000) selected a 90% confidence level, corresponding to a critical t-statistic value of 1.645.

Step 1. Deseasonalization: If the series has a statistically significant seasonal component, it is deseasonalized using the classical multiplicative decomposition method.

Step 2. Decompositoin: In the Classical Theta method, the time series is decomposed into two Theta lines, namely the linear regression line $Z_t(0)$ and the Theta line for $Z_t(2)$.

Step 3. Extrapolation: Subsequently, $Z_t(0)$ is extrapolated using a normal linear regression line, whereas $Z_t(2)$ is extrapolated using SES.

Step 4. Combination: In the Classical Theta method, the final forecasts is an equally weighted combination of the two Theta lines.

Step 5. Reseasonalisation: If in the initial step the series was identified as seasonal, then the final forecasts are subsequently multiplied times their seasonal indices.

The simplicity of the Classical Theta method represents a significant structural advantage (Dudek, 2019). Hyndman & Billah (2003) demonstrated that, albeit under specific conditions as identified by Nikolopoulos *et al.* (2012), the Classical Theta method can be further simplified. This is possible given that the level of the series resulting from the extrapolation of $Z(2)$ is adjusted by half the slope of $Z(0)$ (Hyndman & Billah, 2003). Such simplification establishes a relationship with the SES with drift model.

Another structural advantage of the Classical Theta method lies in its ease of parameterization, making the Classical Theta method computationally efficient as a result of minimal parameter tuning and simplicity (Spiliotis *et al.*, 2020). The Classical Theta method provides an additional structural advantage through its straightforward parameterization, specifically by fixing $\theta = 0$ and $\theta = 2$. This approach enhances computational efficiency by reducing the need for extensive parameter tuning and preserving simplicity (Spiliotis *et al.*, 2020).

Moreover, the decomposition process benefits significantly from amplifying both short-term ($\theta_2 = 2$) and long-term ($\theta_1 = 0$) patterns in the data, enabling the extraction of information often overlooked in simple extrapolations of the original time series (Fiorucci *et al.*, 2016). By adjusting the local curvatures through the value of θ, the Classical Theta method can identify complex patterns within the data. This method allows for the effective extrapolation of each pattern individually, followed by the combination of their results to enhance forecasting performance (Claeskens *et al.*, 2016). The strength of the Theta Method lies in its 'divide and conquer' approach, as no single forecasting model can capture the full spectrum of possible time series patterns. When the series is decomposed into several lines containing less information, even traditional models can experience gains in forecasting accuracy (Assimakopoulos *et al.*, 2020). While the Classical Theta method is suitable for handling data with both trend and seasonality, Thomakos and Nikolopoulos (2014) have demonstrated that it is particularly effective for trended data (Fiorucci *et al.*, 2016).

Furthermore, the Classical Theta method is seen as an extension to the concept of combining, which has been shown to improve forecasting accuracy under certain circumstances and is generally considered a useful practise in the field of forecasting (Clemen, 1989; Assimakopoulos & Nikolopoulos, 2000: Fiorucci *et al*, 2015). The reasoning lies in the averaging of errors produced by each individual forecasting method, which inherently reduces the impact of any single method's bias (Assimakopoulos & Nikolopoulos, 2000).

Altogether, the accuracy and robustness of the Classical Theta method lie in its decomposition process and the application of combining forecasts (Fiorucci *et al.*, 2015; Fiorucci *et al.*, 2016). The accuracy of the Classical Theta method was shown in the M3 Competition, where it exhibited superior performance compared to advanced methods and expert systems (Dudek, 2019). It notably outperformed its competitors, particularly in the realms of monthly series and microeconomic data (Makridakis & Hibon, 2000). Subsequent empirical studies have also affirmed the robust performance of the Classical Theta method, notably by Assimakopoulos *et al.* (2012) and Petropoulos & Nikolopoulos (2013).

Conversely, Theta Methods are by definition characterized by their dynamic nature, which enables the selection of diverse Theta lines and the application of weights that may be either equal or unequal, depending on the specific requirements of the analysis (Fiorucci *et al.*, 2016; Fiorucci *et al.*, 2015). Nevertheless, Assimakopoulos & Nikolopoulos (2000) have restricted this important property by fixing the theta coefficients to predefined values, resulting in a lack of flexibility (Fiorucci *et al.*, 2015). This simplified version, known as the Classical Theta method, uses only two Theta lines: the linear regression (LR) model for the Theta line corresponding to $\theta_1 = 0$ and the SES model for the Theta line corresponding to $\theta_2 = 2$ (Fiorucci *et al.*, 2016). Therefore, the Classical Theta method, as implemented in the M3 Competition, is limited by its focus on specific information within the dataset (Fiorucci *et al.*, 2015). If optimization were applied in the selection of the theta lines, the method would be better equipped to focus on the most important information (Fiorucci *et al.*, 2016).

For instance, Nikolopoulos & Assimakopoulos (2005) and Petropoulos & Nikolopoulos (2013) proposed that employing additional Theta lines, where $\theta \in \{-1, 0, 1, 2, 3\}$, could potentially extract more relevant information from the original data. Furthermore, Constantinidou *et al.* (2012) and Petropoulos & Nikolopoulos (2013) proposed employing unequal weights in the recomposition process of final forecasts, arguing that asymmetric weights, directly associated with forecast horizons, are more likely to yield a more accurate approximation of both short- and long-term components (Fiorucci *et al.*, 2015).

A further limitation of the Classical Theta method is its inability to accurately forecast non-linear trends, such as exponential trends, because it drifts SES forecasts by a constant value of $B_n\left(\frac{\theta-1}{\theta}\right)$ at each point (Assimakopoulos *et al.*, 2019). The consequence of this limitation is potentially poor forecasting accuracy, particularly for long-term forecasts, as the trend component becomes dominant (Assimakopoulos *et al.*, 2020).

Another drawback of the Classical Theta method is the additive combination of the trend and level components, $Z_t(0)$ and $Z_t(\theta)$, respectively (Assimakopoulos *et al.*, 2020). The seasonal component, if present, is combined multiplicatively with the other components. However, time series components do not always adhere to strictly additive or multiplicative relationships (Hyndman *et al.*, 2002).

1.4 Theta-MLP: A Hybrid Method of Theta and Multilayer Perceptrons for Time Series Forecasting

The M4 Competition highlighted the potential of complex methods to significantly enhance forecasting accuracy, notably with the ES-RNN method developed by Slawek Smyl at Uber Technologies. This method combines exponential smoothing (ES) formulas with a recurrent neural network (RNN) forecasting engine (Assimakopoulos *et al.*, 2020). Smyl's method, the top-performing submission in the competition, integrates statistical and machine learning components through a unified gradient descent approach, yielding superior point forecasts (PFs) and prediction intervals (PIs) (Smyl, 2018).

Inspired by the success of the ES-RNN and the findings of the M4 Competition, this study introduces the Theta-MLP method, a novel hybrid approach for time series forecasting. Building on refinements to the Classical Theta method by Legaki and Koutsouri (2020), Fiorucci *et al.* (2015), Assimakopoulos *et al.* (2000, 2020), Constantinidou *et al.* (2001), and Fiorucci (2016), the Theta-MLP method incorporates: (i) a Box-Cox transformation, (ii) both linear and non-linear trend components, (iii) additive and multiplicative Theta lines, (iv) additive and multiplicative seasonality, and (v) optimization of θ using the in-sample Mean Absolute Error (MAE) rather than predefined values, allowing the slope of the trend to either be damped or expanded when needed. These enhancements result in a topology of eight suitable candidate models, from which the best-performing model is selected for forecasting. Additionally, the Theta-MLP method integrates a Multi-Layer Perceptron (MLP) component, regulated by the parameter ϕ, to introduce an error correction term that captures residual non-linearities beyond the scope of the selected Theta model.

The procedure begins with a Box-Cox transformation to stabilize variance and address non-linearities, followed by an assessment of negative values and seasonality to exclude incompatible candidate models. The parameter θ is optimized using the Mean Absolute Error (MAE), and the best-performing Theta model is selected based on in-sample accuracy. Residuals from the final Theta model are then used to train the MLP, which learns and forecasts residual patterns. The MLP forecasts are combined with the Theta model's forecasts, scaled by the optimized parameter ϕ to balance their contributions. Finally, the combined forecasts are inverse-transformed back to the original scale using the inverse Box-Cox transformation.

This section first explores the extensions of the Classical Theta method, as originally proposed by Assimakopoulos and Nikolopoulos (2000) and later refined by subsequent studies. It then examines the architecture and training process of the Multi-Layer Perceptron (MLP), with a focus on how the MLP captures non-linear patterns in the residuals that are not addressed by the Classical Theta method. Next, the section outlines the procedure for applying the Theta-MLP method, including the steps for model selection, residual correction through the MLP, and the final steps of inverse Box-Cox transformation and reseasonalization. The section concludes with a discussion of the structural advantages and limitations of the Theta-MLP method.

1.4.1 Generalizing the Classical Theta Method

The Classical Theta method is renowned for its simplicity, robustness, and accuracy in time series forecasting. Its ability to generate reliable forecasts across diverse data types and frequencies has established it as a benchmark in forecasting competitions, most notably the M3 Competition, where it demonstrated exceptional performance (Constantinidou *et al.*, 2001). Nevertheless, the Classical Theta method has a few notable shortcomings, as previously mentioned: (i) assuming a fixed value for θ, (ii) assuming a linear trend, (iii) assuming an additive relationship between $Z_t(0)$ and $Z_t(2)$, and (iv) relying on multiplicative seasonality.

The first issue, namely the pre-fixed θ value, relates to the fact that the Classical Theta method generates forecasts that drift according to the coefficient B_n. This drift is also a function of θ, calculated as $(\theta - 1)/\theta$ times B_n (Fiorucci, 2016). The parameter θ is the only value that needs to be specified to improve forecasting accuracy across multiple and diverse time series, and if θ is not appropriately selected and the long-term trend continues into the future, the model may produce forecasts that are either overly pessimistic or overly optimistic, depending on the direction of the trend (Assimakopoulos *et al.*, 2020). Therefore, by fixing $\theta = 2$, the Classical Theta method cannot dynamically dampen or expand the trend to suit the specific characteristics of the data.

One possible solution to this problem was introduced by Fiorucci *et al.* (2015), which suggested selecting the optimal θ value using a loss function based on prediction errors over a validation sample, minimized for each series. Following this, the proposed Theta-MLP method determines the θ parameter by minimizing the in-sample (one-step ahead) mean absolute error (MAE) of the model, allowing the slope of the trend to either be damped or expanded, as defined by:

$$MAE = \frac{1}{n}\sum_{n=1}^{n}|y_t - \hat{y}_t|, \tag{26}$$

Where $\hat{y}_t$ denotes the forecasts generated by the model, the minimization is performed using the Brent method, which combines golden section search with successive parabolic interpolation, enabling rapid convergence to a reliable solution (Armstrong *et al.*, 2015). To reduce computational costs, the solution space is constrained to $0 \leq \theta \leq 3$, following a procedure analogous to that of Assimakopoulos *et al.* (2020). By doing so, the forecasts are also ensured not to overestimate the trend, as Armstrong, Green, and Graefe (2015) demonstrated that damping trends and adopting a conservative approach is considered best practice for long-term forecasting.

However, even if the optimal value of θ is successfully determined, the Theta method may still produce inaccurate forecasts for time series with non-linear trends, such as exponential trends (Spiliotis *et al.*, 2019a). This limitation arises from the Classical Theta method's assumption of a linear trend, which can result in poor forecasting accuracy, especially when the trend component becomes particularly dominant.

One solution to this issue was proposed by Legaki & Koutsouri (2020), who suggested applying the Box-Cox transformation to the Classical Theta method. The Box-Cox transformation, introduced

by George Box and David Cox in 1964, is a statistical method designed to stabilize variance and improve adherence to a normal distribution (Saskia, 1992; Osborne, 2010). By extending the concept of power transformations, the Box-Cox method provides a flexible approach to normalize data, offering options beyond traditional transformations such as square root, logarithmic, and inverse methods. This flexibility allows it to address variables exhibiting positive or negative skewness more effectively.

As a result, the Theta-MLP method incorporates the Box-Cox transformation as the first step, which involves transforming the historical data using the parameter λ, which controls the degree of transformation. The optimal value of λ, typically chosen within the range $\lambda \in [-5, 5]$, is identified as the value that best approximates the data to a normal distribution. The transformation is defined as follows:

$$w_t = \begin{cases} \log(y_t) & \text{if } \lambda = 0; \\ \frac{y_t^{\lambda}-1}{\lambda} & \text{otherwise.} \end{cases} \tag{27}$$

Here, y_t represents the original data, and when $\lambda = 0$, the natural logarithm is applied instead. After generating forecasts using the transformed data, the inverse Box-Cox transformation is applied to rescale the data back to its original scale. This step is defined as:

$$y_t = \begin{cases} \exp(w_t) & \lambda = 0; \\ (\lambda w_t + 1)^{1/\lambda} & \text{otherwise.} \end{cases} \tag{28}$$

By integrating the Box-Cox transformation into the Theta-MLP method, this approach enhances the model's ability to handle non-linear trends, improving its forecasting accuracy for time series with dominant trend components.

Another solution to the Classical Theta method's limitations with non-linear trends was proposed by Assimakopoulos & Makridakis (2020), who suggested incorporating both linear and non-linear trends. Building on this idea, the Theta-MLP method extends the Classical Theta method by constructing models that account for both linear and exponential trends:

$$\text{Linear trend:} \quad Z_t(0) = A_n + B_n t; \tag{29}$$

$$\text{Exponential trend:} \quad X_t(0) = a_n e^{b_n t}, \quad \text{or} \quad \log(X_t(0)) = \log(a_n) + b_n t. \tag{30}$$

Where a_n and b_n are estimated by performing a simple linear regression of $\log(y_1), \ldots, \log(y_n)$ against $1, \ldots, n$.

Another shortcoming of the Classical Theta method, as mentioned earlier, is the assumption of an additive connection between the trend component $Z_t(0)$ and the level component $Z_t(\theta)$ by averaging the forecasts produced for $Z_t(0)$ and $Z_t(\theta)$ with equal weights. Furthermore, it is also assumed that the seasonal components interact multiplicatively with the rest of the components, especially if the original series is seasonally adjusted, if needed, using the classical multiplicative decomposition with moving averages. Subsequently, the Classical Theta method can be further extended by incorporating a multiplicative expression of the Theta lines, alongside the existing additive approach, as outlined by Assimakopoulos *et al.* (2020) in the following manner:

$$\text{Additive Theta line:} \quad A_t(\theta) = \theta y_t + (1-\theta)U_t(0); \tag{31}$$

$$\text{Multiplicative Theta line:} \quad M_t(\theta) = \frac{y_t^{\theta}}{U_t(0)^{\theta-1}}, \tag{32}$$

Here, $U(0)$ denotes the trend curve, which, as shown earlier, can be extended in the Classical Theta method to include both a linear trend and an exponential trend. Consequently, the Classical Theta method can be expanded to encompass both additive and multiplicative expressions, as illustrated in the following equations:

$$\text{Additive expression:} \quad y_t = \frac{\theta-1}{\theta}U_t(0) + \frac{1}{\theta}A_t(\theta); \tag{33}$$

$$\text{Multiplicative expression:} \quad y_t = \sqrt[\theta]{U_t(0)}\,\sqrt[\theta]{M_t(\theta)}, \tag{34}$$

Here, $U(0)$, $A(0)$, and $M(0)$ are extrapolated separately. $U(0)$ can either be $Z(0)$ or $X(0)$, which are extrapolated using equations (29) and (30), respectively. $A(0)$ and $M(0)$ are extrapolated using SES.

A total topology of twelve models ($2\times2\times3 = 12$) is derived by considering (i) linear and exponential trends, (ii) additive and multiplicative Theta lines, and (iii) none, additive, and multiplicative seasonality, building on the generalizations of the Classical Theta model. In this topology, each model is denoted by a seasonality type, an expression type, and a trend type. The first letter represents the seasonality type ('N', 'M', 'A'), the second letter denotes the expression ('M', 'A'), and the third letter indicates the trend type ('L', 'E'). Across all cases, 'A' stands for additive, 'M' for multiplicative, 'N' for none, 'L' for linear, and 'E' for exponential. For instance, the Classical Theta model is denoted as MAL, as it assumes multiplicative seasonality, an additive Theta line, and a linear trend.

Out of the twelve theoretically possible Theta models, four are excluded a priori due to issues related to computational stability. Multiplicatively expressed models are inherently limited to datasets with strictly positive values, making them applicable primarily in contexts where negative forecasts are not meaningful (Assimakopoulos & Makridakis, 2020). This restriction is particularly critical when additive seasonal adjustments or the estimation of linear trends yield values that approach or reach zero. Under such conditions, certain multiplicative models become unstable or computationally infeasible. Consequently, the models NML, AML, AME, MML are excluded from consideration, as their application could lead to unreliable or undefined results. By eliminating these configurations, the Theta-MLP method operates with a reduced set of eight candidate models:

Table 1: Topology of Candidate Theta Models: Selection Based on Lowest In-Sample MAE for Forecasting in the Theta-MLP Method M (Multiplicative), A (Additive), N (None), L (Linear), E (Exponential).

Model	**Seasonality**	**Expression**	**Trend**
MAL	M	A	L
MAE	M	A	E
MME	M	M	E
AAL	A	A	L
AAE	A	A	E
NAL	N	A	L
NAE	N	A	E
NME	N	M	E

As previously discussed, the Theta-MLP method begins by applying the Box-Cox transformation to the original time series to better handle non-linear trends. Next, the method tests for negative values in the time series. If negative values are present, the remaining multiplicatively expressed Theta models are excluded to ensure numerical stability and prevent forecasting errors. The Theta-MLP method then performs an autocorrelation function (ACF) test to assess the presence of seasonality. If a statistically significant seasonal pattern is identified, models that do not incorporate seasonality are excluded.

Following these preprocessing steps, a pool of candidate models remains. For each Theta method, the θ parameter is optimized using the in-sample (one-step-ahead) Mean Absolute Error (MAE). The Theta-MLP method then selects the final Theta model for forecasting based on the minimization of the in-sample MAE. The trend, which can either be $Z(0)$ or $X(0)$, is extrapolated using equations (29) or (30), whereas the Theta lines $A(0)$ and $M(0)$ are extrapolated using SES. Finally, the method applies the MLP correction term to the selected Theta forecasts, as will be detailed in the subsequent section.

1.4.2 Multilayer Perceptrons (MLP) Correction Term

After generating the initial forecasts from the selected Theta method, the remaining forecast errors (residuals) are analyzed to capture non-linear patterns that the Theta method may have missed. These residuals are calculated as:

$$\varepsilon_t = y_t - \widehat{y}_{t,\theta} \tag{35}$$

where y_t represents the observed value at time t, and $\widehat{y}_{t,\theta}$ denotes the forecasted value from the Theta method. The residuals, ε_t, reflect underlying data structures not accounted for by the Theta model. To model and predict these residuals, a **Multilayer Perceptron (MLP)** is employed, leveraging its ability to learn complex non-linear relationships.

The Multilayer Perceptron (MLP) is the most widely recognized and frequently used type of artificial neural network (ANN), employing a supervised training process with data examples that have known outputs (Bishop, 1995). The MLP structurally generalizes the artificial neuron known as the Rosenblatt perceptron,' which utilizes the Heaviside activation function (Taud & Mas, 2018). Furthermore, the MLP can approximate any continuous, bounded, differentiable, nonlinear function with defined inputs in a compact space to arbitrary precision (Alves *et al.*, 2023). This is achievable through the additive combination of base functions, which, in the case of an MLP, are ridge functions (Almeida, 1997).

MLP Architecture in the Theta-MLP Method

MLPs structure neurons into a series of layers, with each layer consisting of neurons arranged in parallel (Taud & Mas, 2018). Neurons within the same layer do not interact; they only transmit their outputs forward, which is a defining feature of a *feedforward* structure (Jaiswal, 2024). Since neurons are organized in layers, superscript numbers are used in the notation to indicate their respective layers. MLPs always include an input layer, one or more hidden layers—two hidden layers in the case of the Theta-MLP method—and an output layer (Almeida, 1997). The input layer is denoted by ($l = 0$), where the neurons receive the problem's input variables (Chan *et al.*, 2023). Each neuron in the input layer is a special type that transmits the value of a specific input feature to the neurons in the hidden layer (Alves *et al.*, 2023):

$$\hat{y}_j^{(l=0)} = x_j, \tag{36}$$

In the hidden layers ($0 < l < L$) and the output layer ($l = L$), neurons—standard artificial neurons—receive inputs from the outputs of neurons in the previous layer (Alves *et al.*, 2023). The hidden layer is responsible for mapping the input signal non-linearly into a different space, depending on the problem's requirements (Alves *et al.*, 2023). Since combinations of linear functions are also linear, the activation functions of the hidden nodes are non-linear (Taud & Mas, 2018). The last layer in the MLP is the output layer, where neurons produce the output variable values relevant to the problem; in the Theta-MLP, this involves forecasting the errors of the selected Theta method in the error correction term.

The MLP correction term is implemented with the following architecture:

- **Input Layer**: Accepts residuals from the Theta forecast as input. The input is typically structured as lagged residuals or additional covariates.
- **Hidden Layers**:
  - The first hidden layer consists of 100 neurons with *ReLU (Rectified Linear Unit)* activation.
  - The second hidden layer contains 50 neurons, also employing $ReLU$ activation.
- **Dropout Regularization**: A dropout rate of 20% is applied to each hidden layer to mitigate overfitting by randomly deactivating a fraction of neurons during training.
- **Output Layer**: A single neuron predicts the next residual correction, $\hat{r}_t$.

Training the MLP

The training process of the Multilayer Perceptron (MLP) within the Theta-MLP method is conducted using supervised learning principles (Taud & Mas, 2018). This involves backpropagation, which consists of two phases: the forward pass and the backward pass (Jaiswal, 2024). The MLP takes the residuals from the Theta method forecasts, defined as $\varepsilon_t = y_t - \widehat{y}_{t,\theta}$, as input, and predicts the residual correction ($\hat{r}_t$) as output.

During the forward pass, the residuals are propagated through successive layers of the network, composed of neurons with nonlinear activation functions, to generate the predicted residual corrections ($\hat{r}_t$) (Almeida, 1997). An error signal is calculated by comparing these predictions with the observed residuals using the Mean Squared Error (MSE) loss function:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\varepsilon_i - \hat{r}_i)^2 \tag{37}$$

Here, ε_i represents the actual residual, $\hat{r}_i$ is the predicted residual correction, and n is the number of training examples.

In the backward pass, gradients of the error function are computed with respect to the network's weights. These gradients are propagated back through the network to iteratively update the weights in a direction that minimizes the error (Taud & Mas, 2018). This optimization step is performed using the Adam optimizer, which ensures stable and efficient convergence.

The model is trained with the following hyperparameters: a learning rate of 0.0005, 50 epochs, a batch size of 10, and a 10% validation split to monitor performance on unseen data . Weights are initialized randomly within a uniform distribution between $[-1, 1]$. Training continues until the MSE criterion is met, allowing the MLP to learn complex nonlinear dependencies in the residuals effectively (Chan *et al.*, 2023).

Combining Theta and MLP Forecasts

Once trained, the MLP predicts residual corrections iteratively over the forecast horizon h. The final forecast combines the Theta forecast with the MLP-predicted corrections:

$$\hat{y}_t^{\theta\text{-MLP}} = \widehat{y}_{t,\theta} + \phi \hat{r}_t \tag{38}$$

Here, ϕ is a scaling parameter that balances the contributions of the Theta forecast and the MLP correction, providing dynamic adjustment.

Optimizing ϕ

The parameter ϕ is a critical component in the Theta-MLP method, balancing the contributions of the statistical Theta forecast and the machine learning-based correction term. To optimize ϕ, *Maximum Likelihood Estimation (MLE)* is employed under the assumption that the residual correction term, $\hat{r}_t$, follows a normal distribution. This assumption simplifies the modeling process and ensures mathematical tractability. The likelihood function under this assumption is defined as:

$$L(\phi) = \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_t - (\widehat{y}_{t,\theta} + \phi\hat{r}_t))^2}{2\sigma^2}\right), \tag{39}$$

where y_t represents the actual observed value, $\widehat{y}_{t,\theta}$ is the Theta forecast, $\hat{r}_t$ is the predicted residual correction from the MLP, and σ^2 denotes the variance of the residuals.

To simplify optimization, the negative log-likelihood is minimized, which is equivalent to maximizing the likelihood. The negative log-likelihood function is expressed as:

$$\log L(\phi) = -\frac{T}{2}\log\big(2\pi\sigma^2\big) - \frac{1}{2\sigma^2}\sum_{t=1}^{T}(y_t - (\widehat{y}_{t,\theta} + \phi\hat{r}_t))^2. \tag{40}$$

Minimizing this function with respect to ϕ involves taking the derivative of the log-likelihood with respect to ϕ, setting it to zero, and solving for ϕ. This yields the optimal parameter value as:

$$\phi = \frac{\sum_{t=1}^{T}(y_t - \widehat{y}_{t,\theta})\hat{r}_t}{\sum_{t=1}^{T}\hat{r}_t^2}. \tag{41}$$

Advantages of the MLP Correction Term

The MLP correction term significantly enhances the Theta-MLP method by addressing non-linear residual patterns, thereby enabling more accurate forecasts. Through the dynamic adjustment of the scaling factor ϕ, the method effectively balances the contributions of statistical forecasting and machine learning components. This hybrid approach is particularly effective in scenarios where traditional statistical methods alone fail to capture intricate non-linear patterns inherent in the time series. By leveraging the MLP's capacity for non-linear mapping and the robust statistical properties of the Classical Theta method, the Theta-MLP emerges as a versatile and powerful tool for forecasting across diverse datasets.

1.4.3 Theta-MLP method: steps & procedure

As discussed earlier, the Theta-MLP method extends the Classical Theta method by establishing a topology of eight candidate Theta models. These models consider both linear and non-linear trends, additive and multiplicative Theta lines, and both additive and multiplicative seasonality, while also optimizing θ, allowing the slope of the trend to either be damped or expanded when needed. Furthermore, the Theta-MLP method applies an MLP correction term, which is derived from the residuals of the selected Theta model. This correction term leverages a Multi-Layer Perceptron (MLP) to model and predict residual patterns, effectively addressing complex, non-linear components that the Theta model cannot fully capture.

The method begins by applying a Box-Cox transformation to stabilize variance and address potential non-linearities. The transformed series is then examined for negative values and seasonality, resulting in further exclusions of candidate models. For the remaining models, θ is optimized using the Mean Absolute Error (MAE), and the final Theta model for forecasting is selected based on the in-sample MAE. Next, the residuals of the final Theta model are calculated, and a Multi-Layer Perceptron (MLP) is trained to learn and predict the residual patterns. The trained MLP is then used to forecast future residuals. Subsequently, the scaling factor ϕ is optimized to combine the Theta model's forecasts with the MLP residual corrections. The final forecast is produced by combining the Theta model's forecasts with the MLP residual predictions, scaled by ϕ.

The Theta-MLP method, which combines the extensions of the Classical Theta method by Assimakopoulos & Nikolopoulos (2000) with a Multi-Layer Perceptron (MLP), is outlined by the following steps:

Step 0. Box-Cox Transformation: The Box-Cox transformation is applied to stabilize variance and address non-linearities in the time series, preparing the data for subsequent modeling.

Step 1. Testing for Negative Values: Following the Box-Cox transformation, the time series is tested for negative values. If negative values are present, models with multiplicative expressions are excluded.

Step 2. Seasonality Test: Next, the time series is tested for seasonality by analyzing the autocorrelation function (ACF) at a 90% confidence level. If seasonality is detected, non-seasonal models are excluded from the candidate pool to focus on models that account for seasonal patterns.

Step 3. Optimizing θ: For each remaining candidate Theta model, the parameter θ is optimized by minimizing the in-sample Mean Absolute Error (MAE). This step ensures that the chosen θ value provides the best fit to the historical data, balancing the components of the Theta line and trend appropriately.

Step 4. Model Selection: The Theta model that minimizes the in-sample Mean Absolute Error (MAE) is selected as the final model for forecasting.

Step 5. Generate Forecast: Using the selected Theta model, forecasts are generated for the desired forecast horizon.

Step 6. Residual Extraction: After generating forecasts using the selected Theta model, the residuals are calculated by subtracting the fitted values of the model from the actual observed values.

Step 7. MLP Training: Next, a Multi-Layer Perceptron (MLP) is trained on the residuals to learn these unexplained patterns. The trained MLP serves as a correction mechanism for the Theta model's forecasts.

Step 8. MLP Forecasting: After training, the Multi-Layer Perceptron (MLP) is used to forecast future residuals for the desired forecast horizon. These MLP forecasts represent the patterns and irregularities in the time series that the Theta model could not capture.

Step 9. Optimizing ϕ: In this step, the scaling factor ϕ is optimized to effectively combine the Theta model's forecasts with the MLP's residual predictions. The parameter ϕ determines the weight of the MLP corrections in the final forecast.

Step 10. MLP Correction: Subsequently, the forecasts from the Theta model and the residual predictions from the Multi-Layer Perceptron (MLP) are combined to produce the final forecast. This combination is achieved by adding the MLP-predicted residuals, scaled by the optimized factor ϕ, to the original forecasts generated by the Theta model.

Step 10. Reverse Box-Cox Transformation: Finally, the combined forecasts, which were generated on the Box-Cox transformed scale, are reverted back to the original scale using the inverse Box-Cox transformation.

The Theta-MLP method is a hybrid approach that combines the Theta method, a statistical forecasting technique, with a Multilayer Perceptron (MLP), a type of neural network, to enhance forecasting accuracy. Its key strength lies in merging the advantages of both approaches, enabling the model to capture both linear and non-linear patterns. This flexibility makes the Theta-MLP highly effective for a wide range of time series, whether they exhibit linear or exponential trends, or additive or multiplicative seasonal patterns. The Theta-MLP algorithm includes an MLP correction term that models residuals and non-linear patterns, while the Theta method addresses long-term trend and seasonality. The slope of the trend is subsequently adjusted so that forecasts are either damped or expanded when necessary. The inclusion of the Box-Cox transformation further enhances the model's ability to handle non-normal and heteroscedastic data. Additionally, the method features automatic model selection based on in-sample Mean Absolute Error (MAE), ensuring the optimal combination of linear and non-linear trends is selected for each time series. Finally, applying Brent's method to optimize θ reduces computational time while ensuring a reliable solution. The enhanced forecasting accuracy of the Theta-MLP methods comes from the fact that no single method can fully capture time series patterns adequately. Hybrid models like Theta-MLP combine the strengths of multiple forecasting approaches, each capturing different components of the time series. By leveraging the strengths of each method, these models effectively capture various aspects of the data, and the errors of the individual models tend to cancel out, resulting in improved overall accuracy.

Despite employing Brent's method to reduce computational time, the Theta-MLP algorithm still demands considerable computational resources, primarily due to the MLP training process. This limitation makes it less suitable for large-scale time series applications like the M4 dataset. With 100,000 time series, the M4 dataset presents significant computational challenges for training machine learning models. While the MLP method is relatively simple and more computationally efficient than other neural networks, such as RNNs (Recurrent Neural Networks) or LSTMs (Long Short-Term Memory Networks), the dataset's size still makes the process demanding. Therefore, to mitigate this challenge, focusing on the hourly and daily series of the M4 dataset provides a balanced approach. The hourly series, consisting of 414 time series, captures the complexities of high-frequency data, including strong daily and weekly seasonality patterns. Meanwhile, the daily series, comprising 4,227 time series, emphasizes high-frequency data with strong trend characteristics and negligible seasonality. These subsets allow for a detailed evaluation of forecasting methods without overwhelming computational demands, enabling the testing of hypotheses and exploration of advanced techniques like Theta-MLP.

Another structural disadvantage of the Theta-MLP method, due to its complexity, is the difficulty in achieving 100% replicability. Like other complex models, such as Smyl's ES-RNN, the Theta-MLP method incorporates stochastic elements in the training process, relies on computational resources, and involves intricate transformations, which can result in slight variations in outcomes

across different environments. In the context of the M4 Competition, four out of the five most accurate models are not 100% replicable because of these inherent complexities. Nevertheless, a 98% level of replicability, as seen with the Theta-MLP method, is considered fully replicable by the M4 Competition's standards.

1.5 Performance Measures for Point Forecasts

To rigorously test the proposed hypotheses, a set of established performance metrics from the forecasting literature will be utilized. Over time, various measures have been developed to assess forecasting accuracy (Hyndman & Koehler, 2020). During the M3 Competition, multiple metrics were applied, but no definitive consensus was reached on the relative strengths and weaknesses of each method (Goodwin & Lawton, 1999). To resolve this issue, the M4 Competition introduced the Overall Weighted Average (OWA), which combines the two most widely used accuracy metrics: Symmetric Mean Absolute Percentage Error (sMAPE; Makridakis, 1993) and Mean Absolute Scaled Error (MASE; Hyndman & Koehler, 2006). This combination offers a more objective approach to evaluating forecasting performance (Assimakopoulos *et al.*, 2020). Hence, in this thesis, the OWA will serve as the primary measure for both evaluating the performance of different forecasting methods and testing the proposed hypotheses.

1.5.1 sMAPE

The sMAPE uses a scale-independent percentage error, which is intuitive and part of everyday vocabulary (Makridakis, 1993). One rationale for employing sMAPE lies in its continuity with past M Competitions, particularly after addressing its drawbacks by excluding negative and small positive values, while still maintaining its intuitive, everyday interpretation (Assimakopoulos *et al.*, 2020). For an optimal forecast, the goal is to achieve the lowest possible sMAPE value, as this indicates minimal deviation from actual values and hence a more accurate forecasting model (Assimakopoulos *et al.*, 2020). The sMAPE is defined by the following formula (Makridakis, 1993):

$$sMAPE = \frac{2}{h} \sum_{t=n+1}^{n+h} \frac{|y_t - \hat{y}_t|}{|y_t| + |\hat{y}_t|} \times 100\% \tag{42}$$

Here, y_t represents the actual value of the time series at time t, and $\hat{y}_t$ is the forecasted value. The term h denotes the length of the forecasting horizon, n is the total number of in-sample data points, and m indicates the interval between observations as determined by the organizers for each data frequency in the M4 Competition (M4 Team, 2018).

The rationale for setting specific values of m is to enable consistent measurement and comparison of the accuracy of various forecasting models (Assimakopoulos *et al.*, 2020). This standardization facilitates the calculation of the Mean Absolute Scaled Error (MASE). Yearly, quarterly, and monthly data, for example, have well-defined seasonal cycles. There are 12 months in a year and four quarters (hence, $m = 12$ and $m = 4$, respectively), while for yearly data, $m = 1$ is straightforward. The seasonality of weekly, daily, and hourly data is more ambiguous. A year consists of approximately 52 weeks plus one or two days, depending on leap years. The seasonality of daily data can vary based on how many days are defined in a week (5, 6, or 7), depending on the dataset's context. For hourly data, a daily pattern would imply $m = 24$. However, hourly data might also exhibit double (7 days × 24 h) or even triple (7 days × 24 h × 12 months) seasonality.

By proposing specific m values, the organizers of the M4 Competition aim to simplify the underlying assumption of seasonality and align the treatment of seasonality in the M4 Competition with the

M3 Competition, where data classified as 'other' were considered non-seasonal, despite this data occurring at daily and weekly frequencies (Assimakopoulos *et al.*, 2020). The m values were announced at the beginning of the M4 Competition, emphasizing that the m values only pertained to the estimation of the MASE, whereas participants were free to consider any other alternative value for m when generating forecasts (M4 Team, 2018).

1.5.2 MASE

The MASE is designed to address some issues inherent in the sMAPE and has been recognized as a superior alternative due to its more favorable mathematical properties (Franses, 2016). It is distinguished by a defined mean and a finite variance (Hyndman & Koehler, 2006). Additionally, it is unaffected by scale and can be computed for single forecast horizons (Hyndman & Koehler, 2006). MASE is extensively used in modern forecasting literature due to its independence from data scale, lower sensitivity to outliers, and instances of infinite or undefined values only arising when all historical observations are identical—a situation that is practically unachievable (Assimakopoulos *et al.*, 2020). A value below one indicates that the forecast's performance surpasses that of the average one-step Seasonal Naïve forecast calculated within the sample data (Hyndman & Koehler, 2006). Conversely, a value above one suggests an inferior forecast (Hyndman & Koehler, 2006). Achieving a MASE as low as possible is essential, as a lower value indicates a more accurate and efficient forecast relative to the Seasonal Naïve model (Franses, 2016). The formula representing the MASE is as follows (Hyndman & Koehler, 2006):

$$MASE = \frac{\frac{1}{h}\sum_{t=n+1}^{n+h}|y_t - \hat{y}_t|}{\frac{1}{n-m}\sum_{t=m+1}^{n}|y_t - y_{t-m}|} \tag{43}$$

Where y_t represents the actual value of the time series at time t, while $\hat{y}_t$ denotes the forecasted value. The forecasting horizon is represented by h, the number of in-sample data points by n, and m refers to the interval between observations as defined by the organizers of the M4 Competition.

The method for calculating MASE in the M4 Competition differs from the original approach proposed by Hyndman & Koehler (2006) (Assimakopoulos *et al.*, 2020). Initially, m was fixed at one for all data frequencies, using the in-sample absolute error from the Naïve 1 method (essentially a random walk forecast) as a benchmark to scale the absolute error of the evaluated method (Hyndman & Koehler, 2006). In the M4 Competition, however, the organizers chose the Seasonal Naïve method as the benchmark, considering it more appropriate for seasonal data due to its better scaling (Assimakopoulos *et al.*, 2020). Although Naïve 2 could also be considered an alternative due to similar properties, Seasonal Naïve was preferred, as determining seasonal series and calculating seasonal indices can lack a standardized method (Assimakopoulos *et al.*, 2020). Consequently, using Naïve 2 for MASE benchmarking could make result replication more challenging. Seasonal Naïve, by contrast, is simpler to compute and does not require additional assumptions or information (Assimakopoulos *et al.*, 2020).

1.5.3 OWA

To compute the OWA of sMAPE and MASE, first obtain the relative sMAPE and relative MASE by dividing the MASE and sMAPE of the method by those of Naïve 2. Despite its more complex computation process, Naïve 2 was selected over Seasonal Naïve for the OWA calculations. This choice was based on the fact that Naïve 2 is very popular in time-series forecasting and often displays superior accuracy (Assimakopoulos *et al.*, 2020). Moreover, Naïve 2 has been regularly used as a benchmark in numerous forecasting studies and has been estimated in previous M Competitions, thereby enabling direct comparisons (M4 Team, 2018).

The sMAPE and MASE are initially calculated for each series by averaging the errors computed for each forecasting horizon. These averages are then aggregated to obtain the overall value for the entire dataset. It should be noted that the calculation methods for sMAPE and MASE have evolved from those used in earlier M Competitions (Assimakopoulos *et al.*, 2020). Historically, errors from each series and forecasting horizon were averaged together, meaning subsets with a larger number of series and longer forecasting horizons had a disproportionately bigger impact on the accuracy estimate. To address this, the M4 Competition begins by averaging the errors at the individual series level, ensuring each series within the dataset is given equal weight (M4 Team, 2018).

Next, the simple arithmetic mean of the relative MASE and relative sMAPE is computed. OWA calculations are performed only once, at the end of the evaluation for the entire sample of series. It is important to note that a lower OWA value indicates better performance, as it signifies a smaller overall error across the dataset. Code for computing the OWA can be found in the M4 GitHub repository and is specified by the following formulas:

$$\begin{aligned} \textit{Relative MASE} &= \frac{\text{MASE of the method}}{\text{MASE of Naïve 2}}, \\ \textit{Relative sMAPE} &= \frac{\text{sMAPE of the method}}{\text{sMAPE of Naïve 2}}. \end{aligned}$$

The *Overall Weighted Average (OWA)* is computed as:

$$OWA = \frac{\text{Relative MASE} + \text{Relative sMAPE}}{2}. \tag{44}$$

Recognizing potential limitations, the organizers of the M4 Competition acknowledge that the selected measures might not be the most appropriate to use, anticipating diverse opinions and even objections (Assimakopoulos *et al.*, 2020). Nonetheless, they are confident that the large sample of series used in the M4 Competition helps mitigate the potential impact that different error measures could have on determining the final ranks of the participating methods. Given its more objective approach in combining these critical accuracy metrics, OWA serves as the primary measure for evaluating the performance of different forecasting methods and testing the proposed hypotheses in this thesis.

1.6 Diebold-Mariano Test

The previous section highlighted OWA, the equally weighted average of sMAPE and MASE, as the predominant performance metric for evaluating forecasting accuracy and testing the hypotheses. To advance this analysis, it is essential to determine whether the difference in forecasting accuracy between two forecasts is statistically significant. The Diebold-Mariano (DM) test (Diebold and Mariano, 1995) is a widely used econometric tool designed to address this question (Diebold, 2012).

The DM test is applied to an accuracy measure, which in this thesis is the mean squared error (MSE), defined as $MSE = \frac{\sum_{t=1}^{T} e_t^2}{T}$ (Diebold, 2012). This follows an out-of-sample forecasting study, where S data points are predicted from a dataset of size T (Costantini & Kunst, 2011). By using forecast errors as primitives, the DM test adopts a model-free approach. This means that the forecast errors do not need to originate from specific models, and even if they do, the models themselves do not need to be known to the econometrician (Diebold, 2015). Instead, the test relies on assumptions made directly about the forecast errors or, more specifically, the forecast error loss differentials (Diebold, 2015).

Let $e_{it} = \hat{y}_{it} - y_t, \quad i = 1, 2$, represent the forecast error for model i at time t. The loss associated with e_{it} is denoted by $L(e_{it})$, and for MSE, $L(e_{it}) = e_{it}^2$ (Diebold, 2012). The time-t loss differential is defined as $d_{12t} = L(e_{1t}) - L(e_{2t}) = e_{1t}^2 - e_{2t}^2$. If the loss differential has zero expectation for every t, the two forecasts are considered to have equal predictive accuracy (Mariano, 2000). The only assumption required about the loss differential is covariance stationarity, defined as follows (Diebold, 2012):

$$\text{Assumption DM: } \begin{cases} E(d_{12t}) = \mu, & \forall t, \\ \text{cov}(d_{12t}, d_{12(t-\tau)}) = \gamma(\tau), & \forall t, \\ 0 < \text{var}(d_{12t}) = \sigma^2 < \infty. \end{cases} \tag{45}$$

The null hypothesis H_0 posits that the two forecasting models being compared have equal predictive accuracy, meaning equal expected loss ($E(d_{12t}) = 0$) (Mohammed & Mousa, 2019). Conversely, the alternative hypothesis H_1 suggests that the two forecasting methods have unequal predictive accuracy ($E(d_{12t}) \neq 0$) (Mohammed & Mousa, 2019). Under H_0, the DM statistic follows a standard normal distribution:

$$DM_{12} = \frac{\bar{d}_{12}}{\hat{\sigma}_{\bar{d}_{12}}} \rightarrow N(0, 1), \tag{46}$$

where $\bar{d}_{12} = \frac{1}{T}\sum_{t=1}^{T} d_{12t}$ represents the sample mean loss differential, and $\hat{\sigma}_{\bar{d}_{12}}$ is a consistent estimator of the standard deviation (Diebold, 2012). At a 95% confidence level, the null hypothesis of equal predictive accuracy is rejected if $|DM| > z_{0.025}$, where $z_{0.025} = 1.96$ (Diebold, 2012).

2 Part II: Simulation Study

Part II of this thesis employs synthetically generated data to test the three hypotheses formulated in the introduction. Research frequently uses synthetically generated data because it provides a 'ground truth' that is highly beneficial for the development and assessment of forecasting pipelines (Sun *et al.*, 2021). Although there is substantial interest in synthetic data, a widely accepted definition has yet to be established (Dibben *et al.*, 2016). To encompass all applications and methods associated with synthetic data, the following definition, as proposed by Cohen *et al.* (2022), is adopted:

Definition: *"Synthetic data is data that has been generated using a purpose-built mathematical model or algorithm, with the aim of solving a (set of) data science task(s)." (Cohen et al., 2022).*

Synthetic data, produced through model-driven generation, contrasts with real data, which arises directly from empirical interactions and observations in the natural world (e.g., financial transactions, satellite images, medical tests) (Cohen *et al.*, 2022). In this thesis, the real-world data utilized is derived from the M4 Competition, a topic that will be comprehensively examined in Part III. The application of synthetic data to address specific problems is not a recent development. Instances of its use can be traced back to the 1940s with the pioneering work of Stanislaw Ulam and John von Neumann on Monte Carlo simulation methods (Branciamore *et al.*, 2020).

The simulation study is designed to rigorously evaluate the hypotheses outlined in the introduction by replicating key characteristics of real-world time series using data-generating processes (DGPs). Three hourly DGPs are developed, reflecting varying levels of trend complexity and both additive

and multiplicative seasonality, to comprehensively test the proposed hypotheses under different scenarios. For each DGP, **S = 100 simulations** are conducted to provide sufficient statistical power for performance evaluation. Each simulation generates an hourly time series of length **T = 505 observations**, corresponding to slightly more than 21 days, or just over three full cycles of weekly seasonality. This ensures that the time series captures sufficient variability and seasonal patterns. In addition to performance metrics such as sMAPE, MASE, and OWA, the Diebold-Mariano (DM) test is applied at a 95% confidence level to determine whether the observed differences in forecasting accuracy between methods are statistically significant.

The decision to simulate hourly series is guided by two primary considerations: first, the complexity of hourly data, which incorporates both daily and weekly seasonality, makes it an ideal testing ground for advanced forecasting methods like Theta-MLP; and second, the need to manage the computational demands of the Theta-MLP model. A smaller subset of hourly series strikes a practical balance between data complexity and feasible runtime while still providing sufficient variability to rigorously evaluate the hypotheses. Since the empirical application in Part III includes the hourly series of the M4 dataset, simulating hourly series ensures consistency between Part II and Part III. These DGPs are specifically designed to capture varying levels of trend complexity—ranging from linear to damped and exponential—and to incorporate both additive and multiplicative forms of seasonality:

- **DGP 1**: Hourly time series with a linear deterministic trend and additive daily and weekly seasonality.
- **DGP 2**: Hourly time series with a damped linear deterministic trend and multiplicative daily and weekly seasonality.
- **DGP 3**: Hourly time series with an exponential trend.

The differences in trend characteristics across the DGPs facilitate testing **H3**, which posits that damping the trend improves forecasting accuracy on average across time series but leads to underforecasting in strongly trending series. For **DGP 1** (linear trend), Damped ES is expected to underperform Holt ES in terms of OWA, with the DM test anticipated to reject the null hypothesis of equal accuracy, as the linear trend does not benefit from dampening. In **DGP 2** (damped trend), Damped ES is expected to outperform Holt ES in terms of OWA, with the null hypothesis of equal forecasting accuracy likely to be rejected in the DM test, as Damped ES better aligns with the natural behavior of the trend. Furthermore, for **DGP 3** (exponential trend), Holt ES is expected to outperform Damped ES in terms of OWA. While both methods struggle to fully capture the exponential trend, Holt ES assumes a linear trend, which introduces less bias than the dampened trend assumption of Damped ES, resulting in comparatively better performance. In addition, the DM test is expected to reject the null hypothesis of equal forecasting accuracy between Damped ES and Holt ES at the 95% significance level.

Moreover, **DGP 3** aligns with **H2**, which posits that more complex methods can lead to greater forecasting accuracy compared to simpler ones. The exponential trend in **DGP 3** introduces a high degree of nonlinearity, creating a challenging environment where computational power can be leveraged to predict complex, nonlinear relationships between series. The Theta-MLP method is expected to excel in terms of OWA rank. Additionally, the DM test is anticipated to reject the null hypothesis of equal forecasting accuracy between the Theta-MLP method and both the second most accurate model and the Classical Theta method, not only in **DGP 3** but also in other DGPs. This superior performance is attributed to the Theta-MLP method's ability to extend the Classical Theta method's robust handling of trends and seasonality, as well as to incorporate the Multilayer Perceptron to model nonlinear residuals. This approach effectively leverages the strengths of both statistical and machine learning models, offering superior accuracy in capturing nonlinear trends.

DGP 1 captures a simple linear trend with strong additive seasonality, making it an ideal baseline for testing methods that handle seasonality effectively, such as the Seasonal Naïve, which are

expected to perform relatively well compared to the Theta-MLP hybrid algorithm. To test **H1**, which posits that combinations of statistical methods yield more accurate results than individual methods by leveraging their strengths, the Comb method (a simple arithmetic average of Single, Holt, and Damped Exponential Smoothing) is expected to outperform the individual methods in terms of OWA rank. Furthermore, the null hypothesis of equal forecasting accuracy between the Comb method and Holt, Damped, and SES is expected to be rejected across all DGPs.

2.1 DGP 1: Hourly Time Series with Additive Daily and Weekly Seasonality and a Linear Deterministic Trend

The time series model for **DGP 1** incorporates a linear deterministic trend alongside strong additive daily and weekly seasonal components. The mathematical representation of DGP 1 is given by:

$$y_t = \alpha + \beta t + S_t^D + S_t^W + \varepsilon_t \tag{47}$$

where:

- y_t represents the time series value at time t,
- α is the intercept,
- β is the linear trend coefficient, and
- ε_t denotes the stochastic error term, assumed to follow a standard normal distribution with mean 0 and variance 1.

The seasonal components, S_t^D and S_t^W, capture daily and weekly seasonality, respectively. In DGP 1, S_t^D and S_t^W are simulated with stronger additive effects, ensuring seasonality dominates the time series in a linear context. In contrast, DGP 2 features multiplicative seasonality, where the seasonal fluctuations grow or shrink in proportion to the underlying trend. This design makes DGP 1 ideal for testing methods that excel in capturing strong additive seasonal structures. The components are defined using Fourier series as follows:

Daily Seasonality:

$$S_t^D = \sum_{i=1}^{k} \left(a_i \cos\left(\frac{2\pi i t}{\text{freq}_{\text{daily}}} \right) + b_i \sin\left(\frac{2\pi i t}{\text{freq}_{\text{daily}}} \right) \right) \tag{48}$$

Weekly Seasonality:

$$S_t^W = \sum_{i=1}^{k} \left(a_i \cos\left(\frac{2\pi i t}{\text{freq}_{\text{weekly}}} \right) + b_i \sin\left(\frac{2\pi i t}{\text{freq}_{\text{weekly}}} \right) \right) \tag{49}$$

In these expressions, a_i and b_i are Fourier coefficients, while k represents the number of harmonics used to model the seasonal patterns. For DGP 1, the intercept is $\alpha = 15$ and $\beta = 0.03$. The seasonal components are modeled using the first harmonic ($k = 1$) with Fourier coefficients $a_1 = 3$ and $b_1 = 5$ for daily seasonality (S_t^D), and $a_1 = 1$ and $b_1 = 2$ for weekly seasonality (S_t^W). The frequencies, $\text{freq}_{\text{daily}} = 24$ and $\text{freq}_{\text{weekly}} = 168$, correspond to the periodicities of the daily and weekly seasonality for hourly data.

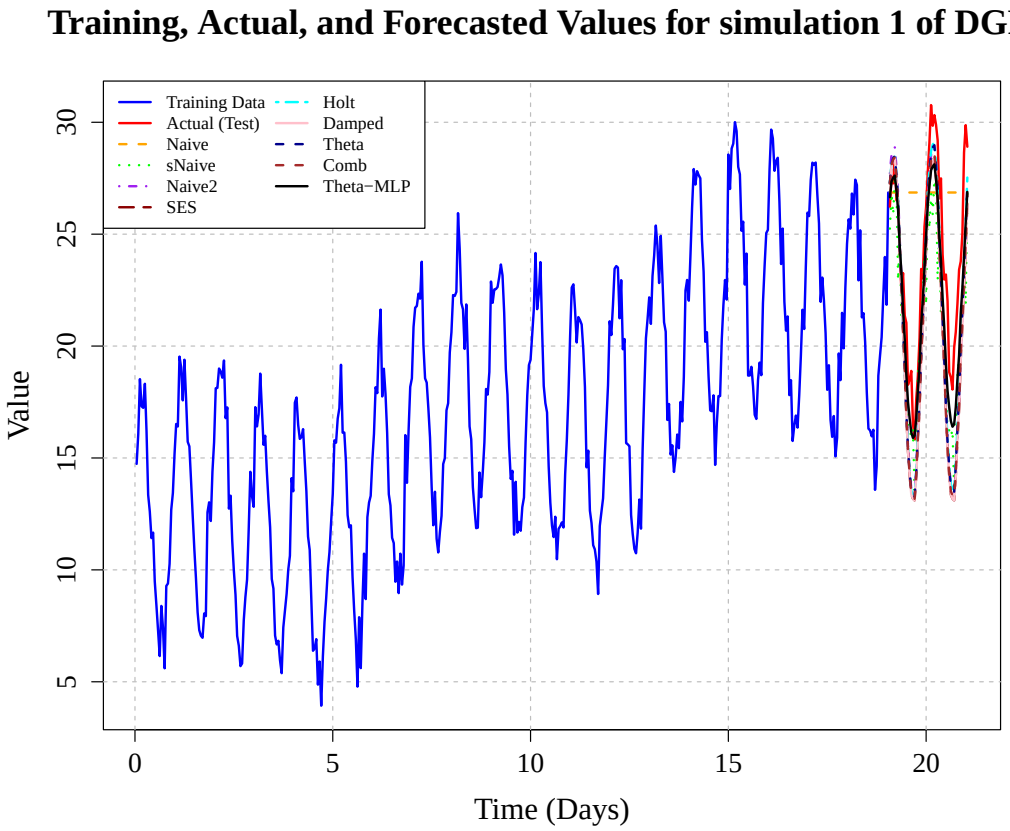

Figure 1: Training data, test data, and forecasted values from the statistical benchmarks and the Theta-MLP method for the first simulation of an hourly time series generated by DGP 1, which features additive daily and weekly seasonality along with a linear deterministic trend.

Table 2: The performance of Naïve 1, Naïve 2, Seasonal Naïve, SES, Holt, Damped, Classical Theta, Comb, and Theta-MLP methods is evaluated in terms of sMAPE, MASE, and OWA. Results are averaged over $T = 100$ simulations of hourly time series generated using DGP 1, which exhibits additive daily and weekly seasonality along with a linear deterministic trend.

Method	**sMAPE**	**MASE**	**OWA**	**Rank**
Naïve	16.537	2.397	1.047	9
sNaïve	12.355	1.651	0.741	2
Naïve2	17.952	2.197	1.000	8
SES	16.574	2.031	0.937	6
Holt	14.626	1.793	0.826	3
Damped	16.848	2.063	0.951	7
Theta	14.929	1.830	0.844	4
Comb	15.994	1.960	0.903	5
Theta-MLP	**8.890**	**1.216**	**0.533**	**1**

The results for DGP 1 demonstrate a notable improvement in the performance of the hybrid Theta-MLP method (OWA: 0.533) compared to the simpler Seasonal Naïve method (OWA: 0.741), which is the second most accurate model for DGP 1, with a 28.06% reduction in OWA. Likewise, the Theta-MLP method outperforms the Naïve and Naïve 2 methods by 49.09% and 46.70% in terms of OWA, respectively, and the DM test rejects the null hypothesis of equal forecasting accuracy between the Theta-MLP and various Naïve methods at a 95% confidence level. The remarkable

performance of the Seasonal Naïve method relative to the Theta-MLP method across the simulations of DGP 1 can be attributed to the strong additive seasonality present in the time series, which the Seasonal Naïve method directly captures by replicating past seasonal patterns. Nevertheless, Naïve 2, which is equivalent to applying the Naïve method to deseasonalized data, performs poorly, ranking eighth. Moreover, the plot in Figure 1 shows that the black line representing the Theta-MLP method closely follows the red line of the test data.

Similarly, the Theta-MLP method achieves a 36.8% reduction in OWA over the Classical Theta method (OWA: 0.844), which highlights the effectiveness of more complex hybrid approaches. By combining the extensions of the Classical Theta method's robust handling of trend and seasonality with an MLP-based correction term that addresses non-linear residuals, the Theta-MLP method improves forecasting accuracy. The DM test also rejects the null hypothesis of equal forecasting accuracy between the Theta-MLP and the Classical Theta method at a 95% significance level. Altogether, this supports **H2**, which posits that more complex methods, such as Theta-MLP, have the potential to achieve greater forecasting accuracy compared to simpler methods.

The performance of the Classical Theta method, which ranks fourth in terms of OWA, highlights the inherent limitation that no single model can comprehensively capture all the patterns present in time series data. The Classical Theta method accounts for both linear trends and multiplicative seasonality by decomposing the time series into a long-term trend (using linear regression) and a smoothed component (via SES). This decomposition provides a clear advantage over simpler statistical approaches that model only trend or seasonality. For instance, the Classical Theta method achieves a 19.4% reduction in OWA compared to Naïve 1 (OWA: 1.047), which does not account for trend or seasonality, and a 15.6% reduction compared to Naïve 2 (OWA: 1.000), which incorporates only seasonal components. The DM test confirms this improvement, rejecting the null hypothesis of equal forecasting accuracy at the 95% confidence level and demonstrating that the Classical Theta method is statistically significantly more accurate than both Naïve 1 and Naïve 2. Additionally, the Classical Theta method shows an 11.3% improvement over the Damped Exponential Smoothing method (OWA: 0.951), which models only a damped linear trend, and the results of the DM test reject the null hypothesis of equal forecasting accuracy. Nevertheless, despite these improvements, the Classical Theta method's OWA score of 0.844 is higher than Holt's linear trend method (OWA: 0.826), indicating a decrease in performance by 2.1%, and it trails the Seasonal Naïve model (OWA: 0.741) by 12.2%. Analysis using the DM test subsequently rejects the null hypothesis of equal forecasting accuracy between the Classical Theta method and both the Seasonal Naïve model and Holt ES.

Interestingly, the results from DGP 1 challenge the assumption in **H1** that combining forecasting methods consistently leads to superior accuracy. The Comb method (OWA: 0.903) ranks fifth, performing 9.3% worse than Holt, which ranks third with an OWA of 0.826. The results from the DM test between the Comb and Holt methods also reject the null hypothesis, showing that Holt ES statistically significantly outperforms the Comb method. Moreover, Holt ES (OWA: 0.826) outperforms Damped ES (OWA: 0.951) by 15.1% in DGP 1, and the null hypothesis of equal forecasting accuracy in the DM test is also rejected at the 95% significance level. This result reinforces **H3**, as damping the trend in a series with a strong linear component, such as DGP 1, introduces unnecessary bias and leads to under-forecasting, whereas Holt ES aligns better with the natural persistence of the linear trend.

2.2 DGP 2: Hourly Time Series with Damped Linear Deterministic Trend and Multiplicative Daily and Weekly Seasonality

The time series model for **DGP 2** incorporates a damped linear deterministic trend combined with multiplicative daily and weekly seasonal components. In contrast to DGP 1, which features a linear deterministic trend with additive seasonality, DGP 2 includes a trend that diminishes over time, aligning with the assumptions of the Damped ES method, as well as multiplicative seasonality, where the seasonal fluctuations grow or shrink in proportion to the underlying trend. Following

this, Damped ES is expected to outperform Holt's linear method in terms of OWA, and the null hypothesis of equal forecasting accuracy in the DM test is expected to be rejected, as Holt assumes a consistent linear trend and tends to over-project the damped behavior of the trend in DGP 2. The mathematical representation of DGP 2 is given by:

$$y_t = (\alpha + \beta\phi^t) \cdot S_t^D \cdot S_t^W + \varepsilon_t \tag{50}$$

where:

- y_t represents the time series value at time t,
- α is the intercept,
- β is the linear trend coefficient,
- ϕ is the damping factor ($0 < \phi < 1$),
- S_t^D and S_t^W capture multiplicative daily and weekly seasonality, respectively,
- ε_t denotes the stochastic error term, assumed to follow a normal distribution with mean 0 and variance 2.

The seasonal components, S_t^D and S_t^W, are modeled using Fourier series as follows:

Daily Seasonality:

$$S_t^D = \prod_{i=1}^{k} \left(a_i \cos\left(\frac{2\pi i t}{\text{freq}_{\text{daily}}} \right) + b_i \sin\left(\frac{2\pi i t}{\text{freq}_{\text{daily}}} \right) \right) \tag{51}$$

Weekly Seasonality:

$$S_t^W = \prod_{i=1}^{k} \left(a_i \cos\left(\frac{2\pi i t}{\text{freq}_{\text{weekly}}} \right) + b_i \sin\left(\frac{2\pi i t}{\text{freq}_{\text{weekly}}} \right) \right) \tag{52}$$

For DGP 2, the seasonal components are modeled using the first harmonic ($k = 1$), with Fourier coefficients $a_1 = 0$ and $b_1 = 0.1$ for daily seasonality (S_t^D), and $a_1 = 0$ and $b_1 = 0.2$ for weekly seasonality (S_t^W). The frequencies $\text{freq}_{\text{daily}} = 24$ and $\text{freq}_{\text{weekly}} = 168$ correspond to the periodicities of the daily and weekly cycles for hourly data. The damped linear trend is defined by the damping factor $\phi = 0.95$, the trend coefficient $\beta = 0.05$, and the intercept $\alpha = 5$.

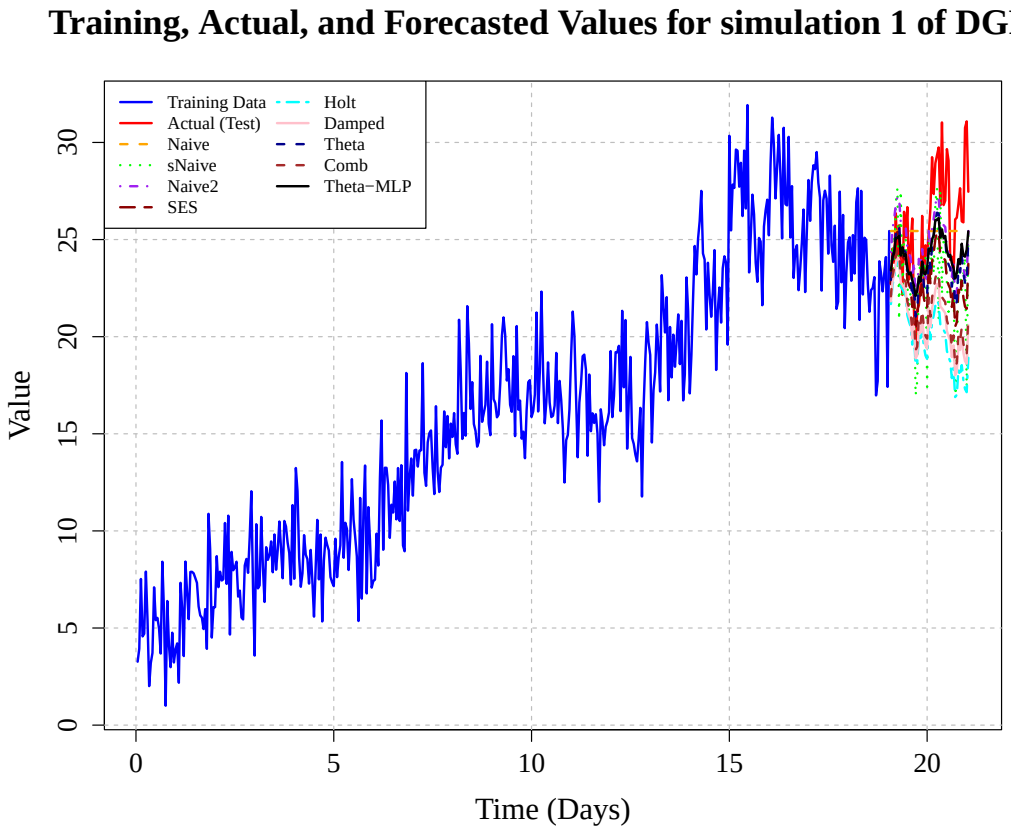

Figure 2: Training data, test data, and forecasted values from the statistical benchmarks and the Theta-MLP method for the first simulation of an hourly time series generated by DGP 2, which features a damped linear deterministic trend and multiplicative daily and weekly seasonality.

Table 3: The performance of Naïve 1, Naïve 2, Seasonal Naïve, SES, Holt, Damped, Classical Theta, Comb, and Theta-MLP methods is evaluated in terms of sMAPE, MASE, and OWA. Results are averaged over $S = 100$ simulations of hourly time series generated using DGP 2, which exhibits a damped linear deterministic trend combined with multiplicative daily and weekly seasonality.

Method	**sMAPE**	**MASE**	**OWA**	**Rank**
Naïve	13.759	1.235	0.946	3
sNaïve	13.920	1.237	1.024	6
Naïve2	14.910	1.321	1.000	5
SES	13.045	1.173	0.960	4
Holt	21.997	1.853	1.558	9
Damped	17.801	1.547	1.287	8
Theta	10.777	0.982	0.799	2
Comb	17.322	1.511	1.251	7
Theta-MLP	**9.565**	**0.879**	**0.711**	**1**

As expected, Damped ES outperformed Holt ES by 17.4% in terms of OWA (OWA: 1.287 vs. 1.558), demonstrating its ability to better capture the diminishing trend in DGP 2. This is demonstrated in Figure 2, where Damped ES (dashed pink line) aligns better with the actual test data (solid red line) than Holt ES (dashed turquoise line) by flattening the trend more effectively. Nevertheless, both ES models exhibit noticeable deviations from the actual values and rank eighth and ninth, respectively,

underperforming relative to the other statistical benchmarks. Results from the DM test also reject the null hypothesis of equal forecasting accuracy between Damped and Holt ES at the 95% significance level. This supports **H3**, as the damping parameter in Damped ES progressively reduces the trend's influence, preventing over-forecasting at longer horizons. By aligning more closely with the damped trend in DGP 2, Damped ES achieves better accuracy than Holt ES, highlighting the benefit of trend damping in such scenarios.

Furthermore, in DGP 2, the Theta-MLP method demonstrates superior forecasting accuracy in terms of OWA, achieving an 11.0% improvement over the Classical Theta method, the second most accurate model in DGP 2 (OWA: 0.711 vs. 0.799). This is evident in Figure 2, where the forecasted line for the Theta-MLP method aligns more closely with the actual test data compared to other methods. However, this improvement is smaller than the 36.8% improvement observed in DGP 1 (OWA: 0.533 vs. 0.844). Results from the DM test reject the null hypothesis of equal forecasting accuracy, demonstrating that the Theta-MLP method is statistically significantly more accurate than the Classical Theta method. This finding reinforces the idea that the additional computational complexity of the Theta-MLP method is justified by its enhanced performance, supporting **H2**.

Similarly, for DGP 2, the improvement of the Theta-MLP method over simpler benchmarks such as Naïve 1 and Naïve 2 is less pronounced. The Theta-MLP method achieves a 24.8% improvement over Naïve 1 (OWA: 0.711 vs. 0.946) and a 28.9% improvement over Naïve 2 (OWA: 0.711 vs. 1.000). By comparison, the improvements in DGP 1 are significantly larger: 49.1% over Naïve 1 (OWA: 0.533 vs. 1.047) and 46.7% over Naïve 2 (OWA: 0.533 vs. 1.000). The results from DGP 2 also challenge **H1**, which posits that combining forecasting methods generally yields superior accuracy. In this case, the Comb method falls short, lagging 30.3% behind SES in terms of OWA (OWA: 1.251 vs. 0.960). Results from the DM test reject the null hypothesis of equal forecasting accuracy, demonstrating that combining forecasts does not always outperform individual methods.

2.3 DGP 3: Hourly Time Series with Exponential Trend

The time series model for **DGP 3** exhibits an exponential trend without seasonal components, making it particularly suitable for testing both **H2** and **H3**. The Theta-MLP method, which has already demonstrated strong performance in DGP 1 and DGP 2, is expected to achieve even greater success in DGP 3 in terms of OWA. As in DGP 1 and DGP 2, the null hypothesis of equal forecasting accuracy in the DM test between the Theta-MLP method and the second most accurate model as well as the Classical Theta method is likely to be rejected at a 95% confidence level. This can be attributed to its advanced MLP correction component, specifically designed to address non-linear patterns such as exponential trends, and the inclusion of an exponential trend equation. Furthermore, the strongly exponential trend in DGP 3 may cause Damped ES to under-forecast, whereas Holt's linear method could exhibit relatively better performance in terms of OWA, as it does not dampen the trend. Nonetheless, both Damped and Holt ES are likely to perform well compared to seasonally focused models, such as Naïve 2 and Seasonal Naïve, due to the strong trend component present in DGP 3. Consequently, it is assumed that the null hypothesis of equal forecasting accuracy between Damped and Holt ES is also likely to be rejected. The mathematical formulation of DGP 3 is detailed below:

$$y_t = \alpha \cdot e^{\beta t} + \varepsilon_t \tag{53}$$

where:

- y_t represents the time series value at time t,
- α is the intercept, which equals $\alpha = 5$,
- β is the exponential growth rate, which equals $\beta = 0.04$,

- ε_t denotes the stochastic error term, assumed to follow a standard normal distribution with mean 0 and variance 1.

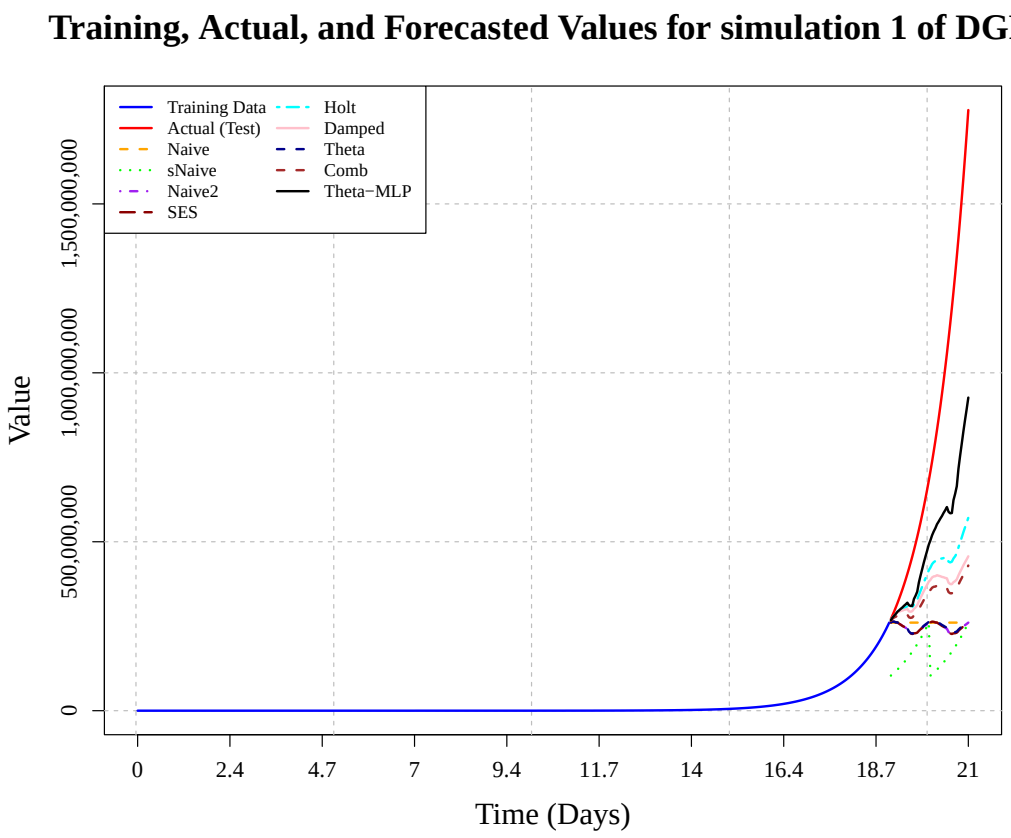

Figure 3: Training data, test data, and forecasted values from the statistical benchmarks and the Theta-MLP method for the first simulation of an hourly time series generated by DGP 3, which features an exponential trend.

Table 4: The performance of Naïve, Seasonal Naïve, Naïve 2, SES, Holt, Damped, Theta, Comb, and Theta-MLP methods is evaluated in terms of sMAPE, MASE, and OWA. Results are averaged over $S = 100$ simulations of hourly time series generated using DGP 3, which exhibits an exponential trend in the absence of seasonal components.

Method	**sMAPE**	**MASE**	**OWA**	**Rank**
Naïve	85.643	57.578	0.966	6
sNaïve	119.052	67.049	1.233	8
Naïve2	89.599	59.000	1.000	7
SES	89.599	59.000	1.000	7
Holt	51.404	41.181	0.636	2
Damped	59.731	45.925	0.722	3
Theta	89.049	58.780	0.995	5
Comb	65.853	48.702	0.780	4
Theta-MLP	**39.194**	**32.662**	**0.496**	**1**

The results highlight the exceptional performance of the Theta-MLP method, achieving a 50.1% improvement over the Classical Theta method (OWA: 0.496 vs. 0.995) and a 22.0% improvement

over the Holt ES model, the second most accurate method in DGP 3 (OWA: 0.496 vs. 0.636). This superior performance is visually evident in Figure 3, where the black line representing the Theta-MLP method closely follows the red line of the steep exponential growth of the actual test data. The DM test further validates these results by rejecting the null hypothesis of equal forecasting accuracy between the Theta-MLP method and both the Classical Theta method and the Holt ES model. By comparison, in DGP 1, the Theta-MLP method achieved a 36.8% improvement over the Classical Theta method (OWA: 0.533 vs. 0.844), while in DGP 2, the improvement was 11.0% (OWA: 0.711 vs. 0.799). The substantial improvement observed in DGP 3 underscores the Theta-MLP method's superior ability to capture non-linear trends, particularly due to the complexity and effectiveness of its MLP component.

Moreover, the results reveal a clear disparity between the Theta-MLP method and the simpler Naïve methods. For DGP 3, the Theta-MLP method (OWA: 0.496) achieved a 48.6% improvement over Naïve 1 (OWA: 0.966) and a 50.4% improvement over Naïve 2 (OWA: 1.000). Similarly, for DGP 1, the Theta-MLP method demonstrated a 49.1% improvement over Naïve 1 (OWA: 0.533 vs. 1.047) and a 46.7% improvement over Naïve 2 (OWA: 0.533 vs. 1.000). For DGP 2, the Theta-MLP method achieved a 24.8% improvement over Naïve 1 (OWA: 0.711 vs. 0.946) and a 28.9% improvement over Naïve 2 (OWA: 0.711 vs. 1.000). These findings, combined with the consistent outperformance of the Theta-MLP method in terms of OWA across the different DGPs, and the DM test consistently rejecting the null hypothesis of equal forecasting accuracy between the Theta-MLP method and the second most accurate method for each DGP, support **H2**, which posits that more complex or statistically sophisticated methods tend to offer superior forecasting accuracy compared to simpler approaches.

Furthermore, both Holt ES (OWA: 0.636) and Damped ES (OWA: 0.722) exhibit strong performance compared to the Seasonal Naïve (OWA: 1.233) and Naïve 2 (OWA: 1.000) models. Specifically, Damped ES achieves a 27.8% improvement over Naïve 2 and a 41.42% improvement over Seasonal Naïve in terms of OWA. Similarly, Holt ES outperforms Naïve 2 by 36.4% and Seasonal Naïve by 48.41% in terms of OWA, underscoring the limitations of seasonality-focused methods when applied to series dominated by strong trends. The DM test rejects the null hypothesis of equal forecasting accuracy between Holt and Damped ES and both the Seasonal Naïve and Naïve 2 models at a 95% significance level. The Holt ES (OWA: 0.636) also outperforms Damped ES (OWA: 0.722), achieving a 13.5% improvement, and the DM test rejects the null hypothesis of equal forecasting accuracy between Holt and Damped ES. Moreover, Figure 3 shows that Holt ES (dashed turquoise line) rises more steeply and aligns better with the strong exponential growth of the actual test data (solid red line) compared to Damped ES (dashed pink line). This aligns with the expectation that, in the presence of a strong exponential trend, as observed in DGP 3, damping the trend results in under-forecasting. These findings strongly support **H3**, which suggests that while trend damping generally enhances forecasting accuracy across time series, it can cause systematic under-forecasting in cases with strong trends.

Lastly, the performance of the Comb method (OWA: 0.780) is notably weaker than that of the individual methods, performing 8.0% worse than Damped ES (OWA: 0.722) and 22.6% worse than Holt ES (OWA: 0.636) in terms of OWA. The DM test rejects the null hypothesis of equal forecasting accuracy between the Comb method and both the Holt ES method and the Damped ES method, directly contradicting **H1**, which posits that combining forecasting methods yields superior accuracy. The Comb method consistently underperforms individual ES models across all DGPs, with the differences in forecasting accuracy consistently found to be statistically significant at a 95% confidence level according to the DM test. In DGP 1, it ranked fifth with an OWA of 0.903, performing 9.3% worse than Holt (OWA: 0.826). Similarly, in DGP 2, it ranked seventh with an OWA of 1.251, trailing SES by 30.3% (OWA: 0.960). These results underscore the limitations of combination methods, which fail to consistently deliver more accurate forecasts than individual models, thereby challenging the validity of **H1**.

2.4 Findings of the Simulation Study

A key finding from the simulation study in Part II is the consistent superiority of the Theta-MLP method, which achieved the highest OWA rank across all DGPs. This superiority is visually demonstrated in Figures 1, 2, and 3, where the solid black line representing the Theta-MLP method closely tracks the actual test data (solid red line) in each DGP. The DM test further validated these results, consistently rejecting the null hypothesis of equal forecasting accuracy between the Theta-MLP method and the second most accurate method for each DGP. These findings highlight the robustness of the Theta-MLP approach in addressing the complexities of trend and seasonality. For instance, in DGP 3, which features an exponential trend, the Theta-MLP method achieved an impressive 50.1% improvement in OWA over the Classical Theta method. This substantial gain underscores the effectiveness of the MLP correction component in capturing non-linear trends, a challenge where simpler models often fall short. Additionally, the Theta-MLP method outperformed Naïve 1 and Naïve 2 by 48.6% and 50.4%, respectively, in DGP 3, while achieving similar improvements in DGPs 1 and 2. Despite the significant computational resources required to train the MLP algorithm, these performance improvements justify the trade-off between computational cost and enhanced forecasting accuracy. The consistent outperformance of the Theta-MLP method across all DGPs provides strong support for **H2**, confirming that more complex forecasting methods lead to greater predictive accuracy compared to simpler methods.

The results from the simulation study provide further support for **H3**, which posits that while trend damping generally improves forecasting accuracy on average, it can lead to under-forecasting in time series exhibiting strong trends. By simulating different trend characteristics across the DGPs, the study highlights the importance of selecting a forecasting method that aligns with the specific trend characteristics of the data. The DM test consistently found the differences in forecasting accuracy between the Damped and Holt ES methods to be statistically significant at the 95% confidence level. For example, in DGPs 1 and 3, where the trend is either linear or exponential, the Damped ES method underperformed Holt ES, showing a 15.1% and 13.5% reduction in OWA, respectively. This underperformance is due to the damping mechanism's inability to capture the full magnitude of the trend, leading to systematic under-forecasting. Conversely, in DGP 2, which features a damped trend, the Damped ES method outperformed Holt ES by 17.4%, reflecting its better alignment with the data's trend characteristics. These findings underscore the critical need to carefully consider the nature of the trend when applying trend damping methods.

Lastly, the performance of the Comb method revealed that combining statistical methods does not necessarily lead to superior forecasting accuracy. Across all DGPs, the Comb method consistently underperformed compared to individual ES models in terms of OWA, with the DM test confirming statistically significant differences in forecasting accuracy at the 95% confidence level. Specifically, in DGP 1, the Comb method performed 9.3% worse than Holt; in DGP 2, it lagged behind SES by 30.3% in OWA; and in DGP 3, it was 8.0% less accurate than Damped ES and 22.6% worse than Holt ES. These results challenge **H1**, which posits that combining statistical methods will always result in improved forecasting accuracy. Altogether, the simulation study in Part II provided a controlled environment to test the hypotheses, offering strong support for **H2** and **H3** while contradicting **H1**. However, real-world time series often involve additional complexities, such as noise, missing data, and greater variability. To ensure the robustness of these findings, Part III extends the analysis to real-world data from the M4 dataset. This extension aims to provide a more comprehensive evaluation of the hypotheses and corroborate the results from the simulation study.

3 Part III: Empirical Application with the Hourly and Daily Series of the M4 Dataset

Part II identified the Theta-MLP method as the best-performing model across all DGPs in terms of OWA, providing strong evidence, supported by the DM test, that more complex forecasting

methods generally achieve higher accuracy compared to simpler ones. Part II also emphasized the role of trend damping in improving forecasting accuracy on average, as confirmed by the DM test, although it highlighted limitations when applied to strongly trending series, where it led to systematic under-forecasting. Conversely, the Comb method consistently underperformed relative to individual ES models in Part II, with the DM test confirming that these differences in accuracy were statistically significant at the 95% level. These findings challenge the assumption that combining methods always results in improved forecasting accuracy, underscoring the importance of aligning forecasting techniques with the specific characteristics of the data.

While the simulation study in Part II provided valuable insights within a controlled environment, it does not fully capture the complexities of real-world time series, such as noise, structural breaks, and variability. Building on the controlled simulations in Part II, Part III extends the analysis to real-world data from the M4 Competition. The M4 dataset, encompassing 100,000 time series across various domains and frequencies, provides a valuable opportunity to evaluate forecasting methods under diverse, real-world conditions. This chapter focuses specifically on the hourly and daily series, which differ significantly in their structural properties, offering a deeper understanding of how forecasting methods perform under different scenarios.

The hourly series, comprising 414 time series, are predominantly characterized by moderate trends and strong seasonality. These characteristics make them well-suited for assessing methods that emphasize seasonal components, such as Seasonal Naïve. In contrast, the daily series, consisting of 4,227 time series, exhibit very strong trends and negligible seasonality, favoring trend-focused approaches like Holt ES. To quantify these distinctions, the Seasonal-Trend decomposition based on Loess (STL) is applied to calculate the strengths of trend (F_T) and seasonality (F_S), providing a detailed understanding of the data's underlying structure.

This chapter begins by introducing the M4 dataset, outlining its composition and detailing the rationale for focusing on the hourly and daily series. It explores the distinct structural characteristics of these series and their implications for forecasting. The performance of various statistical benchmarks and the Theta-MLP method is then analyzed, with a focus on understanding how different methods align with the unique properties of the hourly and daily series. Through this analysis, the chapter bridges the gap between theoretical simulations and practical applications, providing insights into the effectiveness of forecasting methods in real-world scenarios.

3.1 M4 Dataset: Hourly and Daily Series

The M4 dataset was created on December 28, 2017, with Professor Makridakis selecting a seed number to determine the sample of time series used in the M4 Competition. To avoid issues with error measures, each series was scaled by adding a constant to ensure a minimum value of 10, which was necessary for 29 instances across the dataset. This adjustment eliminated negative values or observations below 10. Additionally, identifying information, such as the starting dates of the series, was removed to ensure objectivity, with dates only revealed after the competition ended (Assimakopoulos *et al.*, 2020).

The dataset is categorized by data frequency and application domain, with six data frequencies: low-frequency (yearly, quarterly, and monthly) and high-frequency (weekly, daily, and hourly). It also spans six application domains: Micro, Industry, Macro, Finance, Demographic, and Other (Hyndman, 2020). The distribution of series by frequency and domain reflects the practical forecasting needs of organizations and the strategic importance of each application (M4 Team, 2018). Forecasting horizons were designed to match typical decision-making timelines: yearly data supports long-term planning, while high-frequency data addresses short-term operational needs spanning hours to weeks (Assimakopoulos *et al.*, 2020).

Consistent with earlier M Competitions, low-volume and intermittent series were excluded from the M4 dataset (Assimakopoulos *et al.*, 2020). This exclusion ensured continuity and addressed

challenges posed by zero values in non-continuous series. Participants were required to provide additional forecasts for the series: six for yearly data, eight for quarterly data, and 18 for monthly data (M4 Team, 2018). High-frequency data required 13 forecasts for weekly series and 14 and 18 forecasts for daily and hourly series, respectively (M4 Team, 2018). The dataset was publicly released on December 31, 2017, marking the start of the competition. Initially hosted on the M4 website, it was later included in the *M4comp2018* R package and the M4 GitHub repository (M4 Team, 2018). The training set was accessible from the competition's start, while the test set remained confidential until the competition concluded (Hyndman, 2020). Minimum observation requirements ranged from 13 for yearly series to 700 for hourly series. Compared to the M3 dataset, the M4 dataset includes longer time series, providing more data for training complex models (Assimakopoulos *et al.*, 2020).

For the proposed Theta-MLP hybrid algorithm, which combines the extensions of the Classical Theta method with a Multilayer Perceptron (MLP), this presents both opportunities and challenges. Opportunities arise from the availability of diverse series, enabling sufficient training for the MLP component. However, the iterative training of the MLP significantly increases computational demands. Applying this method to the full M4 dataset of 100,000 time series would result in impractically long runtimes, potentially requiring several weeks. To ensure feasibility, the analysis focuses on a smaller subset of the dataset.

Table 5: Number of M4 series per data frequency and domain.

Time interval between successive observations	**Micro**	**Industry**	**Macro**	**Finance**	**Demographic**	**Other**	**Total**
Yearly	6,538	3,716	3,903	6,519	1,088	1,236	23,000
Quarterly	6,020	4,637	5,315	5,305	1,858	865	24,000
Monthly	10,975	10,017	10,016	10,987	5,728	277	48,000
Weekly	112	6	41	164	24	12	359
Daily	1,476	422	127	1,559	10	633	4,227
Hourly	0	0	0	0	0	414	414
Total	25,121	18,798	19,402	24,534	8,708	3,437	100,000

Due to constraints, the analysis focuses on the hourly and daily M4 data, consisting of 414 and 4,227 series, respectively. To evaluate the strength of trend and seasonality in these time series, the basic decomposition model from Chapter 3 of Makridakis *et al.* (1979), further elaborated by Hyndman *et al.* (2006), is used. This decomposition is performed using the Seasonal-Trend decomposition procedure based on Loess (STL) by Cleveland *et al.* (1990), formulated as follows (Hyndman & Athanasopoulos, 2006):

$$y_t = T_t + S_t + R_t \tag{54}$$

Here, T_t represents the trend component, S_t represents the seasonal component, and R_t represents the remainder component (Hyndman *et al.*, 2006). In data exhibiting strong trends, the seasonally adjusted series, $S_t + R_t$, should exhibit much greater variation than the remainder component, R_t, alone (Hyndman & Athanasopoulos, 2018). As a result, the ratio $\frac{\text{Var}(R_t)}{\text{Var}(T_t+R_t)}$ should be relatively small, with $F_T \approx 1$. Conversely, for data with little or no trends, the variances should be approximately equal, leading to $F_T \approx 0$ (Hyndman & Athanasopoulos, 2018). The strength of the trend is defined as follows (Hyndman & Athanasopoulos, 2018):

$$F_T = \max\left(0, 1 - \frac{\text{Var}(R_t)}{\text{Var}(T_t + R_t)}\right) \tag{55}$$

Where $0 \leq F_T \leq 1$, with values of F_T approaching zero indicating little or no trend, and values approaching one signifying a strong trend (Hyndman *et al.*, 2006). Occasionally, the variance of the remainder may even exceed the variance of the seasonally adjusted data. Therefore, following Hyndman *et al.* (2006), the minimum possible value of F_T is set to zero.

In a similar manner, the strength of seasonality is defined based on the detrended data, $S_t + R_t$, rather than the seasonally adjusted data (Hyndman & Athanasopoulos, 2018):

$$F_S = \max\left(0, 1 - \frac{\text{Var}(R_t)}{\text{Var}(S_t + R_t)}\right) \tag{56}$$

In time series with strong seasonality, $\text{Var}(S_t+R_t)$ will be significantly larger than $\text{Var}(R_t)$, leading to F_S being close to 1 (Hyndman & Athanasopoulos, 2018). Conversely, in series with little or no seasonality, the seasonal strength F_S will be close to 0 (Hyndman & Athanasopoulos, 2018). By applying these measures to both the hourly and daily series of the M4 dataset, it is possible to identify the series that exhibit the most pronounced trends or seasonality. The average strengths of the trends and seasonality across the hourly and daily series of the M4 dataset are provided in the following table:

Series	**Trend Strength** F_t	**Seasonal Strength** F_S
Hourly		
Min	0.004	0.001
1st Qu.	0.419	0.805
Mean	0.683	0.857
3rd Qu.	0.981	0.946
Max	0.999	0.999
Daily		
Min	0.339	0.000
1st Qu.	0.991	0.002
Mean	0.990	0.013
3rd Qu.	0.999	0.016
Max	1.000	0.238

Table 6: Seasonal and Trend Strengths across the Hourly and Daily Series of the M4 Competition.

Based on the results of the decomposition, it is evident that the hourly and daily series in the M4 dataset exhibit fundamentally different structures. The hourly series is predominantly characterized by strong seasonality and moderate trends. The trend strength (F_t) for the hourly series ranges from a minimum of 0.004 to a maximum of 0.999, indicating that some hourly series have negligible trends, while others exhibit nearly perfect trends. The mean trend strength (0.683) suggests that, on average, hourly series display moderate trends, while the 1st quartile (0.419) indicates that at least 25% of the hourly series exhibit weak or negligible trends. The seasonal strength (F_S), on the other hand, has a mean value of 0.857, emphasizing that seasonality is a dominant characteristic of the hourly series, with most series exhibiting strong seasonal patterns. The seasonal strength ranges from a minimum of 0.001 to a maximum of 0.999, reflecting a wide spectrum where some series show minimal or negligible seasonality, while others demonstrate almost perfect seasonal patterns. Furthermore, the first quartile of the seasonal strength (F_S) for hourly series is 0.805, indicating that even the weakest 25% of hourly series exhibit moderately strong seasonality.

In contrast, the daily series is predominantly characterized by very strong trends but weak seasonality. The trend strength in the daily series ranges from a minimum of 0.339 to a maximum of 1.000, indicating that all daily series exhibit at least moderate trends, with many displaying

perfect trends. The 1st quartile (0.991) and 3rd quartile (0.999) show that the vast majority of daily series have trend strengths close to 1.000, while the mean trend strength (0.990) confirms that trends overwhelmingly dominate the daily series. The seasonal strength of the daily series, on the other hand, varies from a minimum of 0.000 to a maximum of 0.238, indicating that most daily series lack significant seasonality. The 1st quartile (0.002) and 3rd quartile (0.016) confirm that at least 75% of the daily series exhibit minimal or negligible seasonality. Consequently, the mean seasonal strength (0.013) highlights that seasonality plays a minimal role in the daily series compared to trends.

The observed differences in trend and seasonal strengths between the hourly and daily series in the M4 dataset carry significant implications for the effectiveness of various forecasting methods. For instance, given the strong seasonality in the hourly series ($F_S = 0.857$ on average), forecasting models that emphasize the seasonal component—such as Seasonal Naïve or Naïve 2—are likely to perform well relative to more complex methods like Theta-MLP. This was exemplified in DGP 1 of Part II, which featured very strong seasonality ($F_S = 0.951$ on average). In this case, the Seasonal Naïve method achieved the second-highest average OWA across the simulations, trailing the Theta-MLP hybrid algorithm by 28.06%. Conversely, the daily series of the M4 dataset, dominated by very strong trends ($F_T = 0.990$ on average), suggest a preference for trend-focused models such as Holt or Damped ES. This was mirrored in the simulations of DGP 3, which exhibited an extremely strong trend ($F_T = 0.999$ on average). Here, Holt ES achieved the second-highest OWA score across the simulations, trailing the Theta-MLP method by 22.0%.

By comparing the forecasting accuracy of statistical benchmarks and the Theta-MLP method across simulated DGPs and real-world hourly and daily data from the M4 Competition, while accounting for the underlying properties of the time series, valuable insights can be gained into the effectiveness of different methods across various scenarios, ensuring robust and generalizable conclusions.

3.2 Results of the hourly series of the M4 Competition

The following table presents the average forecasting performance for the hourly time series in the M4 dataset, evaluated across the eight statistical benchmarks and the Theta-MLP method:

Table 7: The performance of Naïve, Seasonal Naïve, Naïve 2, SES, Holt, Damped, Theta, Comb, and Theta-MLP methods evaluated using average sMAPE, MASE, and OWA for the hourly series of the M4 dataset.

Method	**sMAPE**	**MASE**	**OWA**	**Rank**
Naïve	43.003	11.608	3.593	9
sNaïve	**13.912**	**1.193**	**0.628**	**1**
Naïve2	18.383	2.395	1.000	4
SES	18.094	2.385	0.990	3
Holt	29.474	9.380	2.760	8
Damped	19.277	2.947	1.140	6
Theta	18.138	2.455	1.006	5
Comb	22.114	4.585	1.559	7
Theta-MLP	17.531	1.807	0.854	2

The results from the hourly series of the M4 Competition provide strong support for **H2**, which posits that more complex forecasting methods, such as hybrid approaches, can achieve greater accuracy compared to simpler methods. The Theta-MLP method, a hybrid approach that combines extensions of the Classical Theta method with an MLP correction term, achieved an average OWA of 0.854, securing the second overall rank. This performance demonstrates substantial improvements in OWA compared to simpler methods: a 76.23% improvement over the Naïve method

(OWA: 3.593), a 69.06% improvement over Holt ES (OWA: 2.760), a 25.09% improvement over Damped ES (OWA: 1.140), a 14.60% improvement over Naïve 2 (OWA: 1.000), and a 13.74% improvement over SES (OWA: 0.990). Additionally, the DM test rejects the null hypothesis of equal forecasting accuracy between the Theta-MLP method and these statistical benchmarks, confirming that these differences are statistically significant at the 95% confidence level.

Moreover, the Theta-MLP method outperforms the simpler Classical Theta method (OWA: 1.006) introduced by Assimakopoulos and Nikolopoulos (2000)—which assumes a linear trend, additive components, and multiplicative seasonality without an MLP correction term—by 15.11% in terms of OWA. The results of the DM test also reject the null hypothesis of equal forecasting accuracy between the Theta-MLP method and the Classical Theta method at the 95% significance level, indicating that the additional computational complexity of the Theta-MLP method is justified by its improved performance.

Although the Seasonal Naïve method outperforms the Theta-MLP method by 26.45% in terms of OWA, the results of the DM test do not reject the null hypothesis of equal forecasting accuracy at the 95% significance level. The strong performance of the Seasonal Naïve method is expected due to the hourly series in the M4 dataset exhibiting strong and consistent seasonality, with a high average seasonal strength ($F_S = 0.857$ on average). The Seasonal Naïve method directly utilizes this seasonality by repeating past seasonal patterns, making it particularly effective when trends are moderate or weak ($F_T = 0.683$ on average). This aligns with earlier simulation results from DGP 1, where the Seasonal Naïve method achieved the second-highest average OWA, demonstrating its effectiveness in datasets dominated by strong seasonality.

Furthermore, the results from the hourly series of the M4 Competition reveal that Damped ES outperforms Holt ES by 58.70% in terms of OWA, supporting **H3**, which posits that damping the trend improves forecasting accuracy on average but may lead to under-forecasting in strongly trending series. The results of the DM test also reject the null hypothesis of equal forecasting accuracy between Damped and Holt ES, indicating that the observed differences in accuracy are statistically significant at the 95% confidence level. By incorporating a damping parameter, Damped ES gradually reduces the influence of the trend over the forecast horizon, effectively mitigating the risk of over-forecasting moderate trends. This approach aligns well with the characteristics of the hourly series, which exhibit an average moderate trend strength ($F_t = 0.683$ on average). In contrast, Holt ES assumes a linear and unrestricted trend, making it more prone to over-forecasting in series with weaker or less consistent trends.

Consequently, the results do not support **H1**, which asserts that combining multiple forecasting methods enhances forecasting accuracy on average compared to the individual methods being combined, as evidenced by the substantial underperformance of the Comb method (OWA: 1.559), which ranks seventh. While combining multiple forecasting methods is theoretically appealing—since no single method can fully capture all time series patterns—the Comb method's performance did not align with this premise. It performs 57.5% worse than SES (OWA: 0.990) and 36.1% worse than the Damped ES method (OWA: 1.140). Furthermore, the results of the DM test reject the null hypothesis of equal forecasting accuracy between the Comb method and the individual exponential smoothing methods at the 95% significance level. These findings challenge the validity of **H1** and highlight the limitations of simple, unweighted ensembling in this context.

3.3 Results of the daily series of the M4 Competition

The following table presents the average forecasting performance for the daily time series in the M4 dataset, evaluated across the eight statistical benchmarks and the Theta-MLP method:

Table 8: The performance of Naïve, Seasonal Naïve, Naïve 2, SES, Holt, Damped, Theta, Comb, and Theta-MLP methods evaluated using average sMAPE, MASE, and OWA for the daily series of the M4 dataset.

Method	**sMAPE**	**MASE**	**OWA**	**Rank**
Naïve	3.045	3.278	1.000	6
sNaïve	3.045	3.278	1.000	6
Naïve2	3.045	3.278	1.000	6
SES	3.045	3.281	1.000	6
Holt	3.070	3.231	0.997	4
Damped	3.063	3.234	0.996	3
Theta	3.053	3.262	0.999	5
Comb	**2.985**	**3.205**	**0.979**	**1**
Theta-MLP	3.023	3.231	0.989	2

The results from the daily series of the M4 dataset strongly support **H1**. This is evidenced by the Comb method (OWA: 0.979) achieving the highest overall rank for the daily series. While the improvements in OWA percentages are relatively modest—1.81% over Holt (OWA: 0.997), 1.71% over Damped (OWA: 0.996), and 2.10% over SES (OWA: 1.000)—the DM test confirms that the differences in forecasting errors between the Comb method and the individual ES methods are statistically significant at the 95% confidence level. This finding contradicts the earlier observations on combining models using raw ensembles. In both the hourly M4 series and the simulated hourly DGPs in Part II, the Comb method consistently underperformed relative to individual ES methods, and the DM test confirmed that the differences in forecasting accuracy are statistically significant.

Furthermore, the Damped and Holt ES methods perform relatively well, ranking third and fourth overall, respectively. This outcome is expected, as the daily series of the M4 dataset are characterized by strong trends ($F_T = 0.990$ on average) and weak seasonality ($F_S = 0.013$ on average), conditions under which trend-emphasizing methods tend to excel. Both methods achieve significantly lower OWA scores compared to models that primarily emphasize the seasonal component. Specifically, the Damped ES outperforms both Naïve 2 and Seasonal Naïve by 0.40% in terms of OWA, while the Holt ES outperforms these methods by 0.30% in terms of OWA for the daily series of the M4 dataset. Additionally, the DM test rejects the null hypothesis of equal forecasting accuracy between the Damped and Holt ES methods and the seasonal component-focused models at the 95% significance level, confirming that the observed differences in accuracy are statistically significant. These findings are consistent with the results from DGP 3, which also features strong trends ($F_T = 0.999$ on average). In that context, the Damped and Holt ES methods achieved substantially lower OWA scores compared to the seasonal component-focused models. Specifically, Damped ES outperformed Naïve 2 by 27.8% and Seasonal Naïve by 41.5%, while Holt ES outperformed Naïve 2 by 36.4% and Seasonal Naïve by 48.4% in terms of OWA in DGP 3.

Nonetheless, both the Damped and Holt ES methods perform very similarly in the daily series of the M4 dataset, with slight differences depending on the performance metric. For instance, the Damped ES outperforms Holt ES by 0.23% in terms of sMAPE, while Holt ES performs 0.9% better in terms of MASE. Nevertheless, based on OWA—the primary measure for evaluating the performance of different forecasting methods and testing the proposed hypotheses in this thesis—the Damped ES outperforms Holt ES by 0.10%.

However, the DM test does not reject the null hypothesis of no significant difference in forecasting accuracy between the two methods at the 95% significance level. These results challenge **H3**, and this outcome is unexpected, given that the daily series in the M4 dataset are characterized by very strong trends—a condition where Holt ES would typically be expected to outperform Damped ES, as the damping mechanism may underestimate the trend over longer horizons. By contrast, in DGP 3, which was explicitly designed to test methods under strong trending conditions, the

Damped ES outperformed Holt ES by 13.5% in terms of OWA, with the DM test rejecting the null hypothesis of equal forecasting accuracy.

Lastly, the results from the daily series of the M4 dataset provide strong support for **H2**. The Theta-MLP method achieves an average OWA of 0.989, securing second place overall. It outperforms the simpler Naïve, Naïve 2, Seasonal Naïve, and SES methods by 1.10%, as well as Holt ES and Damped ES by 0.80% and 0.70%, respectively. The DM test at the 95% significance level rejects the null hypothesis of equal forecasting accuracy between the Theta-MLP method and these statistical benchmarks. By building on the Classical Theta method, the Theta-MLP method improves OWA by 1.00%, with the DM test confirming that this improvement is statistically significant. These findings highlight that the accuracy gains achieved by incorporating machine learning techniques in hybrid approaches justify the additional computational expenses. Although the Comb method outperforms the Theta-MLP method by 1.01% in terms of OWA for the daily series, the DM test fails to reject the null hypothesis, indicating that the difference in forecasting accuracy between the Theta-MLP and Comb methods is not statistically significant at the 95% significance level.

3.4 Findings across the hourly and daily series of the M4 Competition

The analysis of the hourly and daily series in the M4 dataset highlights important distinctions in the performance of different forecasting methods, which are largely driven by the underlying characteristics of the series. To quantify these characteristics, STL decomposition was applied to calculate the strengths of the trend and seasonality components. Specifically, the hourly series exhibited strong seasonality ($F_S = 0.857$ on average) and moderate trends ($F_T = 0.683$ on average). In contrast, the daily series displayed very strong trends ($F_T = 0.990$ on average) and weak seasonality ($F_S = 0.013$ on average). These differences are critical because they significantly influence the relative performance of forecasting models and provide valuable insights into the hypotheses tested in this study.

The findings from the hourly and daily series of the M4 Competition strongly support **H2**, which posits that more complex forecasting methods outperform simpler ones. For both series, the Theta-MLP method consistently outperformed most simpler models in terms of OWA. Moreover, the DM test rejected the null hypothesis of equal forecasting accuracy between the Theta-MLP method and various statistical benchmarks at the 95% significance level. Importantly, the Theta-MLP method demonstrated substantial improvements in OWA compared to the Classical Theta method, with these differences confirmed as statistically significant based on the DM test. Although the Theta-MLP method did not achieve the highest overall rank, the DM test failed to reject the null hypothesis of equal forecasting accuracy between the Theta-MLP method and the top-ranking method in both series. These results demonstrate that adopting more sophisticated methods, such as the Theta-MLP hybrid algorithm, is justified despite the computational costs associated with training machine learning models.

Regarding **H3**, which suggests that damping the trend improves forecasting accuracy on average but may lead to under-forecasting in strongly trending series, the findings are mixed. For the hourly series, characterized by moderate trends, Damped ES outperformed Holt ES by 58.70% in terms of OWA. This result aligns with the hypothesis, as the damping parameter effectively mitigated the risk of over-forecasting moderate trends. Furthermore, the DM test confirmed the statistical significance of this finding by rejecting the null hypothesis of equal forecasting accuracy between the two methods.

In contrast, the daily series, dominated by very strong trends, presented a different scenario. Here, it was expected that Holt ES would outperform Damped ES due to its unrestricted trend component better capturing the consistent and persistent trend behavior. However, the results showed that both methods performed comparably. While Damped ES slightly outperformed Holt ES in terms

of sMAPE and OWA, Holt ES achieved better results in MASE. Moreover, the DM test did not reject the null hypothesis of equal forecasting accuracy between the two methods. This unexpected result suggests that factors beyond trend characteristics may influence the relative effectiveness of these methods, highlighting the need for further research to explore the interaction between trend properties and damping mechanisms.

The findings provide mixed evidence for **H1**, with support observed for the daily series but not for the hourly series. In the daily series, the Comb method achieved the best performance, attaining the lowest OWA (0.979) and outperforming all individual ES methods, including Damped (OWA: 0.996), Holt (OWA: 0.997), and SES (OWA: 1.000). The DM test confirmed that these differences—except between the Comb method and the Theta-MLP method—were statistically significant at the 95% confidence level.

However, for the hourly series, the results were less favorable. The Comb method (OWA: 1.559) underperformed compared to the individual ES models, performing 57.5% worse than SES (OWA: 0.990) and 36.1% worse than Damped ES (OWA: 1.140), ranking seventh overall. The DM test further confirmed statistically significant differences in accuracy between the Comb method and the individual ES models. These findings underscore the potential limitations of unweighted combinations and suggest that the effectiveness of combining methods is highly dependent on the characteristics of the time series being forecasted.

In conclusion, the findings from the hourly and daily series of the M4 dataset provide robust support for **H2**, partial support for **H3**, and mixed evidence for **H1**. While the Theta-MLP method demonstrated versatility and accuracy across both series—attributable to the extensions of the Classical Theta model and the complex MLP component—the analysis also highlights the nuanced interplay between trend and seasonality in determining the effectiveness of forecasting methods. The varying performances of other methods underscore the importance of aligning model selection with the specific characteristics of the data. These findings reinforce the critical role of understanding the underlying structure of time series data in advancing forecasting accuracy and provide a strong foundation for future research into hybrid approaches such as the Theta-MLP.

4 Conclusion

Accurate forecasting is essential for informed decision-making across different sectors, enabling effective planning, efficient resource management, and risk mitigation. From guiding economic policies to optimizing business strategies, the ability to predict future outcomes reliably reduces costs, enhances customer service, and boosts overall efficiency (Makridakis & Hibon, 2000). Forecasting competitions, such as the Makridakis series (M Competitions), have significantly advanced the theory and practice of forecasting by providing empirical evidence on how to improve forecasting accuracy and applying these insights to refine forecasting methodologies (Assimakopoulos *et al.*, 2020).

This thesis aimed to identify key factors that enhance forecasting accuracy by conducting a comprehensive evaluation of various methods, focusing on model complexity, trend damping, and combining methods. Central to this research was the development of the Theta-MLP method, a hybrid algorithm that combines extensions of the Classical Theta method with a multilayer perceptron (MLP) to address non-linear residuals effectively. By systematically testing these approaches on both simulated hourly DGPs—designed to capture varying levels of trend complexity and seasonality—and real-world data from the hourly and daily series of the M4 Competition, the study provides valuable insights into how specific conditions and data characteristics influence the relative performance of forecasting methods.

Guided by the research question, *How do model complexity, trend damping, and combining methods influence forecasting accuracy across different time series data?*, this thesis rigorously explored

three hypotheses. These addressed whether combining methods outperforms individual methods on average (**H1**), whether more complex methods lead to greater forecasting accuracy compared to simpler ones (**H2**), and whether trend damping improves accuracy on average across time series but results in under-forecasting for strongly trending series (**H3**). The hypotheses were tested using a dual approach of simulations and empirical analysis, and the results consistently highlighted the importance of aligning forecasting methods with the underlying characteristics of the data.

Regarding model complexity, the findings provide robust support for **H2**, highlighting that computational resources were effectively leveraged to achieve superior forecasting accuracy. Across all three hourly DGPs in the simulation study of Part II, the Theta-MLP method consistently delivered the lowest OWA, demonstrating significant accuracy improvements over simpler benchmarks and the Classical Theta model on which it is based. For instance, the Theta-MLP method outperformed the Naïve model by 49.09% in terms of OWA for DGP 1, by 24.84% for DGP 2, and by 48.65% for DGP 3. Moreover, the DM test consistently demonstrated that the observed improvements in OWA achieved by the Theta-MLP method in Part II were not due to random variation but were statistically significant across all DGPs at a 95% confidence level.

The empirical analysis in Part III further reinforced these findings, as the Theta-MLP method proved highly effective on both the hourly and daily series of the M4 Competition. For the hourly series, which exhibited strong seasonality ($F_S = 0.857$ on average) and moderate trends ($F_T = 0.683$ on average), Theta-MLP achieved an OWA of 0.854, placing second overall. It delivered notable accuracy gains, including improvements of 76.23% over Naïve, 69.06% over Holt, 25.09% over Damped ES, and 15.11% over Classical Theta—all statistically significant as confirmed by the DM test. Furthermore, although the Seasonal Naïve method achieved the lowest OWA (0.628) for the hourly series, the DM test did not indicate a statistically significant difference in forecasting accuracy between the Theta-MLP and Seasonal Naïve methods at the 95% confidence level.

For the daily series, which exhibited very strong trends ($F_T = 0.990$ on average) and negligible seasonality ($F_S = 0.013$ on average), the Theta-MLP method again secured second place with an OWA of 0.989. While the improvements in OWA percentages were relatively modest—outperforming simpler benchmarks by up to 1.1% and surpassing Classical Theta by 1.0%—the results of the DM test confirmed that these differences were statistically significant. Moreover, while the Comb method achieved the lowest OWA (0.979) for the daily series, the DM test failed to reject the null hypothesis of equal forecasting accuracy between the Theta-MLP and Comb methods, indicating that the observed difference was not statistically significant at the 95% significance level. Altogether, despite its computational demands, the Theta-MLP method demonstrated that hybrid approaches integrating statistical and machine learning techniques offer substantial accuracy gains, justifying the costs of additional complexity.

The findings regarding **H3**—which posits that damping the trend generally improves forecasting accuracy but may lead to under-forecasting in strongly trending series—are nuanced and depend on the characteristics of the data. The simulation study in Part II demonstrated strong support for this hypothesis across the hourly DGPs. For DGP 1, which featured a linear trend and additive seasonality, the Damped ES method underperformed relative to Holt ES by 15.1% in terms of OWA, with the DM test confirming that this difference was statistically significant at the 95% confidence level. Similarly, in DGP 3, characterized by a pronounced exponential trend, Damped ES under-forecasted the trend, leading to a 13.5% reduction in accuracy compared to Holt ES. These results underscore the limitations of damping mechanisms in capturing strongly trending patterns, as damping inherently reduces the long-term influence of the trend component. Conversely, for DGP 2, which featured a damped linear trend and multiplicative seasonality, Damped ES outperformed Holt ES by 17.4% in terms of OWA, demonstrating its ability to align more effectively with the underlying characteristics of the data. The DM test confirmed that this improvement was statistically significant. These findings highlight the value of damping mechanisms in mitigating over-forecasting for series where the trend naturally diminishes over time.

The empirical results from Part III further emphasize the importance of aligning trend-damping methods with the specific characteristics of the time series. For the hourly series, featuring strong

seasonality ($F_S = 0.857$ on average) and moderate trends ($F_T = 0.683$ on average), Damped ES outperformed Holt ES by 58.70% in terms of OWA, with the DM test rejecting the null hypothesis of equal accuracy. This result reinforces the utility of damping in preventing over-forecasting for series with moderate trends. However, for the daily series, dominated by very strong trends ($F_T = 0.990$ on average) and negligible seasonality ($F_S = 0.013$ on average), the expected advantage of Holt ES over Damped ES was not observed. Both methods performed comparably, and the DM test failed to detect significant differences in forecasting accuracy. This unexpected outcome suggests that factors beyond trend strength, such as noise or structural breaks, may influence the relative performance of these methods. Altogether, the results for **H3** underscore the importance of understanding the nature of the trend in time series forecasting. While damping mechanisms are highly effective for mitigating over-forecasting in series with moderate or damped trends, their application in strongly trending series requires careful consideration to avoid systematic under-forecasting. These findings highlight the need for further research into adaptive trend-damping methods that can dynamically adjust to varying trend strengths.

Furthermore, the findings for **H1**, which argues that combining forecasting methods enhances accuracy compared to the individual methods, present a nuanced picture. In Part II, the Comb method, a simple arithmetic average of SES, Holt, and Damped ES, consistently underperformed individual methods across all DGPs. For instance, it lagged behind Holt by 9.3% in OWA for DGP 1 and fell 30.3% short of SES in DGP 2. It also underperformed Damped ES by 8.0% and Holt ES by 22.6% in terms of OWA for DGP 3. These differences were statistically significant as confirmed by the DM test. Similarly, in Part III, the Comb method ranked seventh for the hourly series (OWA: 1.559) and delivered results 36.1% worse than Damped ES. However, for the daily series, where trends dominated ($F_T = 0.990$ on average), Comb achieved the best OWA (0.979), marginally outperforming individual methods. The DM test confirmed statistical significance for these differences, except between Comb and Theta-MLP.

These results should not be interpreted as a critique of combining methods but rather as a reflection of the simplicity of raw ensembling, as exemplified by the Comb method. Future research could explore weighted combination schemes, where weights are assigned based on model performance, or adaptive ensembling approaches that adjust dynamically to data characteristics. Such methods could address the limitations of raw ensembling while leveraging the potential benefits of combining forecasting models.

Through a rigorous blend of simulation studies and empirical analyses, this thesis makes significant contributions to the field of forecasting by introducing the hybrid Theta-MLP algorithm—the second true hybrid method applied to the M4 dataset—and investigating the impact of model complexity, trend damping, and combining methods on predictive accuracy. Using both simulations and real-world data from the M4 Competition, the findings for **H2** illustrate the strength of integrating statistical methods with machine learning, as demonstrated by the superior performance of the Theta-MLP method. Additionally, the results for **H1** reveal the shortcomings of simple ensembling approaches and underscore the necessity for more advanced techniques to effectively combine forecasting methods. The results for **H3** emphasize the importance of aligning trend-damping mechanisms with the characteristics of the underlying data, demonstrating their effectiveness in mitigating over-forecasting for moderate trends while cautioning against under-forecasting in strongly trending series. Altogether, these insights address key gaps in forecasting theory and aim to promote the development of novel hybrid approaches.

5 Discussion

While the results across both the simulated DGPs and the M4 dataset provide valuable insights, two key limitations must be acknowledged. The first limitation pertains to the subsets of M4 data that were analyzed. Although the hourly series, comprising 414 time series, and the daily series, consisting of 4,227 time series, are representative of these respective frequencies, they may

not capture the full diversity of patterns present across all 100,000 series in the M4 dataset. This limitation inevitably affects the generalizability of the conclusions drawn. For instance, while the Comb method underperformed in the context of hourly series, broader analyses of the full M4 dataset have demonstrated that Comb methods can outperform individual exponential smoothing (ES) models across a wider range of frequencies and domains (Assimakopoulos *et al.*, 2020). As a result, conclusions related to **H1** (that combination methods do not consistently outperform individual models) may not hold across the entire M4 dataset. Nevertheless, the findings supporting **H2** (that more complex hybrid models achieve higher accuracy) and **H3** (that damping trends improves forecasting accuracy on average but may lead to under-forecasting in strongly trending series) are consistent with prior research and are expected to hold across other data frequencies (Assimakopoulos *et al.*, 2020).

The second limitation concerns the diversity of models considered in the study. The results suggest that the underperformance of the Comb method in the hourly series is likely due to its simplicity, as it combines relatively basic ES models. The study did not investigate more sophisticated combination methods, such as weighted averages based on information criteria or ensemble learning approaches that leverage more complex models. Employing advanced combination strategies that incorporate additional model-specific information or performance metrics could lead to different conclusions regarding **H1**. Therefore, the results should not be interpreted as a critique of combination methods in general but rather as a reflection of the simplicity of the particular Comb method used in this study. More complex combinations of models could offer better forecasting performance, aligning more closely with the finding that more complex hybrid models achieve higher accuracy.

It is also important to note that, prior to this study, only one hybrid model had been applied to the M4 Competition series: the ES-RNN model developed by Smyl (2018), which won the competition. Inspired by Smyl's success, this study introduced the Theta-MLP method as a novel hybrid approach, combining extensions of the Classical Theta method with a multilayer perceptron (MLP) to address non-linear residuals. The results of this study highlight the potential of hybrid models to achieve superior forecasting accuracy, particularly in capturing complex, non-linear patterns in time series data.

In light of these limitations and findings, several avenues for future research are recommended. First, extending the analysis to include additional frequencies from the M4 dataset, such as yearly, quarterly, and monthly data, as well as analyzing different business domains, such as finance and micro, would provide a more comprehensive evaluation of the performance of the Theta-MLP method and other forecasting models. This broader analysis would help validate the findings across different types of time series data, ensuring that the conclusions are not confined to high-frequency data alone. Second, future studies should incorporate more diverse and complex models, particularly in the realm of combination and hybrid methods. The inclusion of advanced combination strategies, such as weighted ensembles or machine learning-based hybrid models, would offer a deeper understanding of how various methods perform in different scenarios. By pursuing these avenues, future research could build on the insights of this study and contribute to the ongoing development of forecasting theory.

6 References

Almeida, L. B. (1997). Multilayer perceptrons. In M.A. Arbib (Ed.), *Handbook of Neural Computation* (C1.2:1-C1.2:6). Oxford University Press.

Armstrong, J. S., Green, K. C., & Graefe, A. (2015). Golden rule of forecasting: Be conservative. *Journal of Business Research, 68*(8), 1717–1731. https://doi.org/10.1016/j.jbusres.2015.03.031

Assimakopoulos, V., Spiliotis, E., & Makridakis, S. (2020). The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting, 36*(1), 54–74. https://doi.org/10.1016/j.ijforecast.2019.04.014

Assimakopoulos, V., & Nikolopoulos, K. (2000). The theta model: a decomposition approach to forecasting. *International Journal of Forecasting, 16*(4), 521–530. https://doi.org/10.1016/S0169-2070(00)00066-2

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *The Journal of the Operational Research Society, 20*(4), 451–468.

Bishop, C.M. (1995) Neural networks for pattern recognition. *ford University Press, Oxford*,482

Brown, R. G. (1959). Statistical forecasting for inventory control. McGraw-Hill.

Costa, L., Guerreiro, M., Puchta, E., Tadano, Y. S., Alves, T. A., Kaster, M., & Siqueira, H. V. (2023). Multilayer Perceptron. In *Artificial Neural Networks*, Universidade Tecnológica Federal do Paraná, 105.

Chan, F., & Pauwels, L. L. (2018). Some theoretical results on forecast combinations. *International Journal of Forecasting, 34*(1), 64–74.

Chan, K. Y., Abu-Salih, B., Qaddoura, R., Al-Zoubi, A. M., Palade, V., Pham, D. S., Del Ser, J., & Muhammad, K. (2023). Deep neural networks in the cloud: Review, applications, challenges and research directions. *Neurocomputing, 545*, 126237. https://doi.org/10.1016/j.neucom.2023.126237

Clemen, R. (1989). Combining forecasts: a review and annotated bibliography with discussion. *International Journal of Forecasting*, 5, 559–584.

Claeskens, G., Magnus, J. R., Vasnev, A. L., & Wang, W. (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting, 32*(3), 754–762. https://doi.org/10.1016/j.ijforecast.2015.12.005

Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). STL: A seasonal-trend decomposition procedure based on Loess. *Journal of Official Statistics, 6*(1), 3–73.

Constantinidou, C., Nikolopoulos, K., Bougioukos, N., Tsiafa, E., Petropoulos, F., & Assimakopoulos, V. (2001). A neural network approach for the theta model. *Information Engineering, Lecture Notes in Information Technology, 25*(1), 116–120.

Constantinidou, C., Nikolopoulos, K., Bougioukos, N., Tsiafa, E., Petropoulos, F., & Assimakopoulos, V. (2012). A neural network approach for the theta model. *Information Engineering, Lecture Notes in Information Technology, 25*, 116–120.

Crone, S. F., Hibon, M., & Nikolopoulos, K. (2011). Advances in forecasting with neural networks? Empirical evidence from the NN3 competition on time series prediction. *International Journal of Forecasting, 27*(3), 635–660. https://doi.org/10.1016/j.ijforecast.2011.04.001

Dudek, G. (2019). Short-term load forecasting using Theta method. E3S Web of Conferences, 84, 01004. https://doi.org/10.1051/e3sconf/20198401004

Fiorucci, J. A. (2016). Time series forecasting: Advances on Theta method. São Carlos: UFS-Car.

Fiorucci, J. A., Pellegrini, T. R., Louzada, F., & Petropoulos, F. (2015). The Optimised Theta Method. Retrieved from https://arxiv.org/abs/1503.03529

Fiorucci, J. A., Pellegrini, T. R., Louzada, F., Petropoulos, F., & Koehler, A. B. (2016). Models for optimising the theta method and their relationship to state space models. *International Journal of Forecasting, 32*(3), 1151-1161. https://doi.org/10.1016/j.ijforecast.2016.05.013

Gardner, E. S., Jr., & McKenzie, E. (1985). Forecasting trends in time series. *Management Science, 31*(10), 1237–1246. https://doi.org/10.1287/mnsc.31.10.1237

Holt, C. C. (1957). Forecasting seasonal and trends by exponentially weighted averages. *Office of Naval Research Research Memorandum, 52.*

Hyndman, R.J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice* (2nd ed.). OTexts. https://otexts.com/fpp2

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*(4), 679–688.

Hyndman, R. J., Koehler, A. B., Snyder, R. D., & Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting, 18*(3), 439-454. https://doi.org/10.1016/S0169-2070(01)00110-8

Hyndman, R. J., & Billah, B. (2003). Unmasking the Theta method. *International Journal of Forecasting, 19*(3), 287-290. https://doi.org/10.1016/S0169-2070(01)00143-1

Jaiswal, S. (2024). Multilayer Perceptrons in Machine Learning: A Comprehensive Guide. *TutorialsPoint.* https://www.tutorialspoint.com/multilayer-perceptrons-in-machine-learning

Jordan, J., Szpruch, L., Hussain, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S. N., & Weller, A. (2022). Synthetic data - what, why and how? The Alan Turing Institute and The Royal Society. *arXiv.* https://arxiv.org/abs/2205.03257v1

Lawler, G. F., & Limic, V. (2010). *Random Walk: A Modern Introduction.* Cambridge University Press.

Legaki, N.-Z., & Koutsouri, A. (2020). Theta - BoxCox: Deseasonalize data, apply Box-Cox transformation, and forecast with Theta. Method Description submitted to the National Technical University of Athens, Forecasting & Strategy Unit.

Makridakis, S. (1993). Accuracy measures: Theoretical and practical concerns. *International Journal of Forecasting, 9*(4), 527–529.

Makridakis, S., & Winkler, R. L. (1983). Averages of forecasts: some empirical results. *Management Science, 29*, 987–996.

Makridakis, S. G., Wheelwright, S. C., & Hyndman, R. J. (1998). *Forecasting: Methods and Applications* (3rd ed.). New York: Wiley.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018a). Statistical and machine learning forecasting methods: concerns and ways forward. *PLoS One, 13*(3), 1–26.

Nikolopoulos, K., Thomakos, D., Petropoulos, F., Litsa, A., & Assimakopoulos, V. (2012). Forecasting S&P 500 with the theta model. *International Journal of Financial Economics and Econometrics, 4*, 73–78.

Nikolopoulos, K. & Assimakopoulos, V. (2005). Fathoming the theta model. In *25th International Symposium on Forecasting, ISF*, San Antonio, Texas, USA. unknown.

Nikolopoulos, K., & Petropoulos, F. (2018). Forecasting for big data: Does suboptimality matter? *Computers & Operations Research, 98*, 322–329. https://doi.org/10.1016/j.cor.2017.05.007

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine, 38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

Osborne, J. W. (2010). Improving your data transformations: Applying Box-Cox transformations as a best practice. Retrieved from https://www.researchgate.net/publication/284261483

Pegels, C. C. (1969). Exponential forecasting: Some new variations. *Management Science, 15*(5), 311–315. http://dx.doi.org/10.1287/mnsc.15.5.311

Petropoulos, F., & Nikolopoulos, K. (2013). Optimizing Theta model for monthly data. In *ICAART 2013 - Proceedings of the 5th International Conference on Agents and Artificial Intelligence* (pp. 190–195). *ICAART 2013 - Proceedings of the 5th International Conference on Agents and Artificial Intelligence; Vol. 1.*

Sakia, R. M. (1992). The Box–Cox transformation technique: a review. *The Statistician, 41*(2), 169–178.

Smyl, S. (2020). A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting, 36*(1), 75–85.

Spiliotis, E., Assimakopoulos, V., & Makridakis, S. (2020). Generalizing the Theta method for automatic forecasting. *European Journal of Operational Research, 284*(2), 550-558. https://doi.org/10.1016/j.ejor.2020.01.007

Spiliotis, E., Assimakopoulos, V., & Nikolopoulos, K. (2019). Forecasting with a hybrid method utilizing data smoothing, a variation of the Theta method and shrinkage of seasonal factors. *International Journal of Production Economics, 209*, 92-102. https://doi.org/10.1016/j.ijpe.2018.01.020

Spitzer, F. (2001). *Principles of Random Walk* (2nd ed.). Springer.

Taud, H., & Mas, J.F. (2018). Multilayer Perceptron (MLP). In M.T. Camacho Olmedo, M. Paegelow, J.-F. Mas, & F. Escobar (Eds.), *Geomatic Approaches for Modeling Land Change Scenarios.*

Taylor, J. W. (2003). Exponential smoothing with a damped multiplicative trend. *International Journal of Forecasting, 19*(4), 715–725. https://doi.org/10.1016/S0169-2070(03)00003-7

Thomakos, D. D., & Nikolopoulos, K. (2015). Forecasting Multivariate Time Series with the Theta Method. International Journal of Forecasting, 34(3), 220-229. https://doi.org/10.1002/for.2334

Van Horne, J. C., & Parker, G. G. C. (1967). The Random-Walk Theory: An Empirical Test. *Financial Analysts Journal, 23*(6), 87–92. https://doi.org/10.2469/faj.v23.n6.87

Wichern, D. W., Makridakis, S., & Wheelwright, S. C. (1979). Forecasting: Methods and Applications. *Journal of the American Statistical Association, 74*(367), 733. https://doi.org/10.2307/2287014

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science, 6*(3), 324–342. https://doi.org/10.1287/mnsc.6.3.324

Xia, F., Liu, J., Nie, H., Fu, Y., Wan, L., & Kong, X. (2020). Random Walks: A Review of Algorithms and Applications. *IEEE Transactions on Emerging Topics in Computational Intelligence, 4*(2), 95–107. https://doi.org/10.1109/TETCI.2019.2952908

7 Appendix: Software and Computing Environment

To ensure the reproducibility of this study's results, this appendix provides an overview of the software environment and computing setup used for both Part II and Part III, including the specific versions of software and packages employed.

All analyses were conducted using **R version 4.4.2**, running on a **Windows 11 x64** operating system within the **Posit Workbench environment licensed to WHU Otto Beisheim School of Management**. The system operated in the **Europe/Amsterdam** timezone.

Key Software and Packages

The primary software and R packages used in this study are outlined below, along with their specific versions:

- **R version**: 4.4.0
- **Operating System**: Windows 11 x64
- **Forecasting and Machine Learning Packages**:
  - `tensorflow` (2.16.0)
  - `keras3` (1.2.0) [Configured within the WHU Posit Workbench environment]
  - `MAPA` (2.0.7)
  - `smooth` (4.0.1)
  - `greybox` (2.0.0)
  - `forecast` (8.22.0.9000)

  - `M4comp2018` (0.1.0)
- **Visualization and Data Processing Packages**:
  - `ggplot2` (3.5.1)
  - `RColorBrewer` (1.1-3)
  - `dplyr` (1.1.4)
- **Utility and Miscellaneous Packages**:
  - `reticulate` (1.38.0)
  - `rmarkdown` (2.27)

Reproducibility and Package Management

To replicate this study, users should ensure that the aforementioned R version and package versions are installed. Using the `renv` or `packrat` package can help manage the environment and ensure compatibility. Additionally, session information can be retrieved using the `sessionInfo()` function in R. This documentation ensures that the computational aspects of this study are transparent and reproducible.