

FORECASTING WITH A LARGE NUMBER OF PREDICTORS: A TREE BASED
MOVING AVERAGE APPROACH

by

Burakhan Sel

B.S., Industrial Engineering, Boğaziçi University, 2018

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Industrial Engineering
Boğaziçi University

2020

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor, Assist. Prof. Mustafa Gökçe Baydoğan, for his continuous support and encouragement throughout the progress of this thesis. He offered me great research opportunities and inspired me to pursue an academic career. This thesis would not have been possible without his help.

I am thankful to my committee members, Prof. Gülgün Kayakutlu and Prof. Refik Güllü, for their valuable comments and advice.

I would like to thank the Scientific and Technological Research Council Turkey for the financial support with the 2210-A National Scholarships Programme for Master's students.

My special thanks go to my friends, Mehmet Fazıl Aydın, Oğuz Yılmaz, Mirhan Ürkmez, and Onur Şero for their support and listening to my complaints during this period. I also thank faculty members in the Department of Industrial Engineering, Boğaziçi University, and in the Department of Industrial Engineering, Istanbul Technical University, who have directly or indirectly helped me during this period.

Last but not least, I am deeply grateful to my parents, Fatma and Şevki Sel, and my brothers, Çağrı and Oğuzhan Sel, because of their love, encouragement and endless support in my whole life.

ABSTRACT

FORECASTING WITH A LARGE NUMBER OF PREDICTORS: A TREE BASED MOVING AVERAGE APPROACH

A large number of predictors brings valuable information to time series forecasting problems, as well as difficulty in capturing this information. Short-term electricity load forecasting (STLF) problems are examples of such problems with many predictors, including several temperature values from different regions, a large number of special days, and multiple time-related variables. In such cases, forecasting with many predictors can be problematic in terms of robustness to redundant predictors, tendency to overfit, and the curse of dimensionality. To resolve these problems, this study proposes a novel tree-based ensemble model, tree-based moving average (TBMA), that provides point and probabilistic forecasts and works as an automated feature extraction method. The proposed model deals with a large number of predictors without sacrificing accuracy and does not require a complicated parameter tuning process as the advantages of being a tree-based ensemble model. The distinctive feature of the proposed model from existing tree-based ensemble models is that the suggested model considers the autocorrelation in time series data with the integrated moving average model and extracts useful features. Our comprehensive experiments show that boosting approaches provide significantly better results when features from TBMA are introduced. The proposed approach also provides competitive results compared to benchmark models in point and probabilistic forecasting of Turkey's electricity load.

ÖZET

ÇOK SAYIDA BAĞIMSIZ DEĞİŞKEN İLE ZAMAN SERİSİ TAHMİNİNE KARAR AĞACI TEMELLİ HAREKETLİ ORTALAMA YAKLAŞIMI

Çok sayıda bağımsız değişken, zaman serisi tahmin problemlerine değerli bilgiler sağlamanın yanında bu bilgilerin kavranması açısından zorluklar getirmektedir. Kısa vadeli elektrik yükü tahmin problemleri, hava koşullarından, özel günlerden ve haftanın günü gibi zamana bağlı birçok değişkenden etkilenmektedir. Çok sayıda bağımsız değişken ile tahmin yapmak, gereksiz değişkenlerin varlığı ve modelin veriye aşırı uyum sağlaması gibi konular dolayısıyla sorun yaratabilmektedir. Bu sorunları çözmek için, bu çalışma noktasal ve olasılıklı tahminler yapabilen ve aynı zamanda otomatik bir değişken üretme yöntemi olarak da çalışabilen yeni bir ağaç temelli topluluk modeli önermektedir. Önerilen model, ağaç tabanlı bir topluluk modeli olmanın avantajı olarak, çok sayıda bağımsız değişkene karşı toleranslıdır ve karmaşık bir parametre ayarlama işlemi gerektirmez. Önerilen modelin mevcut ağaç temelli topluluk modellerinden ayırt edici özelliği, zaman serisi verilerindeki otokorelasyonu modele entegre edilmiş hareketli ortalama modeliyle kavrayabilmesi ve diğer tahmin modelleri için yararlı değişkenler üretebilmesidir. Türkiye'nin saatlik elektrik yükü verisi üzerindeki kapsamlı deneylerimiz, önerilen model tarafından üretilen değişkenlerin, tek başına veya eski değişkenlerle birlikte kullanıldığında tahmin modellerinin doğruluğunu önemli ölçüde artırdığını göstermektedir. Ayrıca, önerilen yaklaşım noktasal ve olasılıklı tahminler yapma konusunda karşılaştırılan modellere kıyasla rekabetçi sonuçlar vermektedir.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	iii
ABSTRACT	iv
ÖZET	v
LIST OF FIGURES	viii
LIST OF TABLES	x
LIST OF SYMBOLS	xii
LIST OF ACRONYMS/ABBREVIATIONS	xiii
1. INTRODUCTION	1
2. BACKGROUND AND LITERATURE SURVEY	7
2.1. Moving Average	7
2.2. Regression Trees	7
2.3. Tree-Based Ensemble Models	9
2.4. Short-Term Electricity Load Forecasting	10
3. METHODOLOGY	15
3.1. Tree-Based Moving Average	15
3.1.1. Point Forecasts	15
3.1.2. Probabilistic Forecasts	19
3.1.3. Feature extraction	19
3.2. Illustrative Example	21
4. EXPERIMENTS AND RESULTS	26
4.1. Experiments	26
4.1.1. Point Forecasts	26
4.1.2. Probabilistic Forecasts	30
4.2. Result and Discussion	31
4.2.1. Performance Metrics	31
4.2.1.1. Root Mean Square Error	31
4.2.1.2. Mean Absolute Percentage Error	31
4.2.1.3. Mean Absolute Error	31

4.2.1.4.	Weighted Mean Absolute Percentage Error	32
4.2.1.5.	Mean Pinball Loss	32
4.2.2.	Point Forecasts	33
4.2.3.	Probabilistic Forecasts	38
4.2.4.	Complexity Analysis	39
5.	CONCLUSION	42
	REFERENCES	44
	APPENDIX A: PARAMETER TUNING WITH CROSS-VALIDATION	52

LIST OF FIGURES

Figure 1.1.	Comparison of a typical knowledge-based model and TBMA model for a STLF problem.	4
Figure 1.2.	Graphical summary of the short-term electricity load point and probabilistic forecasting with TBMA model. C_i stands for the i^{th} candidate forecast.	5
Figure 2.1.	The partition of the train space R into rectangular regions R_1, R_2, R_3 by using the recursive binary splitting method.	9
Figure 3.1.	Point and probabilistic forecasting procedure of TBMA model. . .	16
Figure 3.2.	Tree-based moving average algorithm for point forecasting.	18
Figure 3.3.	Tree-based moving average algorithm for feature extraction.	20
Figure 3.4.	Graphical summary of the feature extraction with TBMA model. .	21
Figure 3.5.	Point forecasting with illustrative data.	23
Figure 3.6.	The decision tree that is fitted to Sample 1.	24
Figure 4.1.	Daily averages of Turkey's hourly electricity consumption in 2016 and 2017.	27
Figure 4.2.	Autocorrelation functions of hourly electricity consumption in a week (4.2a) and daily consumption in a month (4.2b).	27

Figure 4.3.	The train and test set splitting process, and one day enlarging at each iteration.	28
Figure 4.4.	The average ranks of all models for one-day-ahead forecasts (4.4a), four-days-ahead forecasts (4.4b), and seven-days-ahead forecasts (4.4c). The critical distance at significance level 0.05 is 0.872. . . .	37
Figure 4.5.	Graph of one-day-ahead daily WMAPE values of the models in June 2018. Ramadan holiday is between 15.06.2018-17.06.2018. . .	39
Figure 4.6.	Graph of mean pinball loss values in 2018.	40
Figure 4.7.	Feature extraction time with multiple proportions of observations and features. The analyses are conducted on a sample from Turkey's electricity load data with 10000 observations and 100 features. . .	41

LIST OF TABLES

Table 2.1.	The predictors used in the short-term hourly electricity load forecasting of the macro-regions.	13
Table 2.2.	The forecasting models used in the short-term hourly electricity load forecasting of the macro-regions.	14
Table 3.1.	Synthetic time series dataset generated for the illustration of TBMA model. X and Y observations are train and test observations, respectively.	22
Table 3.2.	The sets of similar observations and candidate forecasts that are founded by SMA with order one.	24
Table 4.1.	The parameters and their levels for feature extraction by TBMA model.	29
Table 4.2.	The selected parameter levels of the ET and TBMA models for the electricity load forecasting	30
Table 4.3.	RMSE values of the models in 2018.	33
Table 4.4.	MAPE values of the models in 2018.	34
Table 4.5.	MAE values of the models in 2018.	35
Table 4.6.	The average ranks of all models based on daily WMAPE scores. The numbers in parentheses indicate the order of the average ranks.	38

Table A.1.	Model parameters selected in the cross-validation process.	52
------------	--	----

LIST OF SYMBOLS

C_i	i^{th} candidate forecast
h	Horizon parameter of the tree-based moving average model
l	Number of rectangular regions in train space R
m	Order parameter of the moving average model
N	Number of decision trees in the tree-based moving average model
R	Train space
R_i	i^{th} rectangular region in train space R
S_i	i^{th} set of similar observations to a test observations
$y_{d,h}$	Actual consumption value in day d , hour h
$\hat{y}_{d,h,n}$	Predicted consumption value in day d , hour h from n -days-ahead forecasting horizon
$\hat{y}_{d,h,n,p}$	p^{th} percentile predicted consumption value in day d , hour h from n -days-ahead forecasting horizon

LIST OF ACRONYMS/ABBREVIATIONS

ANN	Artificial Neural Network
ARIMA	Autoregressive Integrated Moving Average
DT	Decision Tree
ET	Extra Trees
ETS	Error Trend Seasonal
GBM	Gradient Boosting Machine
GLM	Generalized Linear Model
KNN	K-Nearest Neighbor
MA	Moving Average
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
RSS	Residual Sum of Square
RMSE	Root Mean Square Error
SARIMA	Seasonal Autoregressive Integrated Moving Average
SES	Simple Exponential Smoothing
SMA	Simple Moving Average
STLF	Short-Term Load Forecasting
SVM	Support Vector Machine
TBMA	Tree-Based Moving Average
WMAPE	Weighted Mean Absolute Percentage Error
XGB	Extreme Gradient Boosting

1. INTRODUCTION

Forecasting problems may need a large number of predictors to capture the dynamics of response variables. A large number of predictors bring valuable information to the forecasting models, along with the curse of dimensionality and difficulty exploiting the information from many predictors [1]. High dimensional data may cause overfitting and require unreasonable computational resources. The presence of redundant predictors can make forecasting complicated. In such cases, utilization from many predictors without being affected by its pitfalls becomes critical for forecasting models.

The statistical time series models such as moving average (MA), autoregressive integrated moving average (ARIMA), and seasonal autoregressive integrated moving average (SARIMA) are used in the forecasting problems to model the autocorrelations in time series data. MA model is one of the computationally simple statistical models, but its capability to capture overall relationships between time series observations is degraded when significant observations are beyond the MA model's scope. ARIMA and SARIMA models overcome the disadvantage of MA model by modeling the trend, noise, or seasonality components of time series data. However, the time series models may not solely be sufficient for the forecasting problems that include many predictors and nonlinear correlations since the mentioned time series models are univariate models and assume linear relationships between observations.

Machine learning models are used in the forecasting problems to capture the nonlinear dynamics of the time series data and to consider predictors. Artificial neural network (ANN) and support vector machine (SVM) models are widely used machine learning models due to their high learning capacity. However, ANN models are often criticized for the sensitivity to the selection of network architecture and the tendency to overfit [2]. SVM model has limitations in model parameter handling [3] and requires long processing time to build [4]. Therefore, ANN and SVM are hard to generalize and require advanced parameter tuning processes.

Tree-based ensemble models are generalized machine learning models in terms of the adaptability to time series forecasting problems without a complicated parameter tuning process. Tree-based ensemble models such as random forest [5], extra trees (ET) [6], extreme gradient boosting (XGB) [7] facilitate forecasting with a large number of predictors. These models can capture the complex correlations between predictors and response variables [8]. They also tend to have low bias and low variance as advantages of aggregating multiple decision trees. However, tree-based ensemble models lack in considering the autocorrelation in time series data by assuming that the observations are independent. Therefore, tree-based ensemble models are needed to be improved for time series forecasting problems. Although the improvement can be achieved by emphasizing the autocorrelation in time series data with some predictors (e.g., a variable indicating the order of observations predictors), there are still rooms to improve the accuracy of tree-based ensemble models for forecasting problems by not only using proper inputs but also improving the algorithms.

Feature extraction is an essential part of data processing to deal with a large number of predictors. Feature extraction methods create new features from a high dimensional dataset to summarize the information from the initial dataset and facilitate forecasting with a manageable number of predictors. In the literature, principal component analysis is used to capture the information of predictors with less dimension in an unsupervised way [9]. A discrete wavelet transform is used to extract new features such as trend and periodic waves of a time series [10]. Machine learning models based on convolutional neural networks [11] and deep learning [12] are used for automated feature extraction by considering both predictors and the response value of time series data.

This thesis proposes a novel tree-based ensemble model called tree-based moving average (TBMA) for forecasting problems with a large number of predictors. TBMA model provides point and probabilistic forecasts and works as an automated feature extraction model. The proposed model is composed of the MA approach's integration into the algorithm of a tree-based ensemble model to consider autocorrelations in time

series data. The integration of MA model does not complicate the computation because of the simplicity of MA. Due to the suggested model's tree-based ensemble nature, it can handle a large number of predictors without sacrificing accuracy and does not require a complex parameter tuning process. The automated feature extraction with TBMA helps existing models to exploit the information from many predictors with less dimension. The extracted features can be used along with the old features or alone to reduce the data dimension.

TBMA model is applied to a short-term load forecasting (STLF) problem, which is an example of forecasting problems with many predictors. STLF plays a key role in power systems, enables proper planning, and reduces costs related to overprediction or underprediction. STLF problems concern the hourly or sub-hourly electricity consumption and cover forecasts with up to one-week-ahead forecasting horizon. As a characteristic, electricity consumption behavior is affected by exogenous factors such as weather conditions and special days (i.e., calendar effects) [13,14]. A large number of predictors is needed in STLF problems considering the heterogeneity of the weather in consumption regions, and the variety of special days.

The proposed model is inspired by knowledge-based models in STLF studies, which identify similarities between consumption observations [15–18]. In knowledge-based models, identification of similar observations is made by using if-else rules based on features such as temperature and day of week [19]. The proposed model utilizes decision trees that are useful to detect similarities among the observations. The observations residing in the same terminal node of a decision tree can be considered as similar to each other, and further analyses for forecasting can be made according to the similarity information. Unlike knowledge-based models, the suggested model is supervised and suitable for big data due to utilizing decision trees. Illustrations of a

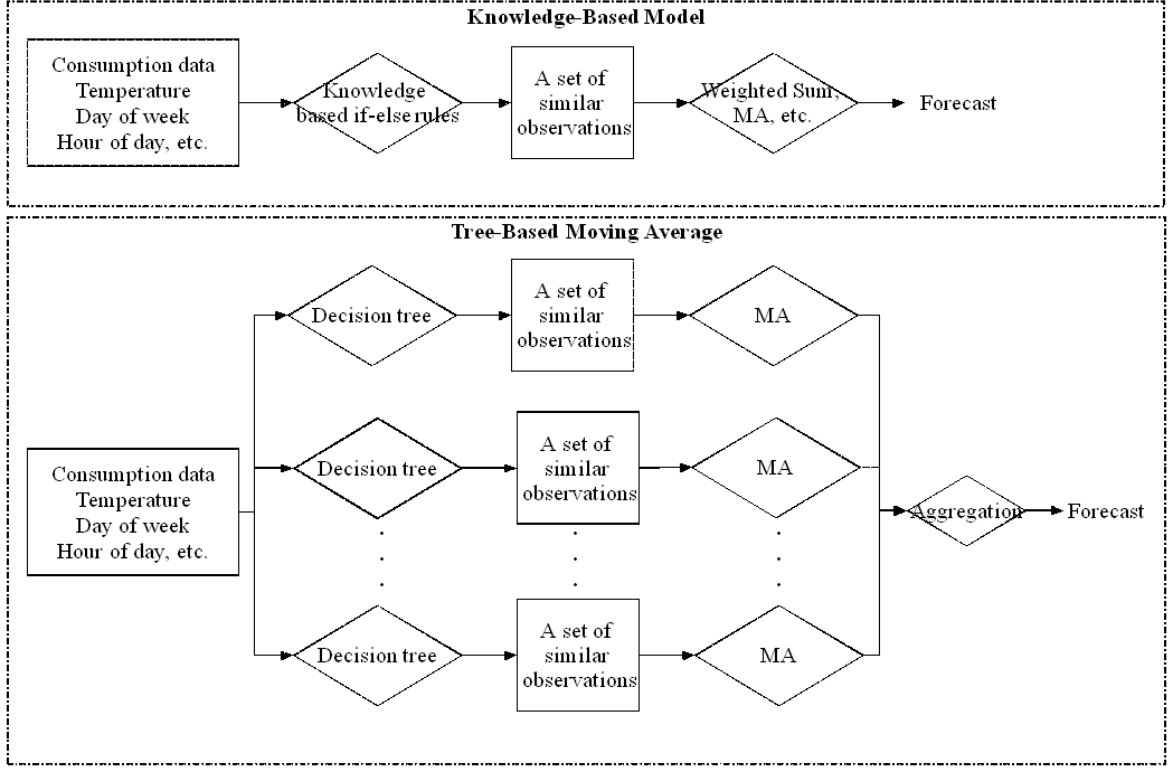


Figure 1.1. Comparison of a typical knowledge-based model and TBMA model for a STLF problem.

typical knowledge-based model and TBMA model for a STLF problem are presented in Figure 1.1.

Turkey's hourly electricity consumption data [20] are used in applying TBMA model to a STLF problem. The application is summarized in Figure 1.2. The data consist of observations from 2016 to 2019. The hourly consumption values in 2018 are predicted from one-day-ahead to seven-days-ahead as point and probabilistic forecasts with 371 consecutive experiments. The predicted hourly temperature values of the seven cities that consume the biggest part of Turkey's electricity and day specifications, such as Republic Day, are included in the predictors. The TBMA model's point forecasting performance is compared with ARIMA, ET, XGB, decision tree (DT), gradient boosting machine (GBM), generalized linear model (GLM), k-nearest neighbor (KNN), and naive models in terms of root mean square error (RMSE), mean absolute percentage error (MAPE), mean absolute error (MAE), and statistical tests based on weighted mean absolute percentage error (WMAPE). Two additional XGB models

(i.e., TBMA-XGB (1) and TBMA-XGB (2)) that use the features extracted by TBMA model are also included in the point forecasting experiment to analyze the effect of feature extraction on forecasting accuracy. TBMA model's performance in probabilistic forecasting is compared with ET, ARIMA, exponential smoothing (i.e., error trend seasonal (ETS)), simple exponential smoothing (SES), and naive models by using pinball loss performance measure.

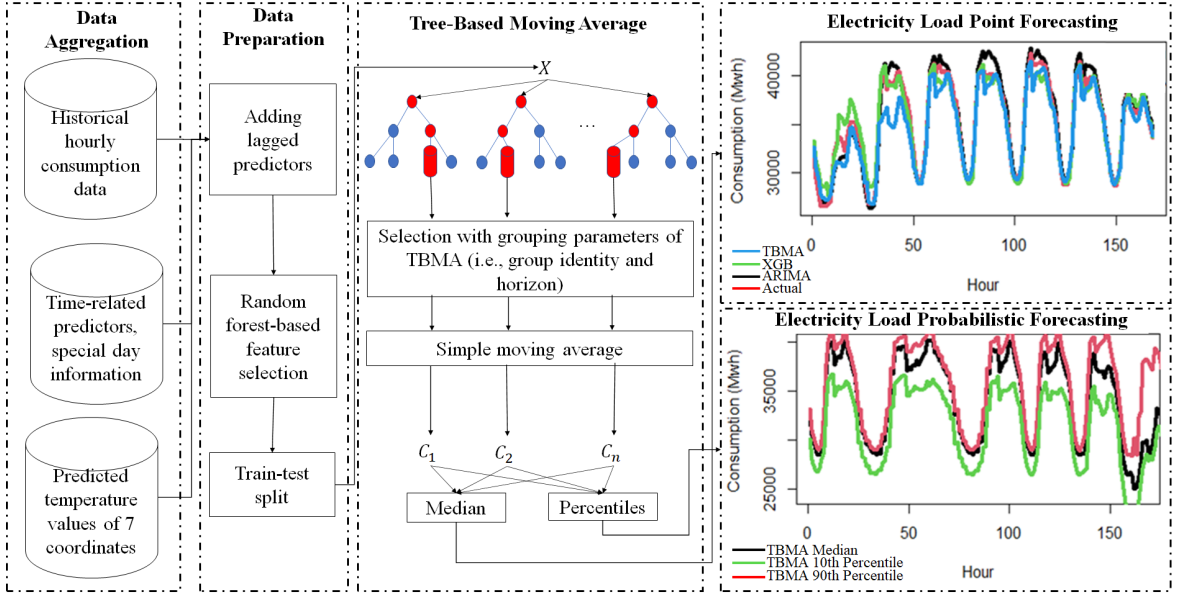


Figure 1.2. Graphical summary of the short-term electricity load point and probabilistic forecasting with TBMA model. C_i stands for the i^{th} candidate forecast.

TBMA model is proposed to deal with a wide variety of forecasting problems, including STLF problems, without the requirements of complex parameter tuning process and high computational power for training. ANN-based and SVM-based models are not included among benchmark models due to computational concerns. Although these models are state-of-art models in STLF literature, using them in 371 consecutive experiments is ineffective considering the required computational resources for parameter tuning and training with a large number of predictors.

The major contributions of this thesis are as follows:

- A novel tree-based ensemble model providing point and probabilistic forecasts and working as an automated feature extraction model is proposed.

- The suggested model can handle a large number of predictors and does not require a complex parameter tuning process or unreasonable computational resources.
- The performance of the proposed model in point forecasting, probabilistic forecasting, and feature extraction is analyzed by using Turkey’s hourly electricity consumption data and benchmark models. The implementations are provided on GitHub [21] to promote reproducibility.
- An R package called “tbma” is published on CRAN [22] for the use of TBMA model in forecasting problems.

The rest of the study is organized as follows. In the next chapter, background information and a review of STLF studies are provided. The methodology of TBMA model is described in Chapter 3. The experiment of TBMA model on an STLF problem and the results are presented in Chapter 4. Lastly, the main conclusions are given in Chapter 5.

2. BACKGROUND AND LITERATURE SURVEY

This chapter introduces the principles of MA, regression trees, and tree-based ensemble models. Afterward, the chapter presents a review of STLF studies.

2.1. Moving Average

MA approach has various versions to be used in time series forecasting problems. The simple equally weighted version of the MA approach, which is called simple moving average (SMA), predicts the next observation with the average of the last m observations. The formula of SMA model is given in Equation 2.1, where m corresponds to the order of SMA model, $\hat{y}_{t+1}$ is the predicted value for time $t+1$, and y_t is the response value in time t .

$$\hat{y}_{t+1} = \frac{1}{m} \sum_{i=t-m+1}^t y_i \quad (2.1)$$

The order parameter m affects the sensitivity of the predictions to the changes in the most recent observation. The smaller the m value, the higher the sensitivity to changes in the recent data. When m is equal to 1, the model becomes naive and predicts an observation with the response value of the previous observation.

2.2. Regression Trees

Decision trees are the machine learning models that can be used for both regression and classification problems. A decision tree model learns from a train dataset by simple decision rules. One of the versions of decision trees is known as Classification

and Regression Trees [23]. The algorithm of regression trees can be divided into two steps:

- (i) Divide the train space R into rectangular regions $R_1, R_2, \dots, R_n$ that cover the entire train set
- (ii) Predict the response values of the observations in the same region (y_i where $i \in R_k$ and $k \in 1, 2, \dots, n$) with their mean or mode $\hat{y}_{Rk}$.

The aim of the division into regions is to come up with a partition that has the lowest impurity criterion, such as the residual sum of square (RSS) (Equation 2.2). Since covering all the possible partitions to find the one with the lowest impurity is not computationally efficient [24], recursive binary splitting (RBS) is used to achieve a partition with a sufficiently low impurity. RBS divides the regions into two subregions by a cutpoint of an input variable that results in the best reduction in impurity. The splitting occurs by a simple if-then-else rule. RBS is known as a greedy approach because it splits a region with the most impurity reducing cutpoint as early as possible and does not consider to split a region which will lead to a better impurity reduction in the further steps. The greedy approach is computationally efficient, even though it may not find the optimal solution. RBS procedure continues to split regions to subregions until one of the stopping conditions, such as minimum node size, is met. Pruning a large regression tree or not allowing a regression tree to grow after some point prevents overfit and decreases the variance of the forecasts with a little cost of bias.

The partition of a train space with the recursive binary splitting method is illustrated with an arbitrary dataset in Figure 2.1, where the plotted values are the response values of the corresponding observations. The observations have two input variables that are V1 and V2. The minimum number of observations in a region is allowed to be 3. The predictions are made with the mean of the observations in a region. RSS values indicate the residual sum of square at each partition.

$$RSS = \sum_{k=1}^n \sum_{i \in R_k} (y_i - \hat{y}_{R_k})^2 \quad (2.2)$$

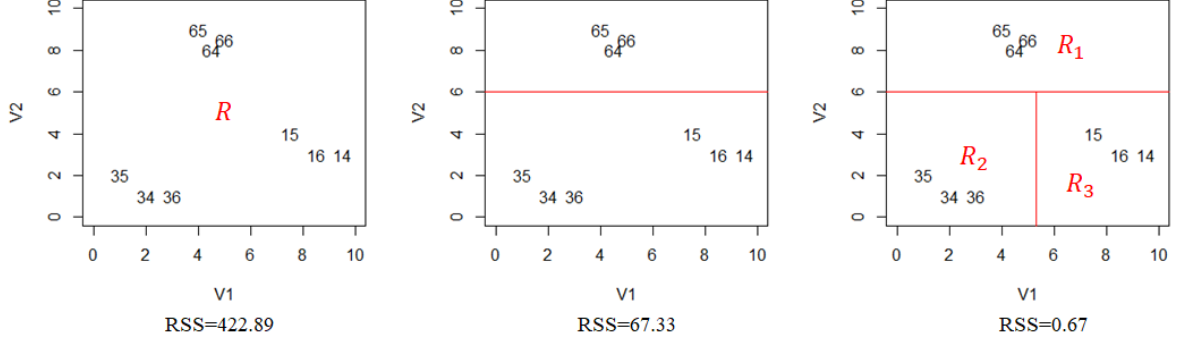


Figure 2.1. The partition of the train space R into rectangular regions R_1, R_2, R_3 by using the recursive binary splitting method.

Regression trees can be easily interpreted and visualized due to the use of simple if-then-else rules. They allow numerical and categorical variables and perform well when used with ensemble methods. However, regression trees have high variance as a disadvantage.

In the field of electricity consumption forecasting, practitioners usually relate a recent consumption behavior to historical consumption behaviors. In that sense, regression trees are useful to detect similarities among observations. The observations residing in the same terminal node (i.e., region) can be considered as similar to each other, and further analyses for forecasting can be made according to the similarity information.

2.3. Tree-Based Ensemble Models

The tree-based ensemble models such as the bagging [25], random forest, and ET build multiple decision trees on random subsets of data and aggregates the results of decision trees. Decision tree models provide forecasts with low bias and high variance due to their tendency to overfit the train data. Tree-based ensemble models reduce the

correlation between decision trees and aggregate the results of decision trees to achieve low bias and low variance. In bagging, multiple decision trees are fitted to bootstrapped samples drawn with replacement, and the trees' predictions are aggregated to obtain a final prediction. The random forest model follows similar procedures with the bagging model except for the random subspace technique [26] that is randomly selecting a set of input variables as candidate input variables to be used for splitting. The ET model, which trains decision trees to the samples drawn without replacement, alters the splitting procedure by adding one more random process that is choosing the cut-points of the candidate input variables randomly. With the mentioned strategies of the tree-based ensemble models, such as the random subspace technique, diversity of decision trees is achieved to obtain low bias and low variance.

The tree-based ensemble models such as boosting [27], GBM [28], and XGB depend on gradually improving the performance of decision trees. In boosting, decision trees are built by focusing on the observations that are hard to predict. The gradient approach in GBM and XGB minimizes the loss function gradually with each decision tree. By sequentially improving decision trees, these models achieve predictions with low bias and low variance.

2.4. Short-Term Electricity Load Forecasting

The electricity load forecasting problems are categorized by forecasting horizons and periodicity of observations [29]. In terms of forecasting horizons, the categories are short-term (up to one-week-ahead), mid-term (one-week-ahead to one-year-ahead), and long-term (more than one-year-ahead) forecasts. The problems in different categories of forecasting horizons require different input variables. For example, short-term electricity consumption behavior may not be significantly affected by the economic factors that change in long periods, while long-term electricity consumption behavior is affected by the economic factors significantly. In terms of the periodicity of observations, the electricity consumption forecasting problems are divided into categories such as half-hourly, hourly, daily, or monthly problems. The hourly or half-hourly problems

reflect different seasonality behavior than the monthly and yearly problems. Therefore, the complexities of the problems and required input variables differ according to the forecasting horizon and periodicity of observations.

This study is positioned in the literature of STLF regarding two aspects: input variables (Table 2.1) and forecasting models (Table 2.2). The studies in the last ten years are reviewed among Science Citation Index/Expanded indexed journals by using “electricity consumption forecast”, “electricity load forecast”, and “electricity demand forecast” keywords. The results are narrowed by the condition of being short-term and being based on hourly observations. Lastly, the studies about electricity consumption in micro-regions (e.g., buildings) are excluded because the forecasting models for micro-regions and macro-regions (e.g., countries) require different input variables.

Table 2.1 and Table 2.2 indicate that temperature values, previous consumption values, special days and time-related predictors are commonly used in a wide variety of forecasting models.

The models that are proposed in STLF literature can be categorized into two main groups: statistical and machine learning models. In the group of statistical models, regression models [29, 35, 39, 40, 46, 50] and time series models such as ARMA [48], ARIMA [49], exponential smoothing [52], and Holt-Winters [33] are used. The statistical models are criticized by their inability to handle nonlinear problems. Machine learning models are used in the literature to capture the nonlinear dynamics of electricity consumption. Among the machine learning models, ANN-based models [31, 37, 44, 54, 55] and SVM-based models [56–58] are the most popular models used in energy economics studies [3]. Although ANN and SVM have high learning capacity, they have limitations in parameter handling [3]. Tree-based ensemble models such as RF [14, 59], GBM [36], and XGB [60] are also used in STLF literature. In [61], RF is found to be favorable to ANN model in terms of the easiness of the model building process. Moreover, fuzzy logic-based models, which are advantageous in interpretability, are employed in STLF problems [38, 45, 46]. Some studies suggest hybrid

models [62, 63] to compensate for the drawbacks of the single machine learning models. Although the statistical time series models are combined with other models in various studies [29, 33, 39, 42, 48, 49], the combination of tree-based ensemble models and statistical time series models is the lack in the reviewed STLF literature.

Table 2.1. The predictors used in the short-term hourly electricity load forecasting of the macro-regions.

Literature	Temperature	Special Days	Lag Consumption	Time-Related Variables	Others	None
[29]	✓	✓	✓			
[30]			✓	✓		
[31]	✓	✓	✓	✓		
[32]	✓			✓	✓	
[33]		✓	✓	✓		
[34]	✓	✓	✓	✓		
[35]	✓	✓	✓	✓		
[36]	✓	✓	✓	✓		
[37]	✓			✓	✓	
[38]	✓		✓	✓		
[9]	✓	✓		✓		
[39]	✓		✓	✓	✓	
[40]						✓
[41]	✓	✓		✓		
[42]		✓	✓			
[43]	✓			✓		
[44]	✓	✓	✓	✓		
[45]	✓		✓	✓	✓	
[46]	✓	✓	✓	✓	✓	
[47]	✓		✓			
[48]		✓	✓			
[49]	✓	✓	✓			
[50]	✓	✓	✓	✓	✓	
[51]	✓		✓		✓	
[52]		✓	✓	✓		
[53]			✓	✓		
This study	✓	✓	✓	✓	✓	

Table 2.2. The forecasting models used in the short-term hourly electricity load forecasting of the macro-regions.

Literature	Proposed Model
[29]	Additive regression model with linear and nonlinear parameters
[30]	Combination of multiple seasonal patterns, neural networks, non-positive constraint theory and Broyden–Fletcher–Goldfarb–Shanno–Firefly algorithm
[31]	Deep belief networks
[32]	Semi-functional partial linear model
[33]	A new rule based Holt–Winters–Taylor method
[34]	Temporal and weather conditional epi-splines based load model using functional approximation
[35]	Regression model
[36]	Gradient boosting
[37]	Bayesian neural network learned by the Hybrid Monte Carlo algorithm
[38]	Fuzzy logic and adaptive neuro-fuzzy inference system
[9]	Combination of genetic algorithm and ant colony optimization for feature selection and multi-layer perception for hourly load prediction
[39]	Regression with reflection of an economic input variable obtained by moving average
[40]	Penalised least-squares
[41]	Sequential aggregation of specialized experts
[42]	Two-stage seasonal autoregressive integrated moving average with exogenous regressors
[43]	Combination of stationary wavelet transforms
[44]	Extreme learning machine with automatic knowledge representation
[45]	Adaptive two-stage hybrid system consisting of Fuzzy & Radial movement optimization
[46]	Multivariate dynamic regression
[47]	Genetic–fuzzy and ant colony–fuzzy hybrid models
[48]	Rule-based autoregressive moving average
[49]	Regression with ARIMA
[50]	Additive quantile regression models
[51]	Integration of long short-term memory and genetic algorithm
[52]	Exponential smoothing
[53]	Components estimation technique
This study	Tree-based moving average

3. METHODOLOGY

3.1. Tree-Based Moving Average

The idea of the TBMA approach is to relate a test observation to train observations multiple times by using diversified decision trees and an additional set of rules based on grouping parameters of TBMA model. The multiple decision trees detect the similarity between test and train observations in a supervised way that is suitable for big data. The use of an additional set of rules enables users to make the similarity information more specific and narrows the sets of train observations that are found to be similar to a test observation.

TBMA model calculates multiple candidate forecasts by using multiple decision trees and a set of rules based on the grouping parameters. The proposed model aggregates the candidate forecasts by their median value to obtain a point forecast or uses the distribution of the candidate forecasts to obtain probabilistic forecasts. TBMA model also extracts features by utilizing the similarity information between observations.

The procedures that are followed to obtain point and probabilistic forecasts are summarized in Figure 3.1, where $S_1, S_2, \dots, S_N$ represent sets of similar observations, and $C_1, C_2, \dots, C_N$ represent candidate forecasts.

3.1.1. Point Forecasts

The procedure of point forecasting a test observation with TBMA model is divided into three main steps:

- (i) Determine N sets of train observations that are similar to the test observation $S_1, S_2, \dots, S_N$ with the help of N decision trees
- (ii) Apply SMA to each set to calculate candidate forecasts $C_1, C_2, \dots, C_N$

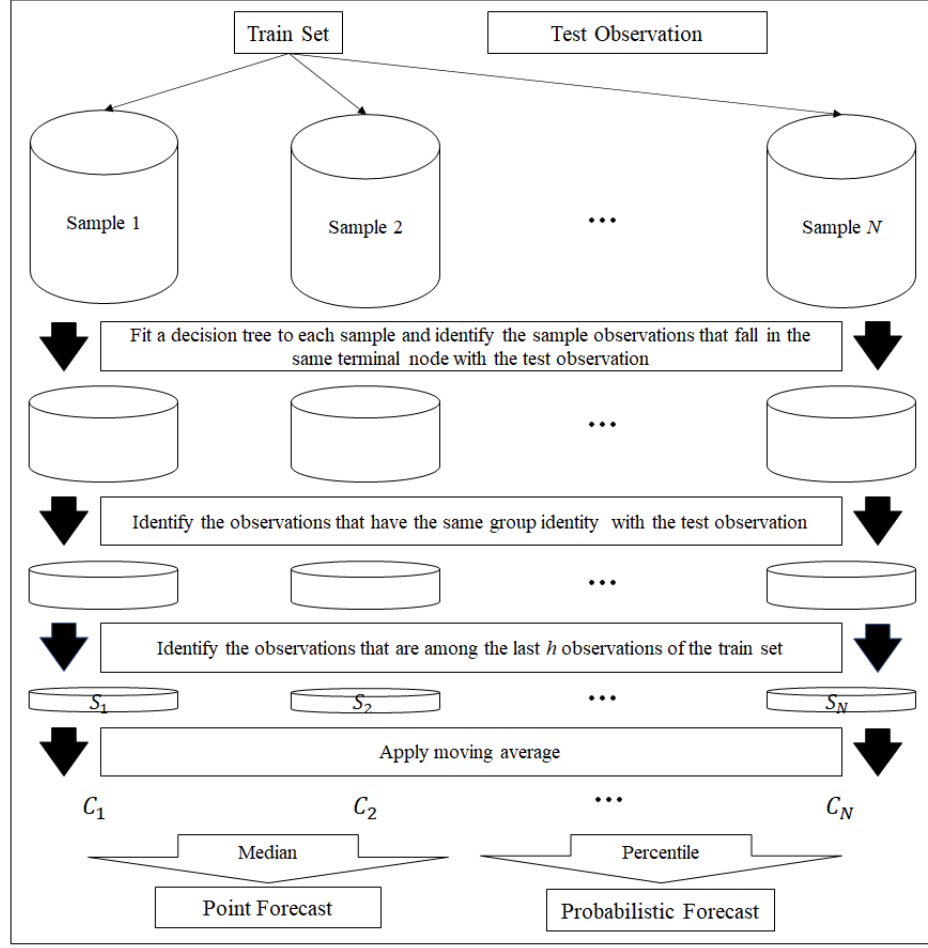


Figure 3.1. Point and probabilistic forecasting procedure of TBMA model.

(iii) Obtain the point forecast $\hat{y}_{test}$ by the median of the candidate forecasts

The determination of S_k where $k \in (1, 2, \dots, N)$ is based on a decision tree and TBMA's grouping parameters, which are group identity (θ) and horizon (h). The observations in S_k satisfy the following conditions:

- Be a member of the sample that the decision tree is fitted to (i.e., in-bag observation)
- Fall in the same terminal node with the test observation
- Have the same group identity with the test observation
- Be among the last h observations of the train set

The TBMA algorithm fits a decision tree to a sample from a train set. The sample observations in the same terminal node with the test observation are detected for the first step in determining the sets of similar observations.

The group identity parameter emphasizes the input variables that have significant effects on the response value. The use of group identity parameter is optional, and it can be set to one or more categorical input variables when differences between the observations of different group identity levels are significant. For example, in hourly electricity consumption forecasting problems, the hour variable is a potential group identity parameter because of its effect on electricity consumption. The hour variable as the group identity parameter brings the condition that only the observations that occur in the same hour of days can be similar.

Similar observations must be close to each other in terms of the order of the observations. The h parameter provides a measure for the closeness. The use of h in TBMA model means that the observations have to be among the last h observations of the train set to be considered similar to a test observation.

The SMA part of TBMA model provides a candidate forecast by averaging the recent observations in a set of similar observations. The order parameter m of the SMA model determines the number of recent observations to average. Since the autocorrelation between the similar observations in a set still exists after the selection procedure, TBMA model uses the mean value of the recent and similar observations as a candidate forecast.

TBMA model finds the median of the candidate forecasts for point forecasting. Since the correlation between the decision trees in tree-based ensemble models is low due to some strategies, such as the random subspace technique, a wide range of candidate forecasts are obtained from the previous steps. The median value becomes a suitable point forecasting method because the median is robust to outliers in the

```

Draw samples from train set;
for each sample do
    Fit a decision tree to the sample;
    Identify the terminal node that each sample observation falls in;
    for each observation in the test set do
        Identify the terminal node that the test observation falls in;
        Select similar observations among the sample observations that fall in the
        same node with the test observation, has the same group identity with the
        test observation, and is among the last  $h$  observations of the train set;
        Apply SMA to the set of similar observations and obtain a candidate forecast;
    end for
end for
Predict each test observation with the median of the corresponding candidate
forecasts;

```

Figure 3.2. Tree-based moving average algorithm for point forecasting.

presence of diverse and outlier candidate forecasts. The algorithm of TBMA model for point forecasting is given in Figure 3.2.

The TBMA model's parameters consist of parameters related to tree-based ensemble methodology and three other parameters: group identity, horizon, and order. The group of tree-based ensemble parameters controls the processes of fitting decision trees and aggregation of multiple decision trees. For example, while minimum node size and maximum depth parameters determine the end of a decision tree, the number of trees parameter is required for the ensemble process. The group identity and horizon parameters rule the process of finding similar observations. The order parameter depends on the version of the MA model integrated into TBMA. In the SMA version of MA, the only parameter is the order, which determines the number of recent observations to average.

3.1.2. Probabilistic Forecasts

The TBMA approach provides probabilistic forecasts by using the distribution of the candidate forecasts. The probabilistic forecasts give information about the variance of future response values. The 50th percentile (i.e., median) of the candidate forecasts corresponds to the point forecast. As an example, the 75th percentile of the candidate forecasts predicts the value that the future responses will be less than with 0.75 probability.

3.1.3. Feature extraction

Regression models require time series features to capture the dynamics of a time series. A standard time series feature extraction involves partitioning a time series into intervals and describing each interval with statistical moments such as mean [64] or time series features such as lag correlation [65]. However, such applications are univariate and do not consider predictors. In the presence of predictors, the utilization of predictors becomes important for time series feature extraction in terms of exploiting the train set information, but identifying relevant predictors among many predictors is not easy. Our data-driven strategy resolves the problem of determining relevant predictors since the proposed model is robust to a large number of predictors due to its tree-based ensemble nature. The proposed model extracts time series features in a supervised way considering both predictors and response variables, which are time series.

TBMA model creates features by making forecasts for each observation in whole data set based on the train set. The difference between the point forecasting and feature extraction approaches of TBMA model is that the feature extraction model predicts the train and test observations by using the information from the train set while TBMA point forecasting model predicts only the test observations. TBMA feature extraction model fits decision trees to samples taken from the whole train set. However, the similar observations to the train observation are selected among only the previous

train observations. With changes in the algorithm of TBMA point forecasting model, the algorithm for feature extraction becomes as in Figure 3.3.

```

Draw samples from train set;
for each sample do
    Fit a decision tree to the sample;
    Identify the terminal node that each sample observation falls in;
    for each observation in the sample and test set do
        Identify the terminal node that the observation falls in;
        Select similar observations among the previous sample observations that fall
        in the same node with the observation, has the same group identity with the
        observation, and is among the previous  $h$  observations;
        Apply SMA to the set of similar observations and obtain a candidate forecast;
    end for
end for
Determine the extracted feature value of each observation with the median of the
corresponding candidate forecasts;

```

Figure 3.3. Tree-based moving average algorithm for feature extraction.

A created feature by TBMA model is affected by the parameter setting of the TBMA model because the created feature is equivalent to the forecasts of TBMA model for the whole dataset. The TBMA feature extraction method creates multiple features by using the multiple combinations of TBMA model parameters. The aim of using multiple parameter settings and creating various features is to comprehensively exploit the information from the train set. Therefore, it is meaningful to use parameter settings that result in diverge forecasts and consequently, diverge features so that different sides of the train set information can be captured. For example, the use of different group identity parameter levels may result in emphasizing the effects of different features on the forecasts and diverging extracted features. The feature extraction with TBMA model is summarized in Figure 3.4.

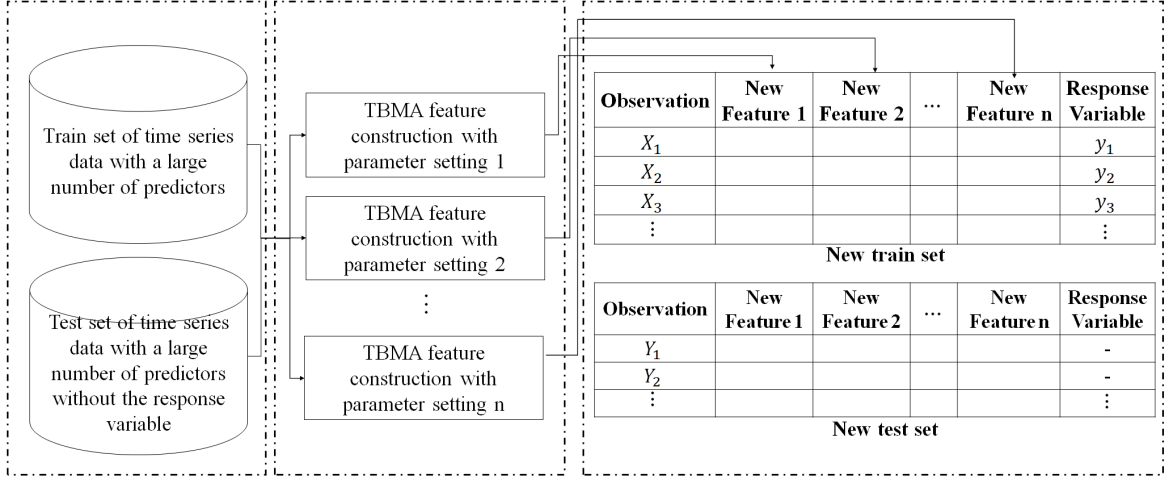


Figure 3.4. Graphical summary of the feature extraction with TBMA model.

3.2. Illustrative Example

A synthetic time series dataset is generated for the illustration of TBMA model. The dataset with four imaginary features and a response variable is shown in Table 3.1. “Feature 1” is a categorical feature that has the group effect on the response variable. The group effect is defined numerically as in Equation 3.1. “Feature 2” and “Feature 3” are two normally distributed random variables with mean 10 and 25, respectively. Both of their standard deviations are two. “Feature 2” and “Feature 3” bring randomness to the dataset. “Feature 4” corresponds to the response value of the previous observation and is defined to provide interdependence between the observations. The relationship between the features and the response variable is formulated as in Equation 3.2. The first 20 observations are separated for the train set, and the last observation is chosen for the test set.

$$\text{Group Effect} = \begin{cases} 1 & \text{if Feature 1 is Group A} \\ 1.5 & \text{if Feature 1 is Group B} \\ 2 & \text{if Feature 1 is Group C} \end{cases} \quad (3.1)$$

Table 3.1. Synthetic time series dataset generated for the illustration of TBMA model.

X and Y observations are train and test observations, respectively.

Observation	Feature 1	Feature 2	Feature 3	Feature 4	Response
X_1	Group A	9.9	22	20	87.7
X_2	Group B	11.5	25.3	87.7	71.5
X_3	Group C	16.5	24.5	71.5	153.5
X_4	Group A	11.5	26.5	153.5	-45.5
X_5	Group B	10.1	26.2	-45.5	199.6
X_6	Group C	8.1	24.2	199.6	0.5
X_7	Group A	6.9	22.1	0.5	98.1
X_8	Group B	7.9	26.3	98.1	49.3
X_9	Group C	11.3	28.7	49.3	155.9
X_{10}	Group A	11.3	23.1	155.9	-45.1
X_{11}	Group B	10.5	19.7	-45.1	206.9
X_{12}	Group C	10.1	21.5	206.9	1.9
X_{13}	Group A	7.6	24.8	1.9	96.1
X_{14}	Group B	8.2	24.8	96.1	53.7
X_{15}	Group C	8.6	27.9	53.7	144.2
X_{16}	Group A	13	26.4	144.2	-31.6
X_{17}	Group B	11.6	25.6	-31.6	190.8
X_{18}	Group C	7	27.9	190.8	2.3
X_{19}	Group A	13	29.8	2.3	106.9
X_{20}	Group B	9.9	23.3	106.9	49.5
Y_{test}	Group C	10.3	25.5	49.5	155.9

$$\begin{aligned} \text{Response variable} = & 3 \cdot (\text{Feature 2}) - (\text{Feature 3}) - (\text{Feature 4}) \\ & + 100 \cdot \text{Group Effect} \end{aligned} \quad (3.2)$$

The number of decision trees N is set to three for simplicity. Therefore, TBMA model will define three sets of similar observations S_1, S_2, S_3 . The steps of TBMA model for forecasting with illustrative data are summarized in Figure 3.5.

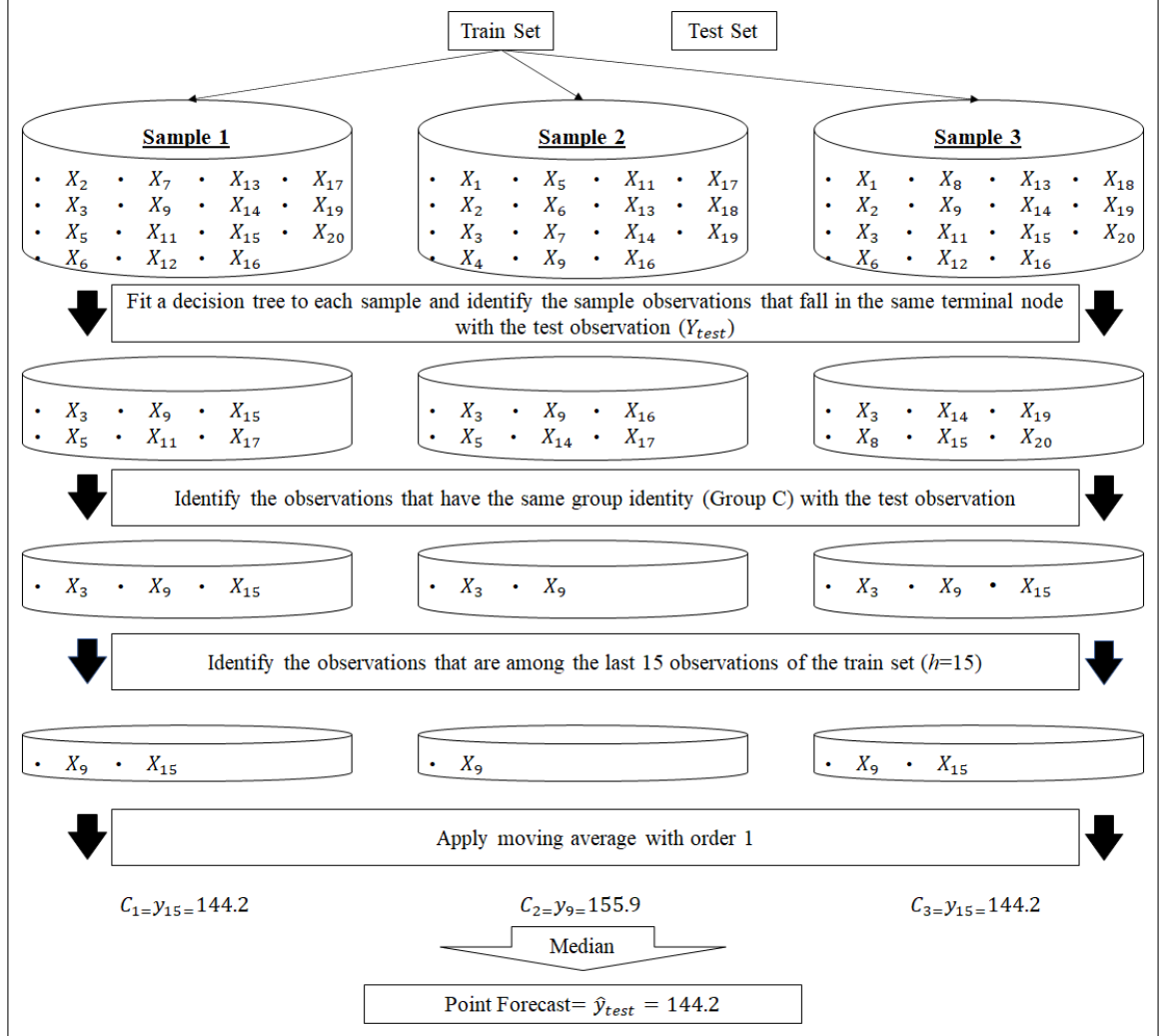


Figure 3.5. Point forecasting with illustrative data.

As the first step for determining the sets of similar observations, three sample samples, Sample 1, Sample 2, and Sample 3, are drawn without replacement from the train set of the synthetic dataset. A decision tree is fitted to each sample to identify the sample observations that fall in the same node with the test observation Y_{test} . The decision tree that is fitted to the Sample 1 is visualized in Figure 3.6 where n indicates the number of observations in each node.

The group identity parameter is set to the “Feature 1”. Therefore, similar observations must have the same “Feature 1” (i.e., Group C) level with the Y_{test} to be considered similar. The h parameter is set to 15 in the synthetic dataset case. It

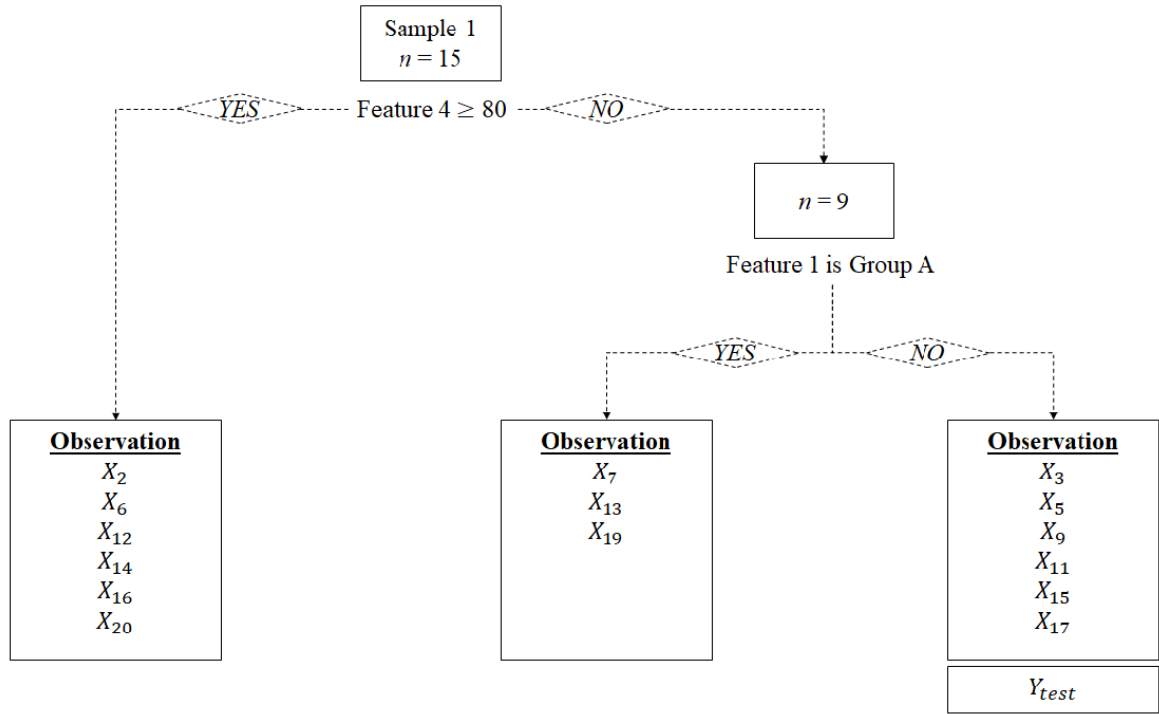


Figure 3.6. The decision tree that is fitted to Sample 1.

means that the observations before the X_5 can not be considered similar to the test observation.

The SMA model is applied to the S_1 , S_2 , and S_3 with the order parameter 1. Therefore, the candidate forecasts are simply the last observations of S_1 , S_2 and S_3 . The candidate forecast values from Sample 1, Sample 2, and Sample 3 are shown in Table 3.2.

Table 3.2. The sets of similar observations and candidate forecasts that are founded by SMA with order one.

Set of similar observations	Response Values	Candidate Forecast
$S_1 = \{X_9, X_{15}\}$	155.9 144.2	144.2 (C_1)
$S_2 = \{X_9\}$	155.9	155.9 (C_2)
$S_3 = \{X_9, X_{15}\}$	155.9 144.2	144.2 (C_3)

The point forecast of the test observation is found to be 144.2 as the median of the candidate forecasts. Since the number of candidate forecasts is low for simplicity, a probabilistic forecast is not provided.

4. EXPERIMENTS AND RESULTS

4.1. Experiments

TBMA model is applied to the hourly electricity consumption data of Turkey. The data consist of 26280 observations from 2016 to 2018. The consumption values in 2016 and 2017 are visualized in Figure 4.1. The electricity consumption dataset is rich in time series components such as seasonality, trend, and noise. The parallel consumption behaviors in 2016 and 2017 show the seasonality component in the data. The consumption values are generally higher in 2017 than in 2016, which indicates the trend component in consumption behavior. The variation of the consumption in short time intervals also represents daily and weekly seasonality. Autocorrelation functions of hourly consumption in a week (Figure 4.2a) and autocorrelation functions of daily consumption in a month (Figure 4.2b) indicate the daily and weekly seasonalities, respectively. Since the electricity consumption dataset is rich in typical times series dataset features, the electricity consumption dataset is suitable for the test of the proposed model.

In Figure 4.1, the sudden decreases in consumption in July 2016 and June 2017 correspond to the Ramadan holiday, which starts ten days earlier every year. The irregular decrease in August 2016 and August 2017 correspond to another religious holiday Festival of Sacrifice, which also starts ten days earlier every year.

4.1.1. Point Forecasts

The consumption values in 2018 are predicted from one-day-ahead to seven-day-ahead periods. The point forecasts are obtained by each of ARIMA, DT, ET, GBM, GLM, KNN, XGB, and naive models. TBMA model is implemented with two different parameter settings. Also, two additional XGB models that use the predictors extracted by the TBMA model are included in the experiment.

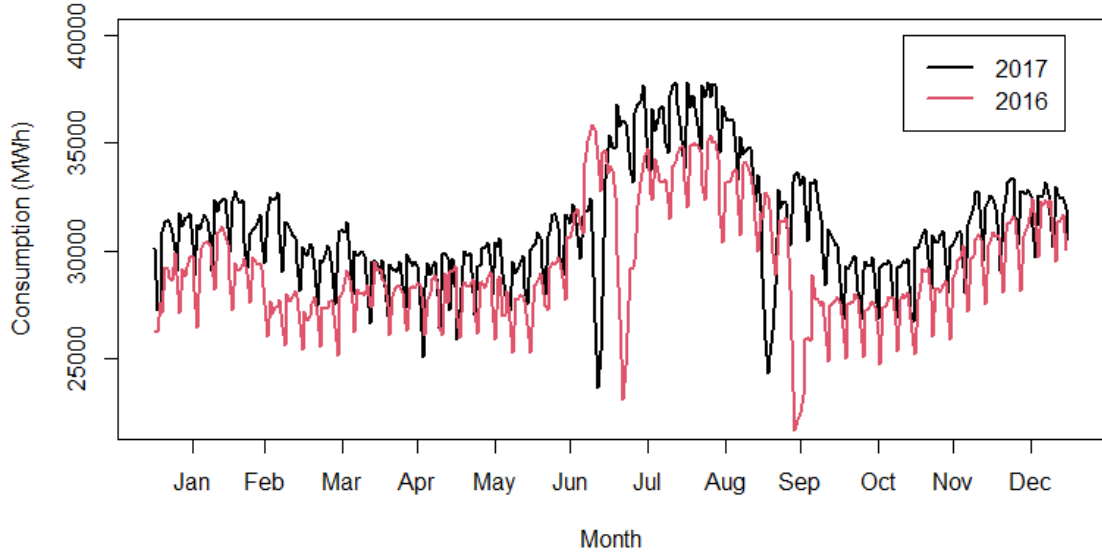


Figure 4.1. Daily averages of Turkey's hourly electricity consumption in 2016 and 2017.

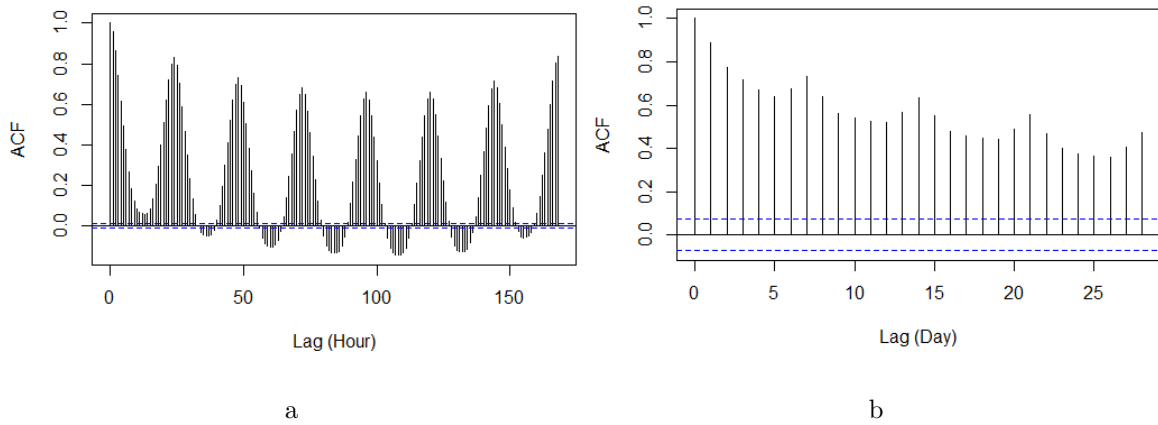


Figure 4.2. Autocorrelation functions of hourly electricity consumption in a week (4.2a) and daily consumption in a month (4.2b).

The train-test splitting procedure consists of 371 iterations to cover all of the days in 2018. At each iteration, the hourly consumption values in the next seven days of the train set are predicted, and the train set is enlarged one day in the next iteration. The iterations are illustrated in Figure 4.3.

The aggregation of time-related predictors such as hour, day specifications, temperature values, and their lagged versions corresponds to 115 predictors. The number

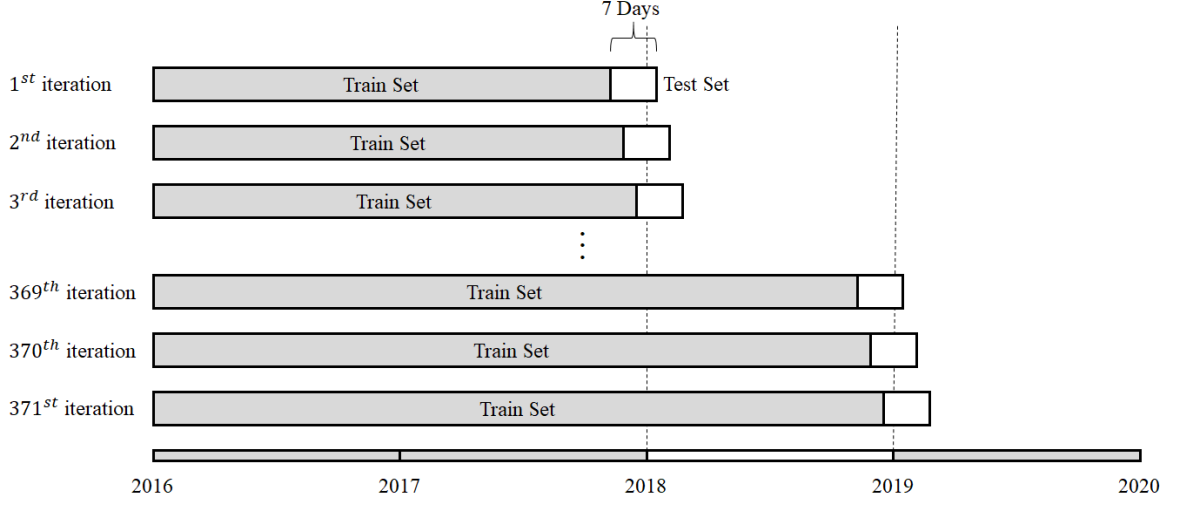


Figure 4.3. The train and test set splitting process, and one day enlarging at each iteration.

of predictors is decreased from 115 to 95 with the embedded RF-based feature selection procedure that selects the features with positive importance scores. The feature selection method is used to increase accuracy, although the proposed model is robust to a large number of predictors. The data from 2018 are saved to the test sets and not used for the feature selection process.

The selected 95 features are used in all models except ARIMA and naive, which are univariate models. The DT, ET, GBM, GLM, KNN, XGB, and TBMA models use the 95 features directly. In the model called TBMA-XGB (1), 16 new features are extracted by the TBMA model, and only the extracted features are given to the XGB model. In the TBMA-XGB (2) model, the extracted predictors are given to the XGB model along with the old 95 predictors. The TBMA model extracted 16 features with 16 different combinations of its four parameters. The selected parameters and their levels for feature extraction are shown in Table 4.1. The levels of the parameters are chosen to diversify the extracted features so that the extracted features bring more information to forecasting models.

The parameters of the DT, ET, GBM, GLM, KNN, XGB, TBMA-XGB (1), and TBMA-XGB (2) models are tuned by the cross-validation method. During the cross-

Table 4.1. The parameters and their levels for feature extraction by TBMA model.

Parameter	Level 1	Level 2
Number of features evaluated at each split	9	18
Always split variables	Hour, Ramadan holiday, Festival of Sacrifice	-
SMA order	1	2
Group identity	Hour	-

validation process, three validation sets are created by the time slice technique, which is choosing train and validation sets sequentially from the train set. The models' chosen levels of parameters are given in Table A.1, along with the candidate levels.

Two different TBMA models, which are called TBMA (1) and TBMA (2), are included in the experiment. The order parameter is set to one and two for TBMA (1) and TBMA (2), respectively. The reason for including both models is to analyze the impact of the order parameter on results. In the parameter setting of the TBMA models, the parameters required for training decision trees are chosen as the same with the parameters of the ET model. In other words, ET and TBMA models use the same decision trees for a fair comparison between them. The parameters of the ET and TBMA models are shown in Table 4.2, along with their selected levels. The detailed information about the parameters of the ET model is available in [6].

The data are split into 24 parts by hours of a day for the application of ARIMA model. The ARIMA models are separately applied to each of the 24 datasets with seven days periodicity. The parameters of the ARIMA models are selected independently for each of the 24 datasets by using the parameter selection method proposed in [66].

The naive model predicts a consumption value with the last week's consumption value in the corresponding day and hour. Although the model is called naive, which is also known as persistence, the naive model in this application has two additional assumptions that electricity consumption behavior in the same days of weeks and

Table 4.2. The selected parameter levels of the ET and TBMA models for the electricity load forecasting.

Model	Parameter name	Parameter level or value
TBMA and ET	Minimum node size	10
	Replacement	TRUE
	Splitrule	Extra trees splitting rule
	Random splits	5
	Number of features evaluated at each split	9
	Always split variables	Hour, Ramadan, Festival of Sacrifice
	Number of trees	500
TBMA	Order	1 for TBMA (1) and 2 for TBMA (2)
	Group identity	Hour
	Horizon	No value is given

the same hours of days are similar. These assumptions are based on the observed consumption behavior in Figure 4.1.

4.1.2. Probabilistic Forecasts

TBMA model generates probabilistic forecasts from the distribution of its candidate forecasts. As each decision tree provides a candidate forecast, the percentiles of the candidate forecasts indicate the probabilistic forecasts. The probabilistic forecasts from the 1st, 5th, 15th, 25th, 50th, 75th, 85th, 95th, and 99th percentiles are obtained by the distribution of the 500 candidate forecasts of TBMA model with order parameter one. A similar strategy is followed for the probabilistic forecasting with the ET model.

ARIMA, ETS, SES, and naive models assume a normally distributed noise with a constant variance around their point forecasts. The probabilistic forecasts of these models are calculated by using the point forecasts, the variance of the noise, and the cumulative distribution function of the standard normal distribution.

4.2. Result and Discussion

4.2.1. Performance Metrics

4.2.1.1. Root Mean Square Error. RMSE is a frequently used performance measure based on residuals. In the STLF experiment, the RMSE values are calculated by using hourly residuals in a year (i.e., between 2018.01.01 - 2018.12.31) as in Equation 4.1, where $y_{d,h}$ indicates the actual consumption in day d and hour h , and $\hat{y}_{d,h,n}$ represents the predicted consumption in day d and hour h from n -days-ahead forecasting horizon. The RMSE values are reported separately for each forecasting model and each forecasting horizon.

$$RMSE_n = \sqrt{\frac{\sum_{d=1}^{365} \sum_{h=0}^{23} (y_{d,h} - \hat{y}_{d,h,n})^2}{365 * 24}} \quad \text{for } n = 1, 2, \dots, 7 \quad (4.1)$$

4.2.1.2. Mean Absolute Percentage Error. MAPE performance metric reports the accuracy of forecasting models in percentages and provides intuitive interpretation. The MAPE scores in our experimentation are calculated by using the hourly absolute percentage residuals of hourly forecasts in 2018. The MAPE values for each forecasting model and forecasting horizon are calculated by using Equation 4.2.

$$MAPE_n = \frac{\sum_{d=1}^{365} \sum_{h=0}^{23} \frac{|y_{d,h} - \hat{y}_{d,h,n}|}{y_{d,h}} * 100}{365 * 24} \quad \text{for } n = 1, 2, \dots, 7 \quad (4.2)$$

4.2.1.3. Mean Absolute Error. MAE is a measure for the closeness of the predicted and actual values. The MAE scores in the STLF experiment are based on the absolute residuals of hourly forecasts in 2018 (Equation 4.3) and reported separately for each forecasting model and forecasting horizon.

$$MAE_n = \frac{\sum_{d=1}^{365} \sum_{h=0}^{23} |y_{d,h} - \hat{y}_{d,h,n}|}{365 * 24} \quad \text{for } n = 1, 2, \dots, 7 \quad (4.3)$$

4.2.1.4. Weighted Mean Absolute Percentage Error. WMAPE measure is based on the percentage of total absolute residuals in a time period. In our experiment, daily WMAPE scores are calculated for each day in 2018, as in Equation 4.4. Non-parametric statistical tests for comparing the performance of forecasting models (i.e., Friedman Test and Nemenyi Test) are employed by using the daily WMAPE values.

$$WMAPE_{d,n} = \frac{\sum_{h=0}^{23} |y_{d,h} - \hat{y}_{d,h,n}|}{\sum_{h=0}^{23} y_{d,h}} * 100 \quad \text{for } n = 1, 2, \dots, 7 \text{ and } d = 1, 2, \dots, 365 \quad (4.4)$$

4.2.1.5. Mean Pinball Loss. Pinball loss measure is a metric to analyze the accuracy of probabilistic forecasts, and based on giving weighted penalties to differences between actual values and probabilistic forecasts. The same absolute deviations of probabilistic forecasts from actual values are penalized with different weights based on the sign of deviation. A lower pinball loss score indicates more accurate probabilistic forecasts. The pinball loss score for the forecast of a specific day and hour with specific horizon and percentile is calculated as in Equation 4.5, where $\hat{y}_{d,h,n,p}$ represents the n -days-ahead, p^{th} percentile probabilistic forecast of the consumption in day d , hour h . In the probabilistic forecasting experimentation, nine different forecasts are obtained from the 1^{st} , 5^{th} , 15^{th} , 25^{th} , 50^{th} , 75^{th} , 85^{th} , 95^{th} , and 99^{th} percentiles. The mean pinball loss values of nine probabilistic features are calculated for each forecasting model and forecasting horizon to obtain an aggregated performance measure (Equation 4.6).

$$Pinball \ Loss_{d,h,n,p} = \begin{cases} (1 - \frac{p}{100})(\hat{y}_{d,h,n,p} - y_{d,h}) & \text{if } \hat{y}_{d,h,n,p} \geq y_{d,h} \\ \frac{p}{100}(y_{d,h} - \hat{y}_{d,h,n,p}) & \text{if } \hat{y}_{d,h,n,p} < y_{d,h} \end{cases} \quad \text{for all } n, d, h, p \quad (4.5)$$

$$Mean \ Pinball \ Loss_n = \frac{\sum_{d=1}^{365} \sum_{h=0}^{23} \sum_p Pinball \ Loss_{d,h,n,p}}{365 * 24 * 9} \quad \text{for } n = 1, 2, \dots, 7 \quad (4.6)$$

4.2.2. Point Forecasts

The annual RMSE values of the models for all forecasting horizons are given in Table 4.3. The TBMA-XGB (2) model, which uses the extracted predictors along with the old predictors, outperforms the existing models in all forecasting horizons. The second best RMSE scores are achieved by the TBMA-XGB (1), which uses only the extracted predictors. Among the models not using any extracted predictors, the XGB model has the lowest RMSE score in one-day to three-days-ahead forecasts, and the TBMA (1) has the lowest RMSE score in four-days to seven-days-ahead forecasts.

Table 4.3. RMSE values of the models in 2018.

	Forecast horizon (days)						
	1	2	3	4	5	6	7
ARIMA	1371.16	1907.10	2217.64	2420.28	2534.01	2581.21	2595.12
DT	2877.86	2911.05	2920.34	2924.37	2926.53	2925.97	2926.04
ET	1494.11	1524.22	1535.15	1541.13	1547.94	1554.07	1569.81
GBM	1505.11	1522.54	1558.34	1576.56	1583.34	1584.63	1585.04
GLM	1684.46	1696.00	1706.02	1713.32	1717.01	1718.54	1719.77
KNN	2570.83	2851.58	2987.38	3063.26	3116.17	3161.15	3226.20
NAIVE	2776.50	2776.50	2776.50	2776.50	2776.50	2776.50	2776.50
XGB	1164.05	1227.52	1307.78	1392.10	1397.40	1387.65	1443.29
TBMA (1)	1270.86	1334.00	1363.49	1361.70	1356.40	1330.97	1353.88
TBMA (2)	1354.15	1402.88	1430.67	1427.36	1428.37	1430.33	1467.38
TBMA-XGB (1)	1093.07	1195.81	1240.32	1252.17	1275.98	1275.75	1291.03
TBMA-XGB (2)	1014.59	1114.90	1167.07	1196.01	1221.44	1229.01	1263.58

The MAPE values of the models are shown in Table 4.4. The TBMA-XGB (2) and TBMA-XGB (1) models achieve the best and the second-best MAPE scores, respectively. For the third-best MAPE scores, the ARIMA, XGB, and TBMA (1) models are competitive. Among these models, the ARIMA model overperforms the other models in one-day ahead forecasts, the XGB model achieved the best MAPE values in two-days to four-days-ahead forecasts, and the TBMA (1) model has the lowest MAPE values in five-days to seven-days-ahead forecasts.

Table 4.4. MAPE values of the models in 2018.

	Forecast horizon (days)						
	1	2	3	4	5	6	7
ARIMA	2.47	3.49	4.02	4.37	4.64	4.79	4.88
DT	6.06	6.12	6.14	6.15	6.15	6.15	6.14
ET	3.53	3.61	3.63	3.64	3.66	3.68	3.71
GBM	3.26	3.32	3.38	3.40	3.42	3.42	3.42
GLM	3.88	3.90	3.93	3.94	3.95	3.96	3.97
KNN	4.58	5.07	5.31	5.48	5.61	5.72	5.85
Naive	4.66	4.66	4.66	4.66	4.66	4.66	4.66
XGB	2.48	2.69	2.83	3.04	3.07	3.11	3.20
TBMA (1)	2.72	2.99	3.08	3.05	3.01	2.94	2.99
TBMA (2)	2.97	3.14	3.21	3.19	3.19	3.20	3.31
TBMA-XGB (1)	2.33	2.61	2.73	2.75	2.81	2.81	2.85
TBMA-XGB (2)	2.12	2.41	2.55	2.60	2.70	2.75	2.79

The results of comparisons based on MAE values (Table 4.5) are parallel to the comparisons based on RMSE and MAPE measure. The best and the second-best performances are again observed with the TBMA-XGB (2) and TBMA-XGB (1) models, respectively. The ARIMA, XGB, and TBMA (1) models are competitive for the third-best performance. As a difference from the RMSE and MAPE comparisons, the TBMA (1) model achieves the third-best MAE scores in four to seven-days ahead forecast.

The procedure suggested by Dem̂sar [67] is used to test the performance of the models statistically. A Friedman test [68] is applied to check whether the median of the daily WMAPE scores of the forecasting models are equal or not. In case of a significant difference, a Nemenyi test [69] is employed to make non-parametric pairwise comparisons based on the ranks of forecasting models.

The Friedman tests for each forecasting horizon results in p-values lower than 0.05 significance level. As the next step, Nemenyi tests are employed for each forecasting

Table 4.5. MAE values of the models in 2018.

	Forecast horizon (days)						
	1	2	3	4	5	6	7
ARIMA	803.41	1120.94	1291.82	1401.00	1485.52	1535.67	1565.44
DT	1976.10	1993.30	2001.04	2003.11	2004.33	2005.03	2004.41
ET	1159.66	1183.23	1192.94	1197.85	1204.36	1211.95	1224.01
GBM	1069.33	1087.38	1106.18	1113.96	1121.41	1119.51	1120.67
GLM	1255.52	1264.66	1272.15	1277.97	1281.22	1283.41	1285.83
KNN	1470.90	1629.22	1713.43	1769.64	1813.33	1855.17	1903.10
Naive	1488.26	1488.26	1488.26	1488.26	1488.26	1488.26	1488.26
XGB	809.58	878.63	919.94	988.07	997.79	1014.61	1036.56
TBMA (1)	871.96	960.43	992.80	987.71	977.00	958.10	976.98
TBMA (2)	952.99	1009.44	1031.93	1029.29	1029.06	1035.40	1072.69
TBMA-XGB (1)	756.61	846.90	889.69	898.16	918.82	919.79	936.22
TBMA-XGB (2)	687.11	781.30	828.06	849.17	882.49	896.02	911.58

horizon, and the critical distance value for significance level 0.05 is found to be 0.872. It means that if the mean rank difference between two models is higher than 0.872, we can conclude that the two models' performances are significantly different, with 0.95 confidence level.

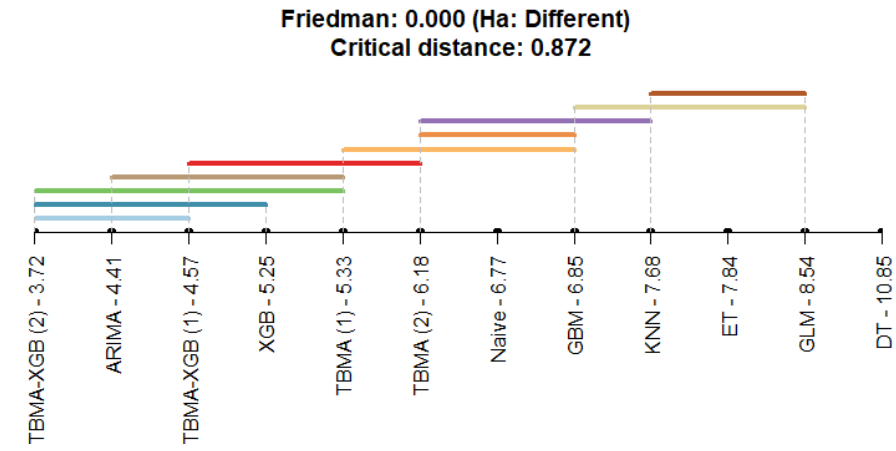
The mean ranks from the Nemenyi tests are shown in Table 4.6. The models of one-days-ahead (Figure 4.4a), four-days-ahead (Figure 4.4b), and seven-days-ahead (Figure 4.4c) forecasts are grouped according to mean rank differences in Figure 4.4. The distribution of the ranks in all forecasting horizons indicates that the TBMA-XGB (2) model has the best rank among other models. However, no significant performance difference between the TBMA-XGB (2), and TBMA-XGB (1) model is found in all forecasting horizons except for the three-days-ahead forecasting. In three-days-ahead forecasting, the TBMA-XGB (2) model performs significantly better than the other models. The TBMA (1) model also takes place in the group of best performing models in five to seven-days-ahead forecasts.

The Nemenyi tests show that the TBMA-XGB (2) model performs significantly better than the XGB model in all forecasting horizons. It can be concluded that when the extracted features are added to the XGB model, the accuracy increases. The mean rank differences between the TBMA-XGB (1) model and the XGB model show the superiority of the TBMA-XGB (1) model in four-days-ahead, six-days-ahead, and seven-days-ahead forecasts. In other forecasting horizons, no significant performance difference is observed, although the TBMA-XGB (1) model has a lower rank than the XGB model in all forecasting horizons.

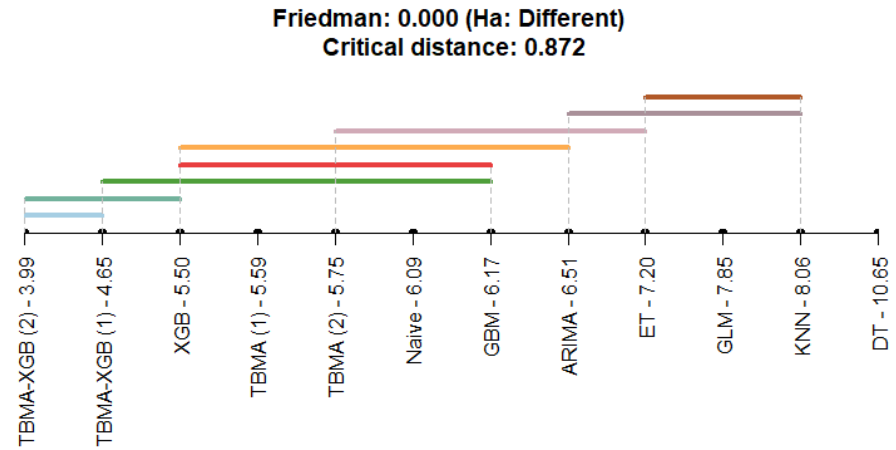
The Nemenyi tests also indicate no significant differences between TBMA (1) and TBMA (2) models, although the TBMA (1) model's rank is lower than the TBMA (2) model's rank in all forecasting horizons. It means that the two levels of the order parameter do not have a significant effect on the experiment results. This feature of the TBMA model is useful for the model building process because users do not need to put too much effort into the tuning process of the order parameter.

Another conclusion from the Nemenyi tests is that the difference between TBMA models and ET model are significant in every forecasting horizon. Both of the TBMA (1) and TBMA (2) models overperform the ET model despite using the same decision trees (i.e., the same parameters for fitting decision trees and ensemble processes).

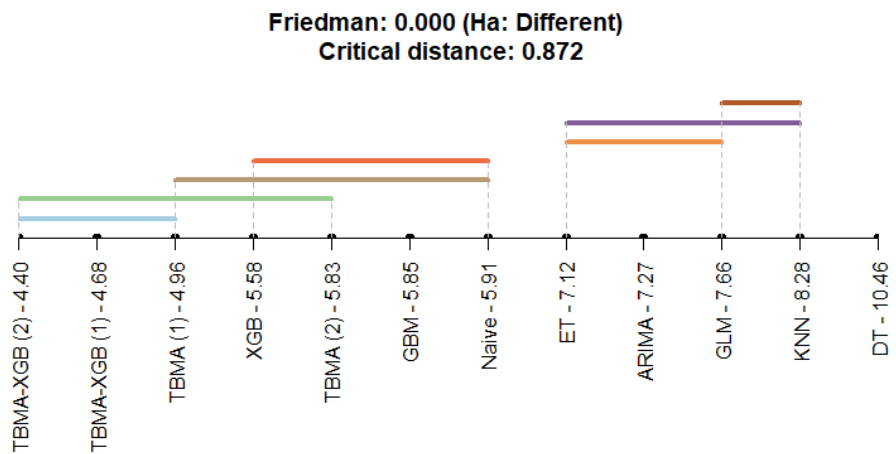
The graph of the daily WMAPE values of one-day-ahead forecasts in June 2018 is shown in Figure 4.5 to see the performances of the models in Ramadan holiday (i.e., 15.06.2018-17.06.2018). According to Figure 4.5, the WMAPE values of all models increases during Ramadan holiday due to the sudden change in consumption behavior. Among all models, the KNN, DT, ARIMA, and naive models have the highest WMAPE scores in Ramadan holiday. The reasons for ARIMA and naive are that the ARIMA model is univariate, and the naive model is incapable of capturing special day effects with the consumption value in the previous week. The WMAPE values of all models also increase one week after Ramadan holiday because of the effect of the previous week's consumption on the models' forecasts.



a



b



c

Figure 4.4. The average ranks of all models for one-day-ahead forecasts (4.4a), four-days-ahead forecasts (4.4b), and seven-days-ahead forecasts (4.4c). The critical distance at significance level 0.05 is 0.872.

Table 4.6. The average ranks of all models based on daily WMAPE scores. The numbers in parentheses indicate the order of the average ranks.

Method	Forecast horizon (days)						
	1	2	3	4	5	6	7
TBMA-XGB (2)	3.72 (1)	3.92 (1)	3.93 (1)	3.99 (1)	4.44 (1)	4.50 (1)	4.4 (1)
ARIMA	4.41 (2)	5.76 (5)	6.39 (8)	6.51 (8)	6.87 (8)	7.15 (9)	7.27 (9)
TBMA-XGB (1)	4.57 (3)	4.62 (2)	4.76 (2)	4.65 (2)	4.82 (2)	4.64 (2)	4.68 (2)
XGB	5.25 (4)	5.24 (3)	5.11 (3)	5.50 (3)	5.44 (4)	5.57 (5)	5.58 (4)
TBMA (1)	5.33 (5)	6.02 (6)	5.8 (4)	5.59 (5)	5.15 (3)	4.85 (3)	4.96 (3)
TBMA (2)	6.18 (6)	5.66 (4)	5.93 (5)	5.75 (4)	5.61 (5)	5.63 (4)	5.83 (5)
Naive	6.77 (7)	6.38 (8)	6.13 (6)	6.09 (6)	5.95 (6)	5.98 (6)	5.91 (7)
GBM	6.85 (8)	6.43 (7)	6.21 (7)	6.17 (7)	6.07 (7)	6.06 (7)	5.85 (6)
KNN	7.68 (9)	7.78 (10)	7.92 (11)	8.06 (11)	8.18 (11)	8.24 (11)	8.28 (11)
ET	7.84 (10)	7.38 (9)	7.24 (9)	7.2 (9)	7.13 (9)	7.08 (8)	7.12 (8)
GLM	8.54 (11)	8.09 (11)	7.88 (10)	7.85 (10)	7.79 (10)	7.75 (10)	7.66 (10)
DT	10.85 (12)	10.72 (12)	10.70 (12)	10.65 (12)	10.55 (12)	10.55 (12)	10.46 (12)

4.2.3. Probabilistic Forecasts

The result of the mean pinball loss comparisons in Figure 4.6 indicates that the TBMA model has the lowest mean pinball loss values in all forecasting horizons except one-day-ahead forecasts, which ETS provides the least mean pinball loss value. Although the TBMA model provided the least scores, there are slight differences between the mean pinball loss values of the ET and TBMA models. The ET and TBMA models are better than other models in terms of providing low mean pinball loss scores with a steady behavior over the forecasting horizons. The reason is that the ET and TBMA models utilize predictors, whereas the other models only depend on the autocorrelation in consumption behavior. By using the predictors, the ET and TBMA models capture the irregularities in consumption behavior better than the other models and provide more accurate probabilistic forecasts.

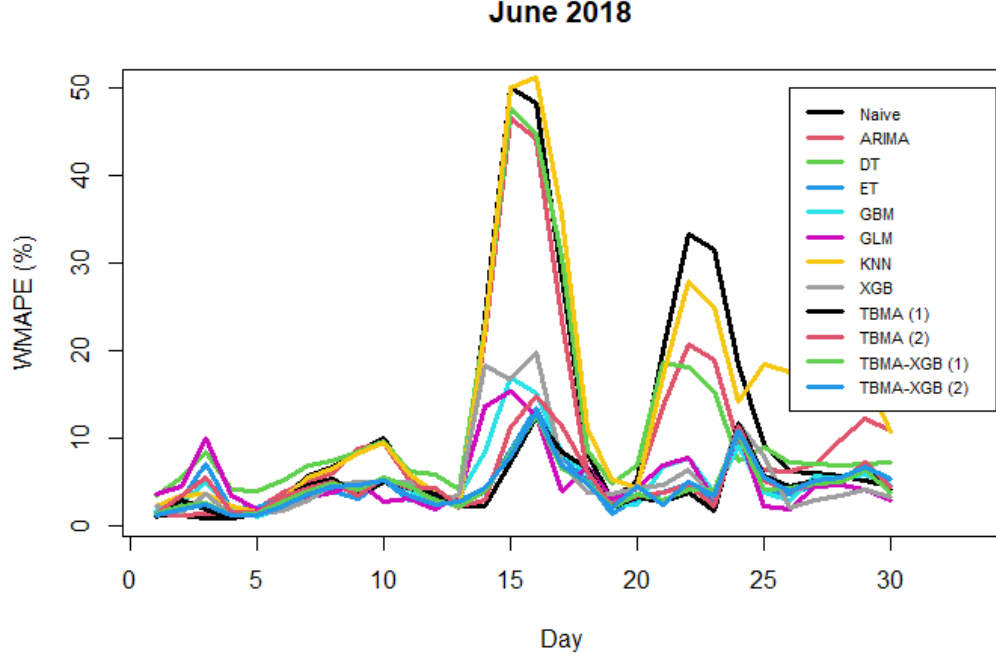


Figure 4.5. Graph of one-day-ahead daily WMAPE values of the models in June 2018. Ramadan holiday is between 15.06.2018-17.06.2018.

4.2.4. Complexity Analysis

The experiments in this study are conducted on Ubuntu operating system with 64GB RAM and 16 cores (Intel 2nd Gen Xeon Scalable Cascade Lake).

A sample from Turkey's electricity load data, containing 1000 observations and 100 features, is used for the empirical analysis. The effects of the number of observations and predictors on feature extraction times are tested with a parameter setting that leads to a longer computation time than those used for experiments in this study. The number of decision trees is set to 500, the minimum node size is set two, the order is selected as two, and the number of features to be considered in each splitting is chosen as 15. No group identity or horizon parameters is used because these parameters do not add considerable complexity to the algorithm. The observations are chosen from the end of the data with proportions of 0.2, 0.4, 0.6, 0.8, and 1. Features are randomly selected with the same proportions for each run. The sample for each run includes ten test observations at the end of the sample. The runs are repeated ten times for a

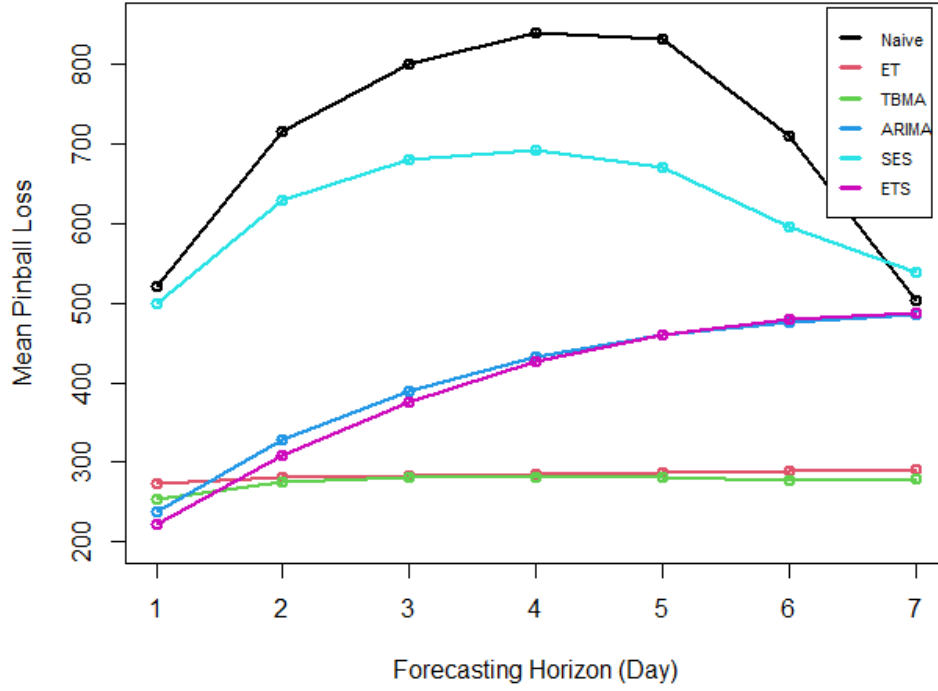


Figure 4.6. Graph of mean pinball loss values in 2018.

reliable analysis. The average feature extraction time for each sample is presented in Figure 4.7.

Figure 4.7 shows that feature extraction time follows a logarithmic growth as the number of observations increase. Run time is proportional to the number of features. Time differences between proportions of features grow as the number of observations increases.

A large part of the complexity of TBMA forecasting and feature construction model comes from the decision tree building process, and relatively shorter time is spent on MA and additional rule-based operations. As already explained, TBMA model creates multiple features by using combinations of its parameters, and TBMA model's parameters consist of parameters that are related to the decision-tree building except for three other parameters, which are group identity, horizon, and order. When features are extracted by using combinations of the parameters not related to the decision trees' building process, the same decision trees can be utilized. Therefore,

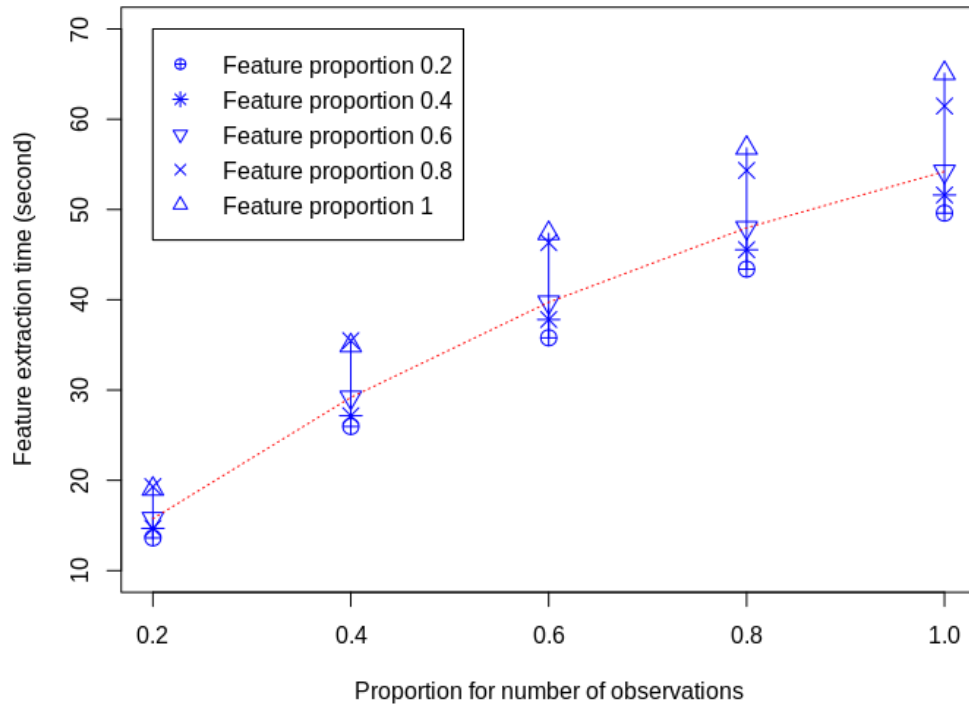


Figure 4.7. Feature extraction time with multiple proportions of observations and features. The analyses are conducted on a sample from Turkey’s electricity load data with 10000 observations and 100 features.

decision trees are created once, and the run time for fitting the same decision trees several times is saved.

TBMA feature construction algorithm and TBMA forecasting algorithms follow similar procedures, which are building decision trees and applying additional rule-based operations. The difference is that the feature extraction model predicts the whole dataset, and the forecasting model predicts only the test dataset. Although this difference does not considerably differentiate their complexities, the point and probabilistic forecasting algorithms potentially have lower complexity than the feature extraction algorithm.

5. CONCLUSION

A new tree-based ensemble model, which is called TBMA, is proposed for time series forecasting problems with a large number of predictors. TBMA model provides point and probabilistic forecasts as well as working as a feature extraction method. The main idea of TBMA model is to integrate MA model into a tree-based ensemble model so that it can handle a large number of predictors and consider autocorrelation in time series data without complicating the computation. In this way, TBMA model makes accurate forecasts and provides useful features to increase the accuracy of forecasting models and reduce the data dimension.

The performances of TBMA model in point forecasting, probabilistic forecasting, and feature extraction are tested on a STL problem. The hourly electricity consumption values of Turkey in 2018 are predicted from one-day-ahead to seven-days-ahead by using statistical time series, and machine learning models. The TBMA-XGB (2) model, which is an XGB model enhanced by the TBMA feature extraction method, is found to be the best performing model in every performance measure of point forecasting. The analysis of probabilistic forecasts shows that the TBMA model provides the lowest mean pinball loss scores in all forecasting horizons except one-day-ahead forecasts in which the ETS model provides the lowest mean pinball loss score.

Further analyses of point forecasting indicate that the TBMA feature extraction method increases the XGB model's accuracy when the extracted features are included in a forecasting model alone or along with the old features. However, using the extracted features together with the old features performs better than using the extracted features alone. Moreover, the TBMA models in the experiment overperform the ET model that uses the same decision trees as a result of integrating autoregressive components into a tree-based ensemble model. Lastly, TBMA model is found to be insensitive to its order parameter with levels one and two. Considering that a big part of

the TBMA's parameters is the same with the tree-based ensemble models' parameters, TBMA model does not require much effort to tune its parameters.

There are several limitations to this study. First, the MA part of TBMA model may lack in capturing the dynamics of a time series dataset because of the simple linearity assumption of the MA model. Thus, the MA part can be altered with other time series models without complicating the computation. Second, the sensitivity of the TBMA model to its parameters can be analyzed in different parameter scenarios. Lastly, the proposed model can be implemented in various benchmark problems to analyze its performance from a broader perspective.

REFERENCES

1. Stock, J. H. and M. W. Watson, “Chapter 10 Forecasting with Many Predictors”, Vol. 1 of *Handbook of Economic Forecasting*, pp. 515 – 554, Elsevier, 2006.
2. Qi, M. and G. P. Zhang, “An investigation of model selection criteria for neural network time series forecasting”, *European Journal of Operational Research*, Vol. 132, No. 3, pp. 666 – 680, 2001.
3. Ghoddusi, H., G. G. Creamer and N. Rafizadeh, “Machine learning in energy economics and finance: A review”, *Energy Economics*, Vol. 81, pp. 709 – 727, 2019.
4. Niu, D., Y. Wang and D. D. Wu, “Power load forecasting using support vector machine and ant colony optimization”, *Expert Systems with Applications*, Vol. 37, No. 3, pp. 2531 – 2539, 2010.
5. Breiman, L., “Random Forests”, *Machine Learning*, Vol. 45, No. 1, pp. 5–32, 2001.
6. Geurts, P., D. Ernst and L. Wehenkel, “Extremely Randomized Trees”, *Machine Learning*, Vol. 63, pp. 3–42, 2006.
7. Chen, T. and C. Guestrin, “Xgboost: A scalable tree boosting system”, *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 785–794, 2016.
8. Cutler, A., D. R. Cutler and J. R. Stevens, “Tree-based methods”, *High-Dimensional Data Analysis in Cancer Research*, pp. 1–19, Springer, 2009.
9. Sheikhan, M. and N. Mohammadi, “Neural-based electricity load forecasting using hybrid of GA and ACO for feature selection”, *Neural Computing and Applications*, Vol. 21, No. 8, pp. 1961–1970, 2012.

10. Meng, M., D. Niu and W. Sun, “Forecasting Monthly Electric Energy Consumption Using Feature Extraction”, *Energies*, Vol. 4, No. 10, p. 1495–1507, 2011.
11. Sadaei, H. J., P. C. [de Lima e Silva], F. G. Guimarães and M. H. Lee, “Short-term load forecasting by using a combined method of convolutional neural networks and fuzzy time series”, *Energy*, Vol. 175, pp. 365 – 377, 2019.
12. Tong, C., J. Li, C. Lang, F. Kong, J. Niu and J. J. Rodrigues, “An efficient deep model for day-ahead electricity load forecasting with stacked denoising auto-encoders”, *Journal of Parallel and Distributed Computing*, Vol. 117, pp. 267 – 273, 2018.
13. Taylor, J. W. and R. Buizza, “Using weather ensemble predictions in electricity demand forecasting”, *International Journal of Forecasting*, Vol. 19, No. 1, pp. 57 – 70, 2003.
14. Lahouar, A. and J. B. H. Slama, “Day-ahead load forecast using random forest and expert input selection”, *Energy Conversion and Management*, Vol. 103, pp. 1040 – 1051, 2015.
15. Ho, K.-L., Y.-Y. Hsu, C.-F. Chen, T.-E. Lee, C.-C. Liang, T.-S. Lai and K.-K. Chen, “Short term load forecasting of Taiwan power system using a knowledge-based expert system”, *IEEE Transactions on Power Systems*, Vol. 5, No. 4, pp. 1214–1221, 1990.
16. Rahman, S. and O. Hazim, “Load forecasting for multiple sites: development of an expert system-based technique”, *Electric Power Systems Research*, Vol. 39, No. 3, pp. 161–170, 1996.
17. Rahman, S. and O. Hazim, “A generalized knowledge-based short-term load-forecasting technique”, *IEEE Transactions on power systems*, Vol. 8, No. 2, pp. 508–514, 1993.

18. Karimi, M., H. Karami, M. Gholami, H. Khatibzadehazad and N. Moslemi, “Priority index considering temperature and date proximity for selection of similar days in knowledge-based short term load forecasting method”, *Energy*, Vol. 144, pp. 928 – 940, 2018.
19. Alfares, H. and N. Mohammad, “Electric load forecasting: Literature survey and classification of methods”, *International Journal of Systems Science - IJSySc*, Vol. 33, pp. 23–34, 2002.
20. Istanbul, E. E., *Turkish Hourly Electricity Consumption*, 2020, <https://seffaflik.epias.com.tr/transparency/>, accessed in March 2020.
21. Sel, B., *Tree-Based Moving Average Application to Electricity Load Forecasting*, 2020, <https://github.com/treebasedmovingaverage/TBMA>, accessed in March 2020.
22. Sel, B. and M. G. Baydoğan, *R Package: tbma*, 2020, <https://CRAN.R-project.org/package=tbma>, accessed in April 2020.
23. Breiman, L., *Classification and Regression Trees*, CRC Press, 2017.
24. Friedman, J., T. Hastie and R. Tibshirani, *The elements of statistical learning*, Vol. 1, Springer, 2001.
25. Breiman, L., “Bagging Predictors”, *Machine Learning*, Vol. 24, No. 2, pp. 123–140, 1996.
26. Tin Kam Ho, “The random subspace method for constructing decision forests”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 8, pp. 832–844, 1998.
27. Schapire, R. E. and Y. Freund, *Boosting: Foundations and Algorithms*, The MIT Press, 2012.

28. Friedman, J. H., “Greedy Function Approximation: A Gradient Boosting Machine”, *The Annals of Statistics*, Vol. 29, No. 5, pp. 1189–1232, 2001.
29. Fan, S. and R. J. Hyndman, “Short-Term Load Forecasting Based on a Semi-Parametric Additive Model”, *IEEE Transactions on Power Systems*, Vol. 27, No. 1, pp. 134–141, 2012.
30. Xiao, L., W. Shao, T. Liang and C. Wang, “A combined model based on multiple seasonal patterns and modified firefly algorithm for electrical load forecasting”, *Applied Energy*, Vol. 167, pp. 135 – 153, 2016.
31. Dedinec, A., S. Filiposka, A. Dedinec and L. Kocarev, “Deep belief network based electricity load forecasting: An analysis of Macedonian case”, *Energy*, Vol. 115, pp. 1688 – 1700, 2016.
32. Aneiros, G., J. Vilar and P. Raña, “Short-term forecast of daily curves of electricity demand and price”, *International Journal of Electrical Power Energy Systems*, Vol. 80, pp. 96 – 108, 2016.
33. Arora, S. and J. W. Taylor, “Short-Term Forecasting of Anomalous Load Using Rule-Based Triple Seasonal Methods”, *IEEE Transactions on Power Systems*, Vol. 28, No. 3, pp. 3235–3242, 2013.
34. Feng, Y. and S. Ryan, “Day-ahead hourly electricity load modeling by functional regression”, *Applied Energy*, Vol. 170, pp. 455–465, 2016.
35. Pielow, A., R. Sioshansi and M. C. Roberts, “Modeling short-run electricity demand with long-term growth rates and consumer price elasticity in commercial and industrial sectors”, *Energy*, Vol. 46, No. 1, pp. 533 – 540, 2012.
36. Taieb, S. B. and R. J. Hyndman, “A gradient boosting approach to the Kaggle load forecasting competition”, *International Journal of Forecasting*, Vol. 30, No. 2, pp. 382 – 394, 2014.

37. Niu, D.-x., H.-f. Shi and D. D. Wu, “Short-term load forecasting using bayesian neural networks learned by Hybrid Monte Carlo algorithm”, *Applied Soft Computing*, Vol. 12, No. 6, pp. 1822–1827, 2012.
38. Çevik, H. H. and M. Çunkaş, “Short-term load forecasting using fuzzy logic and ANFIS”, *Neural Computing and Applications*, Vol. 26, No. 6, pp. 1355–1367, 2015.
39. Lin, C.-T. and L.-D. Chou, “A novel economy reflecting short-term load forecasting approach”, *Energy Conversion and Management*, Vol. 65, pp. 331 – 342, 2013.
40. Shang, H. L., “Functional time series approach for forecasting very short-term electricity demand”, *Journal of Applied Statistics*, Vol. 40, No. 1, pp. 152–168, 2013.
41. Devaine, M., P. Gaillard, Y. Goude and G. Stoltz, “Forecasting electricity consumption by aggregating specialized experts”, *Machine Learning*, Vol. 90, No. 2, pp. 231–260, 2013.
42. Tarsitano, A. and I. L. Amerise, “Short-term load forecasting using a two-stage sarimax model”, *Energy*, Vol. 133, pp. 108 – 114, 2017.
43. Bessec, M. and J. Fouquau, “Short-run electricity load forecasting with combinations of stationary wavelet transforms”, *European Journal of Operational Research*, Vol. 264, No. 1, pp. 149 – 164, 2018.
44. Yeom, C.-U. and K.-C. Kwak, “Short-Term Electricity-Load Forecasting Using a TSK-Based Extreme Learning Machine with Knowledge Representation”, *Energies*, Vol. 10, pp. 1–18, 2017.
45. Mahrami, M., R. Rahmani, M. Seyedmahmoudian, R. Mashayekhi, H. Karimi and E. Hosseini, “A hybrid metaheuristic technique developed for hourly load forecasting”, *Complexity*, Vol. 21, No. S1, pp. 521–532, 2016.

46. Migon, H. S. and L. C. Alves, “Multivariate dynamic regression: modeling and forecasting for intraday electricity load”, *Applied Stochastic Models in Business and Industry*, Vol. 29, No. 6, pp. 579–598, 2013.
47. Lüy, M., V. Ates, H. Polat and E. Çam, “Short-Term Fuzzy Load Forecasting Model Using Genetic–Fuzzy and Ant Colony–Fuzzy Knowledge Base Optimization”, *Applied Sciences*, Vol. 8, pp. 1–30, 2018.
48. Arora, S. and J. W. Taylor, “Rule-based autoregressive moving average models for forecasting load on special days: A case study for France”, *European Journal of Operational Research*, Vol. 266, No. 1, pp. 259 – 268, 2018.
49. Caro, E., J. Juan and J. Cara, “Periodically correlated models for short-term electricity load forecasting”, *Applied Mathematics and Computation*, Vol. 364, pp. 1–13, 2020.
50. Sigauke, C., M. M. Nemukula and D. Maposa, “Probabilistic Hourly Load Forecasting Using Additive Quantile Regression Models”, *Energies*, Vol. 11, No. 9, pp. 1–21, 2018.
51. Santra, A. S. and J.-L. Lin, “Integrating long short-term memory and genetic algorithm for short-term load forecasting”, *Energies*, Vol. 12, No. 11, pp. 1–11, 2019.
52. Caporin, M., F. Fontini and P. Magistris, “Price Convergence within and between the Italian Electricity Day-Ahead and Dispatching Services Markets”, *SSRN Electronic Journal*, pp. 1–21, 2018.
53. Shah, I., H. Iftikhar, S. Ali and D. Wang, “Short-Term Electricity Demand Forecasting Using Components Estimation Technique”, *Energies*, Vol. 12, No. 13, pp. 1–17, 2019.
54. Kulkarni, S., S. P. Simon and K. Sundareswaran, “A spiking neural network (SNN)

- forecast engine for short-term electrical load forecasting”, *Applied Soft Computing*, Vol. 13, No. 8, pp. 3628 – 3635, 2013.
55. Bakirtzis, A. G., V. Petridis, S. J. Kiartzis, M. C. Alexiadis and A. H. Maissis, “A neural network short term load forecasting model for the Greek power system”, *IEEE Transactions on Power Systems*, Vol. 11, No. 2, pp. 858–863, 1996.
 56. Wu, C.-H., G.-H. Tzeng and R.-H. Lin, “A Novel hybrid genetic algorithm for kernel function and parameter optimization in support vector regression”, *Expert Systems with Applications*, Vol. 36, No. 3, Part 1, pp. 4725 – 4735, 2009.
 57. Kavousi-Fard, A., H. Samet and F. Marzbani, “A new hybrid Modified Firefly Algorithm and Support Vector Regression model for accurate Short Term Load Forecasting”, *Expert Systems with Applications*, Vol. 41, No. 13, pp. 6047 – 6056, 2014.
 58. Elattar, E. E., J. Goulermas and Q. H. Wu, “Electric Load Forecasting Based on Locally Weighted Support Vector Regression”, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, Vol. 40, No. 4, pp. 438–447, 2010.
 59. Moon, J., K. Kim, Y. Kim and E. Hwang, “A Short-Term Electric Load Forecasting Scheme Using 2-Stage Predictive Analytics”, *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pp. 219–226, 2018.
 60. Li, C., X. Zheng, Z. Yang and L. Kuang, “Predicting Short-Term Electricity Demand by Combining the Advantages of ARMA and XGBoost in Fog Computing Environment”, *Wireless Communications and Mobile Computing*, Vol. 2018, pp. 1–18, 2018.
 61. Ahmad, M. W., M. Mourshed and Y. Rezgui, “Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy

- consumption”, *Energy and Buildings*, Vol. 147, pp. 77 – 89, 2017.
62. Salami, M., F. M. Sobhani and M. S. Ghazizadeh, “A hybrid short-term load forecasting model developed by factor and feature selection algorithms using improved grasshopper optimization algorithm and principal component analysis”, *Electrical Engineering*, Vol. 102, pp. 437–460, 2020.
 63. Zhang, G. and J. Guo, “A Novel Method for Hourly Electricity Demand Forecasting”, *IEEE Transactions on Power Systems*, Vol. 35, No. 2, pp. 1351–1363, 2019.
 64. Susto, G. A., A. Schirru, S. Pampuri and S. McLoone, “Supervised Aggregative Feature Extraction for Big Data Time Series Regression”, *IEEE Transactions on Industrial Informatics*, Vol. 12, No. 3, pp. 1243–1252, 2016.
 65. Hyndman, R. J., E. Wang and N. Laptev, “Large-Scale Unusual Time Series Detection”, *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pp. 1616–1619, 2015.
 66. Hyndman, R. and Y. Khandakar, “Automatic Time Series Forecasting: The forecast Package for R”, *Journal of Statistical Software*, Vol. 27, No. 3, pp. 1–22, 2008.
 67. Demšar, J., “Statistical Comparisons of Classifiers over Multiple Data Sets”, *J. Mach. Learn. Res.*, Vol. 7, p. 1–30, 2006.
 68. Friedman, M., “The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance”, *Journal of the American Statistical Association*, Vol. 32, No. 200, pp. 675–701, 1937.
 69. Nemenyi, P., “Distribution-free multiple comparisons (Doctoral Dissertation, Princeton University, 1963)”, *Dissertation Abstracts International*, Vol. 25, No. 2, 1963.

APPENDIX A: PARAMETER TUNING WITH CROSS-VALIDATION

Table A.1. Model parameters selected in the cross-validation process.

Model	Parameter	Candidate Levels	Selected Level
GBM	Learning rate	0.1	0.1
	Number of splits in a tree	1, 2, 3	3
	Minimum node size	10	10
	Number of trees	50, 100, 150	50
GLM	Parameter for regularization distribution	0.1, 0.55, 1	0.1
	Regularization strength	8.4, 84.8, 848.4	848.4
DT	Minimum cost improvement of a node for further splitting	0.05, 0.1, 0.52	0.05
KNN	Number of the nearest observations	5, 7, 9	9
XGB	Step size shrinkage	0.3, 0.4	0.3
	Maximum depth of a tree	1, 2, 3	3
	Minimum loss reduction for further partition in a tree	0	0
	Ratio of columns selected when building each decision tree	0.6, 0.8	0.6
	Minimum sum of instance weights for further splitting	1	1
	Sampling ratio of train observations	0.5, 0.75, 1	1
	Number of rounds	50, 100, 150	150

Table A.1. Model parameters selected in the cross-validation process. (cont.)

Model	Parameter	Candidate Levels	Selected Level
TBMA-XGB (1)	Step size shrinkage	0.3, 0.4	0.3
	Maximum depth of a tree	1, 2, 3	2
	Minimum loss reduction for further partition in a tree	0	0
	Ratio of columns selected when building each decision tree	0.6, 0.8	0.6
	Minimum sum of instance weights if splitting occurs	1	1
	Sampling ratio of train observations	0.5, 0.75, 1	0.75
	Number of rounds	50, 100, 150	150
TBMA-XGB (2)	Step size shrinkage	0.3, 0.4	0.3
	Maximum depth of a tree	1, 2, 3	3
	Minimum loss reduction for further partition in a tree	0	0
	Ratio of columns selected when building each decision tree	0.6, 0.8	0.6
	Minimum sum of instance weights if splitting occurs	1	1
	Sampling ratio of train observations	0.5, 0.75, 1	1
	Number of rounds	50, 100, 150	100