Guzel, Tulin; Cinar, Hakan; Cenet, Mehmet Nabi et al.

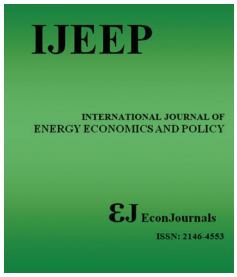
Article

A framework to forecast electricity consumption of meters using automated ranking and data preprocessing

International Journal of Energy Economics and Policy

Provided in Cooperation with:
International Journal of Energy Economics and Policy (IJEEP)

Reference: Guzel, Tulin/Cinar, Hakan et. al. (2023). A framework to forecast electricity consumption of meters using automated ranking and data preprocessing. In: International Journal of Energy Economics and Policy 13 (5), S. 179 - 193. https://www.econjournals.com/index.php/ijeep/article/download/13834/7466/34210. doi:10.32479/ijeep.13834.



International Journal of Energy Economics and Policy

ISSN: 2146-4553

available at http: www.econjournals.com

International Journal of Energy Economics and Policy, 2023, 13(5), 179-193.

A Framework to Forecast Electricity Consumption of Meters using Automated Ranking and Data Preprocessing

Tulin Guzel[1], Hakan Çınar[1], Mehmet Nabi Çenet[1], Kamil Doruk Oguz[1], Ahmet Yucekaya[2]*, Mustafa Hekimoglu[2]

[1]AIMS Analytics Solutions, Istanbul, Turkey, [2]Department of Industrial Engineering, Kadir Has University, Istanbul, Turkey.
*Email: ahmety@khas.edu.tr

Received: 12 November 2022 **Accepted:** 19 July 2023 **DOI:** https://doi.org/10.32479/ijeep.13834

ABSTRACT

Forecasting electricity consumption is crucial for the operation planning of distribution companies and suppliers and for the success of deregulated electricity markets as a whole. Distribution companies often need consumption forecasting for meters to better plan operations and demand fulfillment. Although it is easier to forecast the aggregated demand for a region, meter based demand forecasting brings challenging issues such as non-uniform usage and uncertain customer consumption patterns. The stochastic nature of the demand for electricity, along with parameters such as temperature, humidity, and work habits, eventually causes deviations from the expected demand. In this paper, real meter data from a regional distribution company is used to cluster the customer using their non-uniform usage and automated ranking mechanism is proposed to select the best method to forecast the consumption. The proposed end-to-end methodology includes data processing, missing value detection and filling, abnormal value detection, and mass reading for meters and is applied to regional data for the period 2017-2018 and provides a powerful tool to forecasts the demand in hourly and daily horizons using only the past demand data. Besides proposing effective methodologies for data preprocessing, 10 different regression methods, 7 regressors, 5 machine learning methods that include LSTM and Ar-net models are used to forecast the meter based consumption. The hourly forecasting errors in the demand, in the Mean Absolute Percentage Error (MAPE) norm, are <4% for most customer groups. The meter based forecast is then aggregated to reach a final demand which is then used for operation and demand planning. The proposed framework can be considered reliable and practical in the circumstances needed to make demand and operation decisions.

Keywords: Time Series Analysis, Prediction, Forecasting, Regression, Segmentation, Meter Based Consumption
JEL Classifications: Q47, E17, Q40

1. INTRODUCTION

Demand forecasting has always played an important role in capacity and transmission, generation planning, and pricing. Also, the liberalization and privatization of power markets have increased the importance of demand or load estimation, as market success rates are largely related to their accuracy.

The forecasting of electricity demand or consumption has been studied in numerous studies mostly at an aggregated level for the use of suppliers and system operators. The distribution companies receive power from the suppliers and need to deliver the power to their end customers and meet the demand. Hence, they also need to forecast the total power withdrawn from the meters within their region. The meters in the region are regularly checked for billing purposes and consumption levels. The technological levels for the meters differ, and for the Turkish power market many meters are still regularly checked by an agent to collect the data.

On the other hand, the customers can be classified as household, industrial, commercial, and agricultural users. Each customer's consumption patterns show different characteristics requiring

even more sophisticated forecasting methodologies. Weekday and weekend consumption patterns, peak times, and usage frequencies are different for each customer group. The collected data from the meters might bring some extra issues such as missing data, mass readings, and unexplained abnormal values. Given that the collected data from such a number of meters is huge, a preprocessing of the data is also required.

The forecasting literature for electricity demand and consumption is extensive and it is observed that the studies for consumption forecasting meters are limited. The selected works for general and meter-based forecasting are presented below. Linear models and time series methods are commonly used in the literature for demand forecasting. Anand and Suganti (2012) present literature on forecasting methods including Artificial Neural Networks (ANN), Genetic Algorithms (GA), Support Vector Machines (SVM), and Particle Swarm Optimization (PSO) and other numerical methods. ARMA and ARIMA models are also used to include the stochastic effects for demand forecasting. The electricity demand forecasting methodologies studied in (Andersen et al., 2013; Niu et al., 2010; Lo and Wu, 2003) show that the trends in long and short term forecasts, such as weekly demand patterns and economic growths, are better captured in ARIMA models.

In the literature, time series methods are also combined with other heuristic approaches. Researchers study England and Wales electricity demand data and apply the Holt-Winters method for different periods with an AR model (Taylor and Buizza, 2003). In (Wang et al., 2012), the authors show that the "PSO optimal Fourier method" corrects seasonal ARIMA forecast results, and apply it to the Northwest China electrical network, showing that the combined model's forecasting accuracy is higher than that of single-season ARIMA. Similar works that use ARIMA include different periods for forecasting and demonstrate the effectiveness of the methodology (Ren et al., 2016; Vilar et al., 2012; Filik et al., 2011; Chakhchoukh et al., 2011).

The impact of temperature on electricity demand varies depends on the infrastructure and heating resources, the temperature is used to increase the forecast accuracy though. The different aspects of the influence of the temperature on electricity demand have been analyzed in (Taylor, 2003; De Felice et al., 2013; De Felice et al., 2015; Lusis et al., 2017; Islam et al., 1995; Hor et al., 2005; Momani, 2013; Bašta and Helman, 2013). The seasonal cycles determine the impact of temperature on electricity demand especially if the electricity is used for heating and cooling needs. The impact of temperature can be directly observed on meter based demand. Different studies, classified according to their forecasting methods, are given in Table 1. A similar review is given in (Kök, 2022). The methodologies can be classified as time series analysis, statistical methods, surveys, artificial neural networks and simulation, heuristic approaches, and temperature-based methods. The main assumptions in these studies are to forecast the aggregated demand and not meter based demand.

The main motivation of this research is to forecast the consumption of meters using past consumption data. Forecasting based on individual meter data is more challenging than forecasting aggregated consumption. There are also some studies for meter based demand forecasting that present different aspects of the problem.

Table 1: Overview of the forecasting methods and related resources

Methods	Sources
Time series analysis	(Conejo et al., 2005; Anand and Suganthi, 2012; Clements et al., 2016; Niu et al., 2010; Andersen et al., 2013; Lo and Wu, 2003)
Statistical methods	(Vilar et al., 2012; Taylor, 2010; Fan and Hyndman, 2012; Wang et al., 2012; McSharry et al., 2005; Taylor, 2003; Chakhchoukh et al., 2011; Apadula et al., 2012; Ren et al., 2016; Filik et al., 2011)
Surveys	(Dyner and Larsen, 2001; Anand and Suganthi, 2012; Hahn et al., 2009; Conejo et al., 2005)
Artificial neural network and simulation	(Abumohsen et al., 2023; Zhanga and Dongb, 2001; Wang and Ramsay, 1998; Tarmanini et al., 2023)
Heuristic approaches	(Wang et al., 2012; Zhu et al., 2011; Azadeh et al., 2007; Pai and Hong, 2005)
Temperature based methods	(Taylor and Buizza, 2003; Felice et al., 2013; Felice et al., 2015; Crowley and Joutz, 2003; Lusis et al., 2017; Islam et al., 1995; Hor et al., 2005; Momani, 2013; Bašta and Helman, 2013)

Gajowniczek and Ząbkowski (2016, 2017) present a segmentation approach to forecast the electricity load at individual household levels for smart meters. Ghofrani et al. (2011) use real time measurement data from smart meters to forecast short term demand for a residential customer using Kalman filtering methodology. Arora and Taylor (2016) use conditional kernel density to estimate consumption for smart meters. They aim to estimate probability density for consumption. Similar works such as Wijaya et al. (2015), Hsiao (2014), and Taieb (2016) present methods to forecast and analyze consumption drivers based on smart meter data. Dewangan et al. (2023) present a recent review on load forecasting models. They focus on smart meter data in smart grids. Wang et al. (2023) focus on anomaly detection in real time load forecasting. However, the scope of these researches is not extensive, and they do not include the detailed analysis presented in this research.

In presented studies, mostly residential consumption forecasting is considered and the customer types, nonuniform consumption profiles, and data collection, processing, and clustering have not been included. The customer segmentation, missing value detection and filling, abnormal value detection, and then using numerous methodologies to determine the best methodology are not proposed for the authors' best knowledge. The main research questions of this study are, clustering the customers correctly, proposing the best methods for detecting and filling the missing values, and detecting the abnormal and mass reading values and then proposing an end-to-end novel methodology to forecast the meter based consumption using different algorithms to determine the most suitable alternative. The proposed methodologies that include regressions, regressors, and machine learning methods are extensive. Furthermore, using the test period and determining

the best forecasting methodology for each analysis period, i.e., selecting different methodology for each day using automated ranking mechanism, is unique and novel. The specific objectives of this work can be further described as:

- To develop methodologies to preprocess consumption data from different meters, that will identify mass readings, detect and fill missing values, and identify the abnormal values
- To propose clustering methods to cluster consumers as residential, industrial, commercial, and agricultural then further classify them based on usage profiles
- To use regressions, regressors, and machine learning methods for hourly and daily consumption forecasting of each meter for each period and automated ranking mechanism based on accuracy level
- To combine this end-to-end methodology that includes data collection, processing, clustering, and forecasting to be used for daily operations for consumption forecasting in each meter
- To identify possible implications for system operators and distribution companies.

To achieve the aforementioned objectives, in this work, an extensive end-to-end novel methodology is proposed that include data processing, customer segmentation, missing value detection and filling, abnormal value detection, and regression methods to forecast the meter based consumption using different algorithms. The forecasting methods include numerous regressions, regressors, and machine learning methods, and the best methodology is selected for each analysis period using method shifting and forecast accuracies are evaluated.

In Section 2, an overview of the market mechanism, the data used for the validation of the model, and a discussion of the structure of the daily variation curves are presented. Then, the details of the proposed models are explained in Section 3. Hourly and daily forecasting details for each customer group are given in Section 4. Section 5 presents the conclusion and suggestions for future directions respectively.

2. DATA COLLECTION AND PROCESSING

The Turkish power market has experienced a significant development on both the demand and supply side and no significant shortage has occurred in the last decades. The liberalization process started in the early 2000s and privatization led to new capacity investments in the market. Once the market is settle by the independent system operator, the electricity is generated and transmitted via state owned transmission lines. As a part of the privatization stage in Turkey, the distribution right is also transferred to companies. The country is divided into different regions and distribution companies are responsible for the distribution, maintenance, and billing. Figure 1 shows the regions of the distribution companies in which the distribution regions are represented with different colors.

The distribution companies oversee the operation within their region and are responsible for sufficient electricity to be provided to meet the demand. The demand forecasting is challenging but a required step in this process. Once the demand is forecasted, the required electricity needs to be provided from the spot market or through bilateral contracts.

The electricity market in Turkey is deregulated while household electricity consumption is approximately 68%, and the electricity is not commonly used (about 8.6%) for heating in winter but it is used for cooling needs in summer. The electricity is mainly used for household, industrial, commercial, agricultural, and street lighting purposes.

Although the aggregated consumption for the country can be forecasted using different methods, and there are already studies in the literature, consumption forecasting at a distribution company level requires special efforts. The distribution companies collect data from electric meters for billing and maintenance purposes. Such data is also crucial for consumption forecasting as it is first hand data from the end consumer. The customers in a distribution region include households, commercial units, industrial facilities, irrigation and agricultural meters, and illumination units. Each consumption type has different characteristics and usage patterns. Hence, the collected data from the meters leads to different data structures.

The reading times for meters vary depend on the customer types. A technician still visits most meters, and the consumption level is read and compared with the previous reading to calculate the usage and determine the bill. Although most such meters are visited monthly, some other meters are checked during other periods. There are usually hourly meters for users with high consumption, while for residential users, there are generally monthly meters. It is also possible for some meters to collect the data automatically. The data is sent regularly to a server, or a technician receives the data whenever necessary.

As the consumption data that is collected from the meters are not uniform, careful and detailed data processing should be performed before creating a forecasting model for the meters. The data used in this work is provided by a distribution company. When the meters belonging to different customer groups are analyzed, different usage profiles, unread meters, meters that are not read at the specified time and read collectively, and damaged meters are observed. The data need to be processed to determine the customer type, reading cycle, and data collection periods.

As the raw data is collected from the meters with different technological levels and used by various customer groups, there are some minor issues with the data, such as missing cells and double readings that need to be refined before the forecasting model is built. The data is reviewed and the variables such as temperature and expected consumptions are found to be complete as the temperature is an important parameter for the consumption. The data is grouped based on the customers and located cities for 267 m as given in Table 2.

The meters with missing values need to be treated separately. Reliable past values are essential for appropriate forecasting. The previous 24 h of the data is usually fed to the forecasting models, and the model is tested using the past values. If a

Figure 1: The distribution companies and regions in Turkey (Tedas, 2022)

data series in a meter has some unfilled values, value filling methodologies can be applied using previous and later actual values. However, if too many consecutive readings data are missing for a meter, then this meter should be excluded. Hence, a prescreening process is applied and the meters with a missing series of 240 h are eliminated. Table 3 shows some of the meters that are excluded.

2.1. Abnormal Records and Extreme Values

Apart from the missing values, another problematic issue is the abnormal records, which can directly affect the success of the models. Abnormal records may occur for many reasons. Situations such as mass readings, power outages, meter failures, maintenance, and abnormal consumption are some of the reasons.

These values should be determined and filled with the most appropriate value for a successful operation. If the reasons for the formation of the records are known, they should definitely be specified in the models.

In order to detect abnormal values in the data, standard deviation, interquartile range, Isolation Forest, Minimum Covariant Determinant, and DBSCAN statistical methodologies are used. Each hourly data from each meter was scored with these 5 different methods. If 4 or more models mark the relevant hour as extreme values, those points were accepted as extreme values. Outlier values were detected in 846 h of 175 m. Figure 2 shows a representation of extreme values. The process is replicated for each meter for each analysis period and it is integrated into the overall methodology.

2.2. Missing Value Identification and Filling

After the collective reading of meters is determined, the remaining missing values in meters data are identified. Such missing values need to be filled using proper methodologies. Linear Interpolation, Quadratic Interpolation, Cubic Interpolation, Moving average, and Moving median are used to fill the missing values, and the best method is selected. Figure 3 shows a representation of missing values.

In order to find the most suitable model, 10 different meters were selected randomly from different customer groups and 4 sequential hourly data were randomly removed from each day of the week. Then the proposed models are applied for testing. Table 4 shows Pearson correlation values for selected meters for each method.

When we look at the Pearson Correlation Values and scatter plots of the filled values with the real values, we can see that the filling values of the linear interpolation method are more successful than the other methods. Hence missing values and extreme values in the meters were filled with Linear Interpolation method. Figure 4 shows the scatter plots for filled values.

Figure 2: A view of extreme values for September-October, 2018

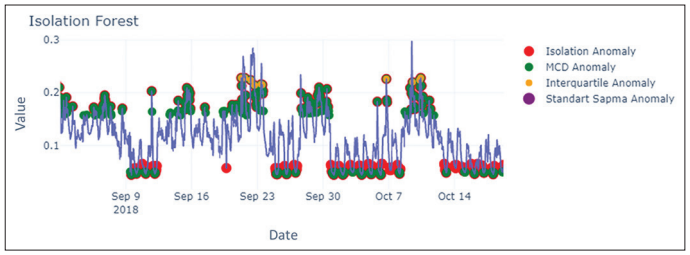


Table 2: The classifications of the meters

Customer	Meters	City	Meters
Commercial	174	City 1	21
Industrial	85	City 2	100
Household	1	City 3	32
Agricultural usage	7	City 4	114
Total	267	Total	267

Table 3: Meter examples with extreme lost values

Meter number	Missing hours	Customer
S203	1673	Industrial
S211	990	Commercial
S254	827	Commercial
S197	753	Commercial
S246	748	Agricultural
S248	744	Agricultural
S59	720	Commercial
S252	240	Industrial
S200	62	Commercial
S110	57	Industrial

3. CUSTOMER SEGMENTATION AND MODEL BUILDING

3.1. Meter Based Customer Segmentation

The consumer types and hence the meters in a distribution region, as well as their profiles are variable. The segmentation of the meters based on usage characteristics is expected to be useful in developing a better forecasting approach. The segmentation is useful and required in meter based forecasting for many essential reasons. It allows to characterize the consumption patterns of each meter as well as to observe the times the consumption is highest or lowest at the meter. Such knowledge can be used in pricing, load distribution, or operation planning such as maintenance or power outage planning. On the other hand, if a meter is new or limited data is available due to less frequent readings, the segmentation allows to develop a common approach to consumption forecasting for these kinds of meters. Segmentation also helps to follow the consumption characteristics of the meters and determine the meters that have changed its consumption characteristics due to some reasons such as power theft and significant consumption decrease. Such meters are identified, and action is taken if necessary.

In order to determine the most efficient segmentation, we consider agglomerative hierarchical clustering, K-means, and mean shift clustering methodologies with different parameters. Agglomerative hierarchical clustering combines similar objects in the dataset starting from (sub)-cluster including only one object. Pairs of clusters are iteratively combined into a larger cluster until the algorithm reaches a cluster of all objects (James et al., 2014). Once a pair of clusters are combined. The similarity of cluster pairs is measured using (weighted) Euclidean distance between *representative points* of each cluster. Depending on the selection of representative point agglomerative hierarchical clustering is employed with single, full, centroid linkage options (James et al., 2014). Outputs of hierarchical clustering algorithms are represented with *dendograms.*

K-means is a well-known iterative algorithm that starts with a given *K* amount of clusters with randomly chosen *centroids* and cluster centers. At each iteration, objects in the dataset are assigned to the closest centroid based on a distance metric. Once the assignment of objects is complete, the centroid of each cluster is updated by taking the average of coordinates of each object assigned to a cluster. This marks the end of an iteration, and the algorithm proceeds to the new iteration with centroids from the previous iteration. It runs until the resulting clusters converge at the end of an iteration.

Mean-Shift is a center-based clustering algorithm that divides the data into groups, taking into account the distribution density. This algorithm is based on the concept of Kernel Density Estimation, which is a way of estimating the probability density function of a random variable. Due to this feature, clusters can be assigned to the data without the need to define the number of clusters.

For the segmentation of hourly meters, we experiment with different parameters of the three clustering algorithms. In these experiments, the data of each meter was selected for 28 days, that is, 4 h from each day of the week. It was observed that the usage of the meters differed according to the consumer groups on weekdays and weekends. While modeling the consumption characteristics, each meter was clustered in two different ways as weekdays and weekends.

The consumption amounts on Monday, Tuesday, Wednesday, Thursday, and Friday were calculated by taking the grand and the hourly total. Weekend segment variables were created by summing and proportioning the overall and hourly energy consumption on Saturday and Sunday. Since segment variables are calculated from the ratio of hourly usage to general usage through normalization, segment variable values are values that do not contain extreme values that range from 0 to 1. Weekday and weekend variables are represented as weekday hour and weekend hour respectively.

We utilize the silhouette coefficient for measuring segmentation performance to compare the outputs of different segmentations.

Table 4: The Pearson correlation values for each method

Meters	Linear	Quadratic	Cubic	Moving average	Moving median
S1	0.97	−0.05	−0.08	0.38	0.43
S11	0.97	−0.43	−0.46	0.82	0.83
S116	0.76	−0.07	−0.07	0.67	0.62
S132	0.90	−0.04	0.00	0.76	0.72
S139	0.70	−0.09	−0.10	0.65	0.64
S21	0.93	0.03	0.05	0.94	0.94
S34	0.91	−0.01	−0.01	0.82	0.80
S345	0.98	0.01	−0.01	0.75	0.61
S40	0.93	−0.33	−0.35	0.81	0.80
S50	0.81	−0.12	−0.14	0.48	0.48

Figure 3: A representation of missing values and filling

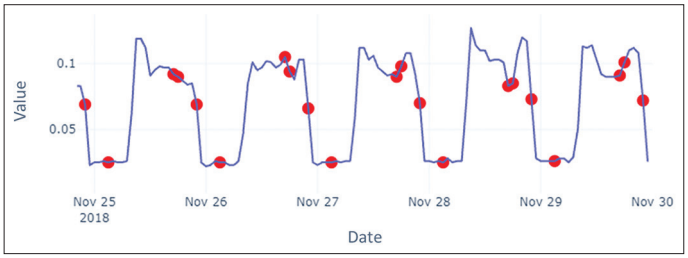
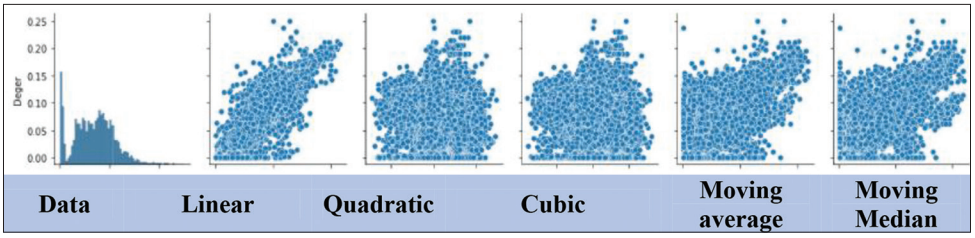


Figure 4: The actual and filled values comparisons



The value of the silhouette coefficient is between [−1, 1] while close to 1 means that the data point is within the cluster it belongs to and far away from other clusters. The worst value is −1 and values close to 0 indicate overlapping clusters. The results of our segmentation experiments are given in Tables 5 and 6 for weekday and weekend consumption, respectively.

Considering the results obtained for the weekday consumption characteristic, agglomerative hierarchical clustering with 4 clusters leads to the highest silhouette coefficient indicating the most compact segments are significantly distinguished from each other.

Considering the results obtained for the weekend consumption characteristics, when both the silhouette statistics and the low number of clusters and whether they distinguish significantly from each other, the agglomerative algorithm with 3 clusters establishes the ideal cluster structure. Hence, the agglomerative hierarchical clustering method is selected for the Energy Consumption Characteristics of the meters. Accordingly, the hourly energy consumption of meters on weekdays and weekends is given in Table 7.

Table 5: Weekday segment scenarios

Methodology	Segments	Parameter	Siluet coefficient
Mean shift	5	Quantile=0.1	0.42
Mean shift	4	Quantile=0.2	0.4
Mean shift	5	Quantile=0.3	0.56
Mean shift	5	Quantile=0.4	0.46
Mean shift	5	Quantile=0.5	0.42
Agglomerative	5	Euclidean-ward	0.24
Agglomerative	4	Euclidean-ward	0.46
Agglomerative	3	Euclidean-ward	0.45
Agglomerative	3	Manhattan-complete	0.29
Agglomerative	4	Manhattan-complete	0.29
Agglomerative	4	Manhattan-complete	0.27
Agglomerative	4	Euclidean-average	0.55
Agglomerative	5	Euclidean-average	0.45
K-means	3		0.41
K-means	4		0.29
K-means	5		0.28

Table 6: Weekend segment scenarios

Methodology	Segments	Parameter	Siluet coefficient
Mean shift	7	Quantile=0.1	0.51
Mean shift	6	Quantile=0.2	0.52
Mean shift	3	Quantile=0.3	0.69
Mean shift	5	Quantile=0.4	0.66
Mean shift	6	Quantile=0.5	0.5
Mean shift	6	Quantile=0.6	0.47
Mean shift	2	Quantile=0.7	0.71
Agglomerative	4	Euclidean-ward	0.35
Agglomerative	5	Euclidean-ward	0.29
Agglomerative	3	Euclidean-ward	0.35
Agglomerative	3	Manhattan-complete	0.56
Agglomerative	4	Manhattan-complete	0.50
Agglomerative	5	Manhattan-complete	0.50
Agglomerative	3	Euclidean-average	0.67
Agglomerative	5	Euclidean-average	0.59
K-means	3		0.39
K-means	4		0.26
K-means	5		0.28

The agglomerative hierarchical clustering is applied to meters and the meters are clustered based on consumption. The distribution of meters on weekdays and weekends is given in Table 8.

3.2. Methodology and Modelling

The collected data from hourly meters include missing values, mass readings, and abnormal values and meters might differ from each other in terms of the total amount of consumption or hourly consumption patterns. The proposed methodology detects whether there are missing values, mass readings, abnormal values and fill or remove the data points and then applies the consumption clustering for the meters. Then forecasting methodologies are used for hourly and daily forecasts. Figure 5 presents the pseudo code of the proposed methodology.

The proposed methodology uses different forecasting techniques and selects the best one based on an automated ranking mechanism. This selection process is updated for each forecasting period, hence the selected methodology might be different for each hour or day. The bias, variance and feature selection should be carefully taken into account in this process. Any predictive model includes bias, variance, and irreducible errors. Depending on the model complexity bias-variance trade-off might be different over various machine learning methods. The irreducible error stems from exogenous randomness which cannot be removed from the system (James et al., 2014).

In machine learning and statistics, feature selection, also known as variable selection, feature selection, or variable subset selection, is the process of selecting a subset of related features (variables, predictors) for use in a model setup. Feature selection techniques are used to simplify models so that their outputs can be easily interpreted, shorten training times, and solving the size problem (Liu and Motoda, 1998). The proposed methodology employs different regression methods from statistics and machine learning literature with various regressors. An overview of the regression methods employed in our study is given in Table 9 below.

The regression methodologies are commonly used for forecasting. The electricity consumption is correlated to different parameters such as temperature, and hence regression methodologies provide quite good results (Draper and Smith, 1998). Although linear regression is the most common methodology, ridge regression (Hoerl et al., 1970), lasso regression (Tibshirani, 1996), elastic net regression (Zou and Hastie, 2005), least angle regression (Efron, 2004), huber regression (Huber, 1964), orthogonal matching pursuit (Pati et al., 1993), Bayesian ridge regression (Grinstead and Snell, 2006), and regression with decision trees algorithms (Quinlan, 1987) are different combinations and provide forecasting results depend on the nature of the data and industry.

The usage of regressors can lead to successful forecasts especially for industries in which the demand is related to dominant parameters. Industrial consumption and climatic factors can be considered as such. Hence, AdaBoost Regressor (Wyner et al., 2017), Random Forest Regressor (Wyner et al., 2017), Gradient Boosting Regressor (Hastie et al., 2009), CatBoost Regressor (Hastie et al., 2009), Extra Trees Regressor, K Neighbors Regressor (Altman, 1992), and Passive Aggressive Regressor (Blondel, 2014) are included as forecasting methodologies.

Table 7: Consumption characteristics based on agglomerative clustering

Hours	Weekday consumption				Weekend consumption		
	Equal	Double Peak	Daytime	Morning	Equal	Double Peak	Night
0	3.8	0.4	2.3	3.6	3.9	0.9	14.9
1	3.8	0.3	2.2	4.0	3.9	0.9	15.1
2	3.7	0.3	2.2	4.2	3.8	0.9	14.2
3	3.7	0.3	2.2	5.5	3.8	0.8	14.4
4	3.7	0.4	2.3	7.1	3.8	1.0	13.8
5	3.8	0.6	2.5	7.6	3.8	2.0	5.7
6	4.1	4.1	2.6	7.1	3.9	4.9	1.8
7	4.3	7.8	5.2	6.5	4.2	7.8	0.9
8	4.5	9.9	6.0	6.2	4.5	9.1	1.1
9	4.6	10.4	6.3	6.1	4.5	9.5	1.0
10	4.5	9.4	6.3	5.8	4.5	8.9	1.0
11	4.4	4.9	6.1	5.1	4.4	5.9	0.9
12	4.4	7.5	6.3	4.3	4.3	7.6	0.9
13	4.4	10.1	6.2	3.3	4.3	9.0	0.8
14	4.4	10.2	5.9	2.6	4.3	8.6	0.8
15	4.3	10.2	5.6	2.2	4.3	7.9	0.8
16	4.3	7.6	5.0	2.2	4.4	5.9	0.7
17	4.4	2.3	4.5	2.2	4.5	2.1	0.6
18	4.4	1.0	4.1	2.2	4.4	1.4	0.7
19	4.3	0.7	3.8	2.2	4.4	1.1	1.0
20	4.2	0.5	3.3	2.2	4.2	1.0	1.6
21	4.1	0.4	2.9	2.3	4.1	1.1	2.0
22	4.1	0.4	2.7	2.5	4.0	0.9	2.7
23	4.0	0.3	2.6	2.8	3.9	0.8	2.6

Table 8: The distribution of meters based on the consumption

Segment	Weekend	Weekday
Equal consumption	174	224
Double peak	14	21
Daytime consumption	53	
Morning consumption	7	
Night consumption		3

Additionally the machine learning methodologies might fit forecasting meter based electricity forecasting context. Extreme Gradient Boosting (XGBOOST), Light Gradient Boosting Machine (Hastie et al., 2009), Long Short Term Memory (Hochreiter and Schmidhuber, 1997), and Ar-net (Triebe et al., 2019) are also used for the consumption forecasting. Finally, a methodology that compare and combines the results of all algorithms and provide a new forecast is also included.

In the assessment of the performance of the forecasting methodologies, Mean Absolute Percentage Error (MAPE) is used. If S_h and y_h are the actual demand and the forecast demand for hour h, then MAPE can be defined as given in Eq(1) and:

Table 9: Algorithms and forecasting methodologies

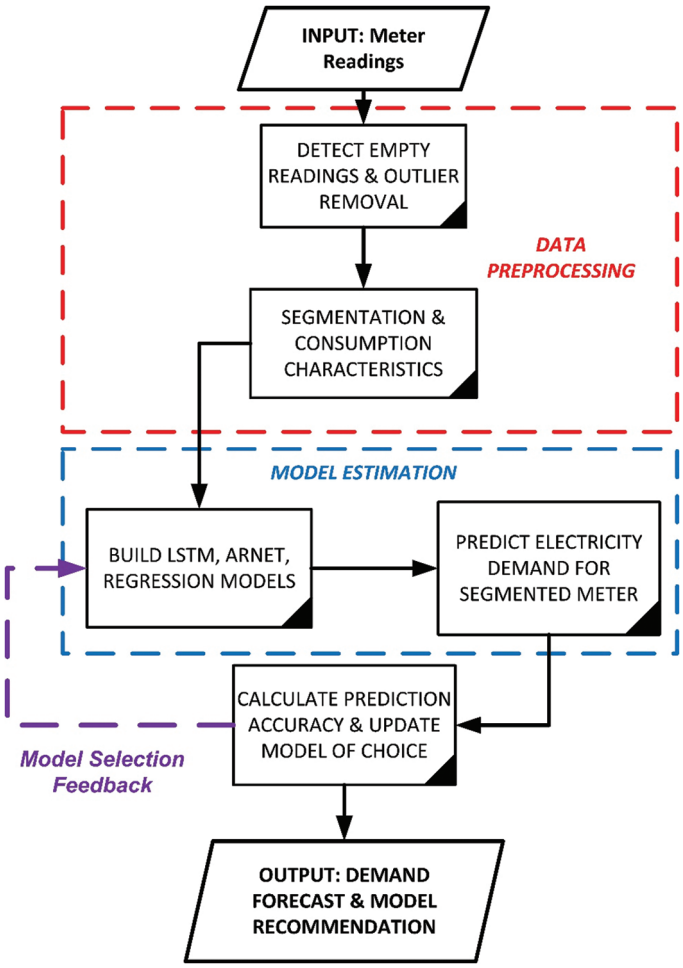
Regressions	Regressors	Machine learning methods
Linear regression Ridge regression Lasso regression Elastic net regression LARS Lasso least angle regression Huber regression OMP Bayesian ridge regression Regression with decision trees	AdaBoost Regressor Random forest regressor Gradient boosting regressor CatBoost regressor Extra trees regressor K-neighbors regressor Passive aggressive regressor	XGBOOST Light gradient boosting machine LSTM Ar-net Comparative and combines results with all algorithms

LARS: Least angle regression, OMP: Orthogonal matching pursuit, XGBOOST: Extreme gradient boosting, LSTM: Long short term memory, Ar-net: Autoregressive neural networks

$$MAPE = \frac{100}{N}\sum_{h=1}^{N}\frac{|y_h - S_h|}{S_h} \quad (1)$$

Where N is the total number of estimated values.

Figure 5: Flow chart for the data processing and forecasting



4. CONSUMPTION FORECASTING OF ELECTRICITY METERS

The data is provided from an electricity distribution company in Turkey, and the meters belong to different customers in different cities. The collected data from each meter is not uniform with some missing values. The forecasting is performed considering different customers groups and time horizons as such:

- Hourly consumption forecasts of meters
- Daily consumption forecasts that are converted to hourly estimates
- Hourly consumption forecasts based on segment-customer-city groups

The details of each forecasting result are explained and presented below.

4.1. Hourly Consumption Forecasting of Meters

While making hourly forecasts of the meters, 24 h (hold out) before the last hour are selected as test data and the data is used to measure the success of the model. In hourly meter forecasting models, the model is created before midnight of the day and the hours of the next day are scored with the relevant model. For example, if the day is 00:00 on November 29, 2018, the hours after November 28, 2018 00:00 are reserved for testing purposes and the model of the meter is created from the data up to this date. The consumption is forecasted for the meter using the proposed models for all hours of November 28, 2018. The actual consumption data on November 28, 2018 and the estimates are collected separately, the MAPE values are calculated and the most successful model is selected while the success level of each method is scored using the data of the day before the actual day. Hence, hourly LSTM, 16 different regression and AR-Net models were created for each of the meters.

When creating LSTM models, the model was created using only the series itself. Since LSTM is a deep learning algorithm, it is a very successful model for finding patterns since it keeps the effects of previous values in memory. Since the model training takes too long, it can negatively affect the working performance of the model in terms of time, and this time may be longer as the number of meters increases. For the sake of clarity, an analysis for meter S11 is presented.

It is observed that using 16 different regression models for each meter provides useful results. Each algorithm has a different content, and there are regression models that are more capable of

Figure 6: The actual and forecasted results for each methodology

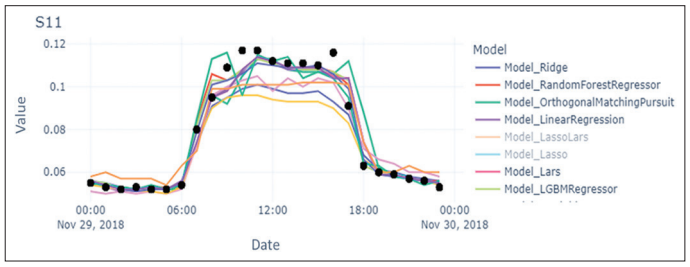


Figure 7: The actual and forecasted consumption

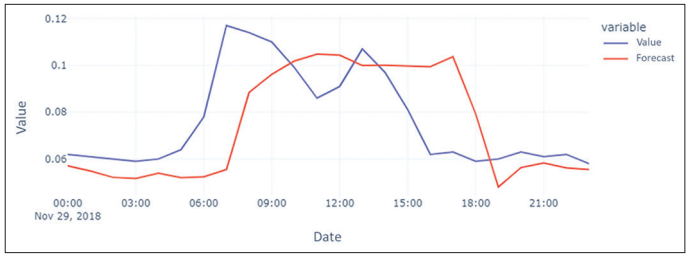
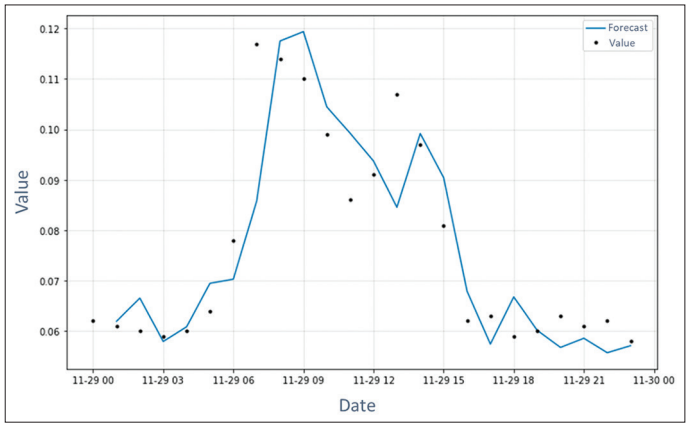


Figure 8: The actual and forecasted consumption



finding nonlinear patterns. The disadvantage is that since there are independent variables, data manipulation and processing are difficult and long, and modelling takes a long time due to the complex structure of some regression models. Figure 6 shows the forecasted results from each methodology, and Figure 7 shows the results obtained from the best performing methodology.

The MAPE values for meter S11 are given in Table 10. Decision Tree Regressor (Turquoise) seems to be the model that best matches the actual consumption (black dots) based on the MAPE, which is closest to the total consumption of November 29. There is a decision tree-based algorithm in the second best model that shows that the energy consumption of this meter can be better explained with independent variables.

Table 10: The MAPE values for each methodology

Model	MAPE (%)
Decision tree regressor	0.1
Random forest regressor	0.1
Orthogonal matching pursuit	0.4
LGBM regressor	0.5
Lars	0.9
Bayesian ridge	0.9
Gradient boosting regressor	1.1
Ada boost regressor	1.4
Linear regression	1.4
K-neighbors regressor	4.2

MAPE: Mean absolute percentage error

The details of the results provide much useful information. The Ar-net results for S11 meter show the trend changes, the impact of delayed temperature on the consumption and the consumption on special days such as holidays. Figure 8 provides the results for the meter S11 but they can be replicated for all meters.

Another important issue is that the positive effect of situations such as holidays and special days on the model can be observed. Figure 9 provides results for S11. The model is quite successful in capturing special day effects.

The MAPE value of the AR-Net model in the meter S11 is very low (0.1%) and it is slightly worse than the Decision Tree Regressor. We conducted control tests for November 2008 and the successful models and their mean MAPE values are given in Figure 10.

It is shown that the MAPE values of the hourly models of the S11 meter in November 2018 and the models used on the relevant day. While the MAPE value remained below 1% throughout November, it had a MAPE value of 1.5% with the orthogonal matching pursuit

Figure 9: The model reaction for the special days

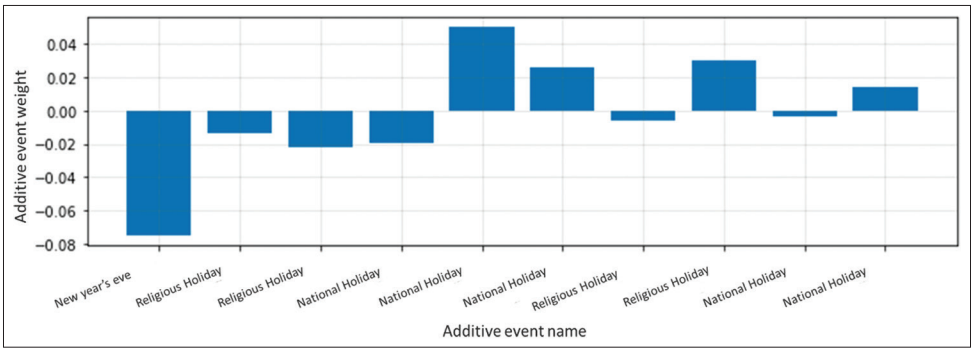
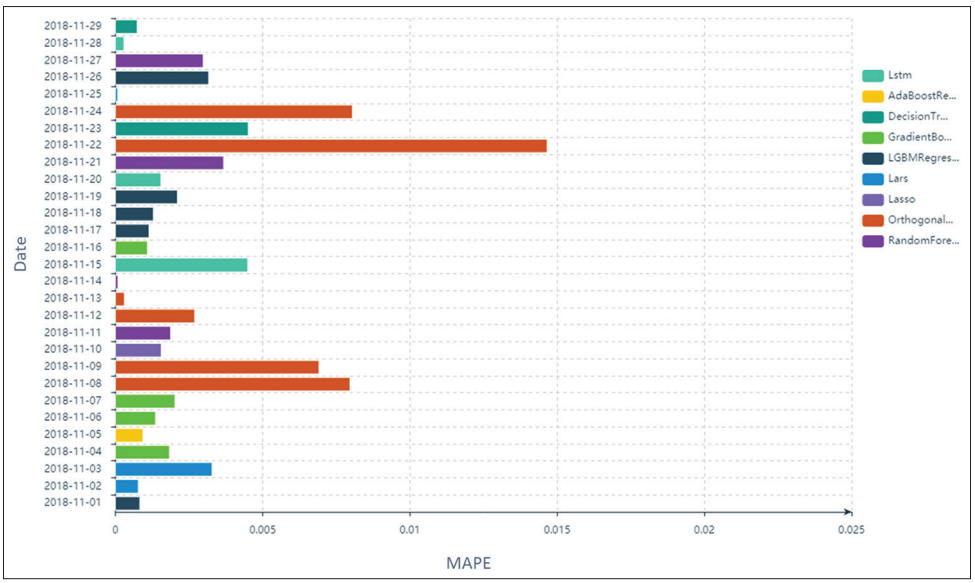


Figure 10: The MAPE levels for each selected methodology for November, 2018



model on November 22, 2018. 9 different models were selected successfully for November, which shows the effectiveness of the automated ranking mechanism.

The most successful methodologies based on MAPE values for 248 m are given in Table 11. The results show that the MAPE values of 241 out of 248 m are below 4%. While the trend and partial trends can be taken into account, it can be used in the model by adding possible trend change situations. The number of training iterations of the model can be affected, and the model can make more successful predictions. The disadvantage is that it takes much longer than all other models, and it takes about 35 s to create a 1-day model per meter.

4.2. Daily Consumption Forecasting to Convert to Hourly Estimates

The same methodologies were used when estimating the meters daily and then converting them to hourly consumption. The consumption value of each meter was collected daily while the daily models were set up, and then the hourly consumption values were estimated from the aggregate forecasts. 3-month data was used for the LSTM and 6 months of data were used for the regression models for training purposes. The last completed day is used to test the model results and compare them with the actual hourly consumption.

Figure 11: The actual and forecasted consumption for random Forest regressor

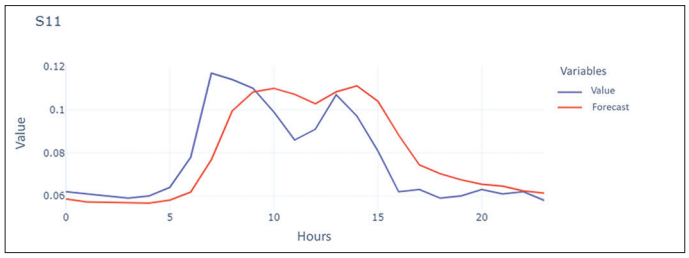


Table 11: The frequencies of high-performing models for forecasting

Models	Meters
Decision tree regressor	48
LSTM	45
Orthogonal matching pursuit	31
LGBM regressor	23
Gradient boosting regressor	18
Random forest regressor	15
Huber regressor	15
Ada boost regressor	10
Bayesian ridge	10
Lars	8
Linear regression	7
Ridge	7
K-Neighbors regressor	5
Passive aggressive regressor	2
ElasticNet	1
Lasso	1
Lasso lars	1
Dummy regressor	1

LSTM: Long short term memory

The methodologies are sorted based on the MAPE values for November 29, 2018, and they are given in Table 12 below. Although Decision Tree Regressor and LSTM take the first two places in the model ranking, as in the hourly models, it is seen that the numbers and places of the other models have changed. Although the temperatures are mostly taken as the average of

Table 12: The frequencies of high-performing models for forecasting

Models	Meters
Decision Tree Regressor	47
LSTM	28
Random Forest Regressor	27
Huber Regressor	20
Ada Boost Regressor	19
Orthogonal Matching Pursuit	16
LGBM Regressor	16
Gradient Boosting Regressor	14
K-Neighbors Regressor	13
Passive Aggressive Regressor	12
ElasticNet	7
Linear Regression	7
Bayesian Ridge	7
Ridge	7
Lasso	3
Dummy Regressor	2
Lasso Lars	2
Lars	1

LSTM: Long short term memory

Table 13: The MAPE values and winning scenario for each day of November, 2018

November 2018	Hourly model MAPE (%)	Daily model MAPE (%)	Winner scenario
1	8.00	38.20	Hourly
2	2.00	3.70	Hourly
3	2.10	8.90	Hourly
4	1.50	5.80	Hourly
5	2.00	8.20	Hourly
6	0.80	3.60	Hourly
7	0.60	3.50	Hourly
8	0.60	2.80	Hourly
9	1.10	5.60	Hourly
10	3.00	11.70	Hourly
11	3.10	21.30	Hourly
12	0.50	2.40	Hourly
13	0.80	4.20	Hourly
14	3.10	12.70	Hourly
15	1.00	5.00	Hourly
16	0.80	3.30	Hourly
17	1.00	5.50	Hourly
18	0.80	2.90	Hourly
19	0.80	2.40	Hourly
20	0.80	3.80	Hourly
21	0.70	5.00	Hourly
22	1.40	5.90	Hourly
23	0.50	2.10	Hourly
24	0.70	2.40	Hourly
25	0.70	4.60	Hourly
26	0.60	3.60	Hourly
27	1.60	7.30	Hourly
28	1.30	8.80	Hourly
29	1.20	4.30	Hourly
Mean	1.50	6.90	Hourly

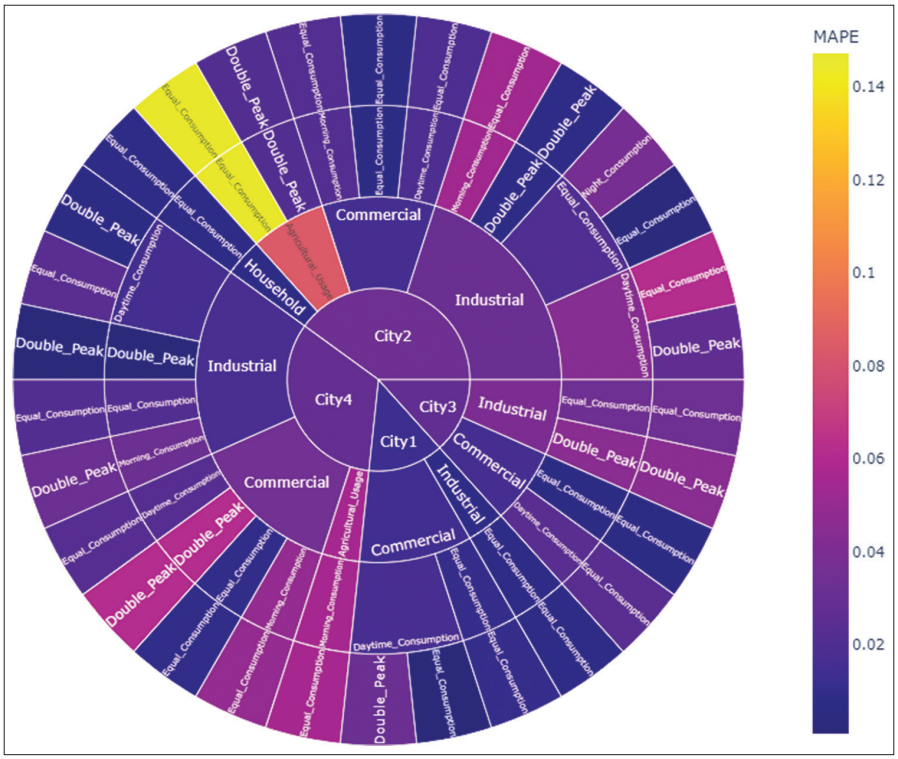
MAPE: Mean absolute percentage error

Table 14: The segments, models, and hourly MAPE values for November 29, 2018

City	Subscriber group	Weekday consumption	Weekend consumption	Model	Value	Forecast	MAPE
City1	Commercial	Equal	Equal	Huber regressor	38552	38431	0.31%
City1	Commercial	Morning	Equal	Orthogonal Matching Pursuit	8814	8813	0.01%
City1	Commercial	Morning	Double Peak	Random Forest Regressor	2866	2965	3.45%
City2	Household	Equal	Equal	Huber Regressor	2823	2841	0.64%
City2	Industrial	Equal	Equal	Huber Regressor	116269	113081	2.74%
City2	Industrial	Equal	Night	Huber Regressor	74588	74559	0.04%
City2	Industrial	Morning	Equal	Random Forest Regressor	2746	2699	1.71%
City2	Industrial	Morning	Double Peak	Random Forest Regressor	2024	2012	0.59%
City2	Industrial	Equal Peak	Double Peak	Random Forest Regressor	16508	8855	46.36%
City2	Industrial	Morning	Equal	Orthogonal Matching Pursuit	1483	1432	3.44%
City2	Tarimsal	Equal	Equal	Random Forest Regressor	0.15	0.142	5.33%
City2	Tarimsal	Equal Peak	Double Peak	Random Forest Regressor	1422	1381	2.88%
City2	Commercial	Equal	Equal	HuberRegressor	499539	500861	0.26%
City2	Commercial	Morning	Equal	Ridge	92062	103925	12.89%
City2	Commercial	Morning	Double Peak	Orthogonal Matching Pursuit	1657	1532	7.54%
City2	Commercial	Equal Peak	Double Peak	Random Forest Regressor	4745	2879	39.33%
City2	Commercial	Morning	Equal	LGBM Regressor	3904	3541	9.30%
City3	Industrial	Equal	Equal	Random Forest Regressor	182728	191949	5.05%
City3	Industrial	Morning	Equal	Orthogonal Matching Pursuit	1278	1292	1.10%
City3	Industrial	Morning	Double Peak	Random Forest Regressor	2616	2634	0.69%
City3	Commercial	Equal	Equal	Orthogonal Matching Pursuit	78311	77531	1.00%
City3	Commercial	Morning	Equal	Gradient Boosting Regressor	7981	7734	3.09%
City4	Industrial	Equal	Equal	Huber Regressor	247511	241842	2.29%
City4	Industrial	Equal	Night	Huber Regressor	72.24	72097	0.20%
City4	Industrial	Morning	Equal	Gradient Boosting Regressor	40426	39423	2.48%
City4	Industrial	Equal Peak	Double Peak	Random Forest Regressor	19073	10127	46.90%
City4	Industrial	Morning	Equal	Random Forest Regressor	1835	1708	6.92%
City4	Industrial	Morning	Double Peak	Orthogonal Matching Pursuit	1396	1416	1.43%
City4	Tarimsal	Morning	Equal	Orthogonal Matching Pursuit	0.194	0.205	5.67%
City4	Commercial	Equal	Equal	Gradient Boosting Regressor	245319	241577	1.53%
City4	Commercial	Morning	Equal	Random Forest Regressor	22239	22428	0.85%
City4	Commercial	Equal Peak	Double Peak	Gradient Boosting Regressor	5331	5941	11.44%
City4	Commercial	Morning	Equal	Huber Regressor	1514	1322	12.68%

MAPE: Mean absolute percentage error

Figure 12: The distribution of mean absolute percentage error values for best performing models, November, 29th



the day in daily forecasts, variables such as the day of the week become more meaningful.

After the most successful models were selected, daily forecasts were distributed to hours by taking into account the hourly ratios of the meters that had previously been calculated according to the days of the week for the last 28 days. The percentages are calculated by normalizing the consumption of each hour based on the total daily consumption. The Random Forest Regressor results of S11 returned to be the best performing for the meter, and the results are given the Figure 11 below.

The conversion process of daily forecasts to hourly consumption was successfully applied for 176 of 248 m, while the number of meters with a MAPE below 4% is 121. It is shown that the hourly forecasts are much more successful results. The MAPE details for each day are given in Table 13 below. It is shown that forecasting the daily consumption and then converting to hourly forecasts using the calculated ratios return quite successful results.

4.3. Hourly Consumption Forecasts Based on Segment-customer-City Groups

Another scenario applied to estimate the consumption of meters is to make an hourly forecast on the basis of the consumption characteristic of the customer group and the city in which the meter is located. Energy companies may need estimations on the basis of regional or customer groups in line with their workflow. A representation of the distribution of meters on the basis of segment, city, and subscriber group is given in Table 14 below. The table also shows the results of some forecasting methodologies and the best methodology of each meter will be selected based on MAPE.

The models to be created are LSTM, 16 different regression and AR- Net models and they will be applied to 34 different groups. The air temperature variable in the data is city-based and it was not significant for some groups. 3 months long data has been used to create and train the models. The best-performing methodologies results are obtained after running the forecasting models. Figure 12 shows the distribution of MAPE values for best-performing methods classified based on different groups.

Figure 13: The frequencies of best performing methodologies

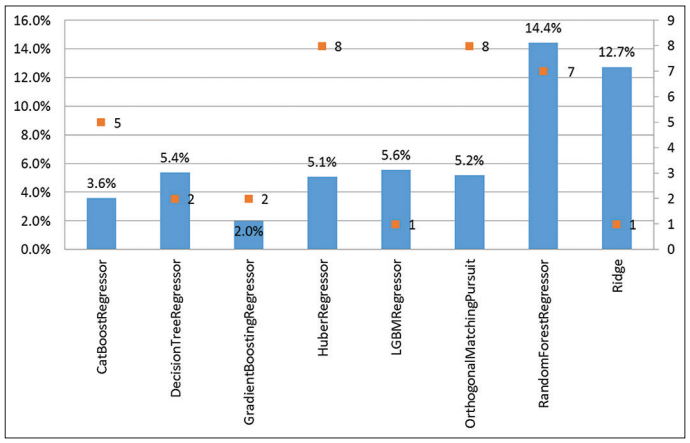


Table 15: The segments, models, and daily MAPE values for November 29, 2018

City	Subscriber group	Weekday consumption	Weekend consumption	Value	Forecast	MAPE (%)
City 1	Industrial	Equal	Equal	2.465	2.453	0.50
City 1	Commercial	Equal	Equal	52.627	51.636	1.90
City 1	Commercial	Daytime	Equal	6.882	6.744	2.00
City 1	Commercial	Daytime	Double peak	3.114	3.057	1.80
City 2	Household	Equal	Equal	68.801	68.56	0.30
City 2	Industrial	Equal	Equal	112.444	113.793	1.20
City 2	Industrial	Equal	Night	146.828	147.662	0.60
City 2	Industrial	Daytime	Equal	1.231	1.317	7.00
City 2	Industrial	Daytime	Double peak	1.131	1.187	4.90
City 2	Industrial	Double peak	Double peak	21.795	20.061	8.00
City 2	Industrial	Morning	Equal	1.835	1.901	3.60
City 2	Agriculture	Equal	Equal	0.15	0.161	7.50
City 2	Agriculture	Double peak	Double peak	1.422	1.436	1.00
City 2	Commercial	Equal	Equal	272.784	272.023	0.30
City 2	Commercial	Daytime	Equal	53.958	53.212	1.40
City 2	Commercial	Morning	Equal	5.593	5.369	4.00
City 3	Industrial	Equal	Equal	157.34	158.656	0.80
City 3	Industrial	Double peak	Double peak	1.581	1.542	2.50
City 3	Commercial	Equal	Equal	61.385	61.922	0.90
City 3	Commercial	Daytime	Equal	5.784	5.804	0.40
City 4	Industrial	Equal	Equal	495.034	478.545	3.30
City 4	Industrial	Daytime	Equal	54.758	52.719	3.70
City 4	Industrial	Daytime	Double peak	4.589	4.618	0.60
City 4	Industrial	Double peak	Double peak	18.622	18.325	1.60
City 4	Industrial	Morning	Double peak	0.946	0.972	2.80
City 4	Agriculture	Morning	Equal	0.194	0.248	28.00
City 4	Commercial	Equal	Equal	167.535	169.612	1.20
City 4	Commercial	Daytime	Equal	50.202	48.962	2.50
City 4	Commercial	Double peak	Double peak	2.299	2.543	10.60
City 4	Commercial	Morning	Equal	1.514	1.448	4.40

The groups with higher MAPE values are represented in yellow color. It is shown that the MAPE values are <0.04 for the majority of the segments and cities. The higher MAPE values usually come from agricultural meters, which possess uncertain behaviors and are challenging to forecast. It is shown that the proposed forecasting methodologies are successful in estimating consumption. The details of the groups and segments for each city are given in Table 15.

For the regression models, hourly data from the last 3 months were used. 16 regression models were created for each group, and the best one was selected according to the MAPE criterion. While the MAPE values of 15 of 34 groups is >4%, the MAPE value of 7 groups is <1%. Orthogonal Matching Pursuit and Huber Regressor are the most frequently used models with 8 different groups. Gradient Boosting and Cat Boost Regressor models were the models with an average MAPE below 4%. Although the Random Forest Regressor model is the best model in 7 different groups, the MAPE average seems to be an unsuccessful model with 14.4%, as can be seen in the previous table. The MAPE values for the City2-Industry-Daytime consumption-double peaks and City4 Commercial-Morning consumption-Equal consumption groups have very high MAPE values that lead to higher average MAPE. City 2-Industry-Daytime consumption consists of 2 m (121 and 232) and has a least success for estimation. The S121 meter was estimated with Hourly Huber Regression model with a MAPE of 1% and with the Hourly Orthogonal Matching Pursuit on S232 with MAPE of 1%. Although segment-based MAPE was 14%, it was estimated at 1% per h. Figure 13 below shows the model selection frequencies for best performing methodologies and their average MAPE values. It is shown that the success of the models increases in groups with a higher number of meters which shows the importance of data availability in model success.

5. CONCLUSION

Electricity demand forecasting plays a key role for power companies as they need to develop long- and short-term strategies. On the other hand, the distribution companies need to forecast the consumption for meters. The meters are distributed to different regions and belong to different customer groups such as commercial, household, industrial, and agricultural. The characteristics of the consumers are different, and the data gathered from the meters need extra processing. Household electricity consumption is dominated by illumination, heating, and cooling needs; hence it has strong periodic components whose amplitudes depend on climatic conditions. Holidays and special events are irregular but predictable events that affect electricity consumption to a great extent. In particular, in the Islamic world, religious holidays are determined according to the lunar calendar and they start 10 days earlier each year. These types of problems require special methods for dealing with special days and events.

In this work, meter based consumption data which is provided from a regional distribution company in Turkey was analyzed. Then an end-to-end methodology that includes data processing, missing value detection and filling, abnormal value detection, customer clustering and segmentation, and eventually forecasting is proposed and successfully applied. The meter consumption data of the distribution company need to be processed as not all the meters have the same data structure. The abnormal value detection, missing value detection and filling, and mass reading identification operations are performed for data cleaning. The test data include 267 m from 4 different cities and belong to the commercial, industrial, household, and agricultural customers. Standard deviation, interquartile range, Isolation Forest, Minimum Covariant Determinant, and DBSCAN methods were used to detect the abnormal values. On the other hand, Linear Interpolation, Quadratic Interpolation, Cubic Interpolation, Moving average, and Moving median are used to fill the missing values and Linear interpolation is selected as the best methodology.

The customers are clustered using hierarchical, K-means, and mean shift clustering methodologies and the hierarchical clustering methodology is selected as the most suitable alternative. Then the customers are segmented based on their city, consumption patterns, weekday and weekend consumptions. Such an approach increased the forecast accuracy.

The forecasting is planned as hourly, daily, and segment based and 10 different regression methods, 7 regressors, 5 machine learning methods that include LSTM and Ar-net models are used and the best performing methodologies are selected for each customer group in each segment. The meter based forecasting for hourly, daily, and segment-customer groups are presented along with the MAPE values. It is shown that the results are quite satisfactory with MAPE values <4% for the majority of the groups and the best methodologies return around 1% of MAPE. Although the main idea is to develop an automated and data-centered end-to-end approach for distribution companies and the same forecasting methodology might not be the best one for each case, the decision tree regressor and LSTM are determined as the methods with lowest MAPE values. The proposed method is able to forecast the hourly and daily consumption quite satisfactorily, and the results presented for test data for hourly, daily and segment based cases.

The proposed model is applicable to the electricity demand data for any meter, provided that sufficient data is provided and customer segmentation is performed. The methodologies are more extensive than that of previous researches. As a matter of fact, the real data gathered from the meters make the research more realistic as it is difficult to obtain data from the meters if it is not smart grid. The anomaly detection and missing value filling methodologies are novel and not being used in previous researches. Special days such as holidays are also identified and treated separately. Nevertheless, better performance is expected to occur in cases where the customer consumption is more uniform such as household as the agricultural consumption is more challenging to forecast due to nonunform consumption patterns.

6. ACKNOWLEDGMENT

The authors acknowledge the financial assistance from the Tubitak for the support to energy and power market research.

7. FUNDING

This research received partial funding from Tubitak, Turkey.

REFERENCES

Abumohsen, M., Owda, A.Y., Owda, M. (2023), Electrical load forecasting using LSTM, GRU, and RNN algorithms. Energies, 16(5), 2283.

Altman, N.S. (1992), An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician, 46(3), 175-185.

Anand, S.A., Suganthi, L. (2012), Energy models for demand forecasting-a review. Renewable and Sustainable Energy Reviews, 16(2), 1223-1240.

Andersen, F.M., Larsen, H.V., Gaardestrup, R.B. (2013), Long term forecasting of hourly electricity consumption in local areas in Denmark. Applied Energy, 110, 147-162.

Apadula, F., Bassini, A., Elli, A., Scapin, S. (2012), Relationships between meteorological variables and monthly electricity demand. Applied Energy, 98, 346-356.

Arora, S., Taylor, J.W. (2016), Forecasting electricity smart meter data using conditional kernel density estimation. Omega, 59, 47-59.

Azadeh, A., Ghaderi, S.F., Tarverdian, S., Saberi, M. (2007), Integration of artificial neural networks and genetic algorithm to predict electrical energy consumption. Applied Mathematics and Computation, 186(2), 1731-1741.

Bašta, M., Helman, K. (2013), Scale-specific importance of weather variables for explanation of variations of electricity consumption: The case of Prague, Czech republic. Energy Economics, 40, 503-514.

Blondel, M., Kubo, Y., Naonori, U. (2014), Online Passive-aggressive Algorithms for Non-negative Matrix Factorization and Completion. In: Artificial Intelligence and Statistics. New York: PMLR. p96-104.

Chakhchoukh, Y., Panciatici, P., Mili, L. (2010), Electric load forecasting based on statistical robust methods. IEEE Transactions on Power Systems, 26(3), 982-991.

Clements, A.E., Hurn, A.S., Li, Z., (2016), Forecasting day-ahead electricity load using a multiple equation time series approach. European Journal of Operational Research, 251(2), 522-530.

Conejo, A.J., Contrerasa, J., Espínolaa, R., Plazasb, M.A. (2005), Forecasting electricity prices for a day-ahead pool-based electric energy market. International Journal of Forecasting, 21, 435-462.

Crowley, C., Joutz, F.L. (2003), Hourly Electricity Loads: Temperature Elasticities and Climate Change. In: The 23rd US Association of Energy Economics North American Conference Mexico City.

De Felice, M., Alessandri, A., Catalano, F. (2015), Seasonal climate forecasts for medium-term electricity demand forecasting. Applied Energy, 137, 435-444.

De Felice, M., Alessandri, A., Ruti, P.M., (2013), Electricity demand forecasting over Italy: Potential benefits using numerical weather prediction models. Electric Power Systems Research, 104, 71-79.

Dewangan, F., Abdelaziz, A.Y., Biswal, M. (2023), Load forecasting models in smart grid using smart meter information: A review. Energies, 16(3), 1404.

Draper, N.R, Smith, H. (1998), Applied Regression Analysis. 3rd ed. Hoboken, New Jersey: John Wiley.

Dyner, I., Larsen, E. (2001), From planning to strategy in the electricity industry. Energy Policy, 29, 1145-1154.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. (2004), Least angle regression. Annals of Statistics, 32(2), 407-499.

Fan, S., Hyndman, R.J., (2012), Short-term load forecasting based on a semi-parametric additive model. IEEE Transactions on Power Systems, 27(1), 134-141.

Filik, Ü.B., Gerek, Ö.N., Kurban, M. (2011), A novel modeling approach for hourly forecasting of long-term electric energy demand. Energy Conversion and Management, 52(1), 199-211.

Gajowniczek, K., Ząbkowski, T. (2016), Short term electricity forecasting based on user behavior from individual smart meter data. Journal of Intelligent and Fuzzy Systems, 30(1), 223-234.

Gajowniczek, K., Ząbkowski, T. (2017), Electricity forecasting on the individual household level enhanced based on activity patterns. PLoS One, 12(4), e0174098.

Ghofrani, M., Hassanzadeh, M., Etezadi-Amoli, M., Fadali, M.S. (2011), Smart Meter Based Short-term Load Forecasting for Residential Customers. In: 2011 North American Power Symposium. New York City: IEEE. p1-5.

Grinstead, C.M., Snell, J.L. (2006), Introduction to Probability. 2nd ed. Providence, RI: American Mathematical Society.

Hahn, H., Meyer-Nieberg, S., Pickl, S. (2009), Electric load forecasting methods: Tools for decision making. European Journal of Operational Research, 199, 902-907.

Hastie, T., Tibshirani, R., Friedman, J.H. (2009), 10. Boosting and additive trees. In: The Elements of Statistical Learning. 2nd ed. New York: Springer. p337-384.

Hochreiter, S., Schmidhuber, J. (1997), Long short-term memory. Neural Computation, 9(8), 1735-1780.

Hoerl, A.E., Robert, W.K. (1970), Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12, 55-67.

Hor, C.L., Watson, S.J., Majithia, S. (2005), Analyzing the impact of weather variables on monthly electricity demand. IEEE Transactions on Power Systems, 20(4), 2078-2085.

Hsiao, Y.H. (2014), Household electricity demand forecast based on context information and user daily schedule analysis from meter data. IEEE Transactions on Industrial Informatics, 11(1), 33-43.

Huber, P.J. (1964), Robust estimation of a location parameter. Annals of Statistics, 53(1), 73-101.

Islam, S.M., Al-Alawi, S.M., Ellithy, K.A. (1995), Forecasting monthly electric load and energy for a fast growing utility using an artificial neural network. Electric Power Systems Research, 34(1), 1-9.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013), An Introduction to Statistical Learning. New York: Springer.

Kök, A., Yükseltan, E., Hekimoglu, M., Aktunc, E.A., Yücekaya, A., Bilge, A. (2022), Forecasting hourly electricity demand under COVID-19 restrictions. International Journal of Energy Economics and Policy, 12(1), 73-85.

Liu, H., Motoda, H. (1998), Feature Selection for Knowledge Discovery and Data Mining. New York: Springer.

Lo, K.L., Wu, Y.K. (2003), Risk assessment due to local demand forecast uncertainty in the competitive supply industry. IEE Proceedings-Generation, Transmission and Distribution, 150(5), 573-581.

Lusis, P., Khalilpour, K.R., Andrew, L., Liebman, A. (2017), Short-term residential load forecasting: Impact of calendar effects and forecast granularity. Applied Energy, 205, 654-669.

McSharry, P.E., Bouwman, S., Bloemhof, G. (2005), Probabilistic forecasts of the magnitude and timing of peak electricity demand. IEEE Transactions on Power Systems, 20(2), 1166-1172.

Momani, M.A. (2013), Factors affecting electricity demand in Jordan. Energy Power Engineering, 5, 50-58.

Niu, J., Xu, Z.H., Zhao, J., Shao, Z.J., Qian, J.X. (2010), Model predictive control with an on-line identification model of a supply chain unit. Journal of Zhejiang University Science C, 11(5), 394-400.

Pai, P.F., Hong, W.C. (2005), Forecasting regional electricity load based on recurrent support vector machines with genetic algorithms. Electric Power Systems Research, 74(3), 417-425.

Pati, Y., Rezaifar, R., Krishnaprasad, P. (1993), Orthogonal Matching Pursuit: Recursive Function Approximation with Application to Wavelet Decomposition. In: Asilomar Conference on Signals,

Systems and Computers. p40-44.
Quinlan, J.R. (1987), Simplifying decision trees. International Journal of Man-Machine Studies, 27(3), 221-234.
Ren, Y., Suganthan, P.N., Srikanth, N., Amaratunga, G. (2016), Random vector functional link network for short-term electricity load demand forecasting. Information Sciences, 367, 1078-1093.
Taieb, S.B., Huser, R., Hyndman, R.J., Genton, M.G. (2016), Forecasting uncertainty in electricity smart meter data by boosting additive quantile regression. IEEE Transactions on Smart Grid, 7(5), 2448-2455.
Tarmanini, C., Sarma, N., Gezegin, C., Ozgonenel, O. (2023), Short term load forecasting based on ARIMA and ANN approaches. Energy Reports, 9, 550-557.
Taylor, J.W. (2003), Short-term electricity demand forecasting using double seasonal exponential smoothing. Journal of the Operational Research Society, 54(8), 799-805.
Taylor, J.W. (2010), Triple seasonal methods for short-term electricity demand forecasting. European Journal of Operational Research, 204, 139-152.
Taylor, J.W., Buizza, R. (2003), Using weather ensemble predictions in electricity demand forecasting. International Journal of Forecasting, 19(1), 57-70.
Tedas. (2022), Available from: https://www.tedas.gov.tr/#!dagitim_srkt
Tibshirani, R. (1996), Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B (Methodological), 58(1), 267-288.
Triebe, O., Laptev, N., Rajagopal R. (2019), AR-Net: A Simple Auto-Regressive Neural Network for Time-Series.
Vilar, J.M., Cao, R., Aneiros, G. (2012), Forecasting next-day electricity demand and price using nonparametric functional methods. International Journal of Electrical Power and Energy Systems, 39(1), 48-55.
Wang, A.J., Ramsay, B. (1998), A neural network based estimator for electricity spot-pricing with particular reference to weekend and public holidays. Neurocomputing, 23(1-3), 47-57.
Wang, J., Li, L., Niu, D., Tan, Z. (2012), An annual load forecasting model based on support vector regression with differential evolution algorithm. Applied Energy, 94, 65-70.
Wang, X., Yao, Z., Papaefthymiou, M. (2023), A real-time electrical load forecasting and unsupervised anomaly detection framework. Applied Energy, 330, 120279.
Wijaya, T.K., Vasirani, M., Humeau, S., Aberer, K. (2015), Cluster-based Aggregate Forecasting for Residential Electricity Demand Using Smart Meter Data. In: 2015 IEEE International Conference on Big data (Big data). New York City: IEEE. p879-887.
Wyner, A.J., Olson, M., Bleich, J., Mease D. (2017), Explaining the success of AdaBoost and random forests as interpolating classifiers. Journal of Machine Learning Research, 18(48), 1-33.
Zhanga, B.L., Dongb, Z.Y. (2001), An adaptive neural-wavelet model for short term load forecasting. Enery Power Systems Research, 59(2), 121-129.
Zhu, S., Wang, J., Zhao, W., Wang, J. (2011), A seasonal hybrid procedure for electricity demand forecasting in China. Applied Energy, 88(11), 3807-3815.
Zou, H., Hastie, T. (2005), Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B, 67(2), 301-320.