Turkish Journal of Electrical Engineering and Computer Sciences

Volume 25 | Number 1 Article 3

1-1-2017

Naive forecasting of household natural gas consumption with sliding window approach

MUSTAFA AKPINAR

NEJAT YUMUŞAK

Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2017) 25: 30 – 45

doi:10.3906/elk-1404-378

Naïve forecasting of household natural gas consumption with sliding window approach

Mustafa AKPINAR*, Nejat YUMUŞAK
Department of Computer Engineering, Faculty of Computer and Information Science, Sakarya University, Sakarya, Turkey

Received: 15.04.2014 • **Accepted/Published Online:** 13.11.2015 • **Final Version:** 24.01.2017

Abstract: Household consumption has a significant importance for natural gas wholesale companies. These companies make one-day-ahead forecasting daily. However, there are penalties depending on the error of the estimates. These penalties increase exponentially depending on the error rate. Several studies have been done to develop mathematical models to forecast natural gas consumption and minimize the error rate. However, before mathematical model predictions, a previous step, data preparation, is also important. The data must be prepared correctly before the mathematical model. At this point, prior to the mathematical model, selecting the appropriate data set size has a vital role. In this study, one-day-ahead household natural gas consumption is forecasted for different data sizes. Forecasts have been made for the year 2012. For removing insignificant variables, multiple linear regression (MLR) is applied to all data. In this research, 2 particular scenarios are applied for forecasting. In the first scenario, 2 different data set models are prepared. These sets consist of the data collected 6 weeks before the forecasted day. Daily outcomes are added to the data set and the set is applied in a model called Model A. The other model is depicted based on a sliding window idea having 6 weeks of fixed data size with dynamic data inside (Model W6). For the two models, MLR is applied and error rates are compared. Here, Model A has 7 times higher mean absolute percent error (MAPE) than Model W6. In scenario 2, 6 models are studied and compared for the sliding window approach. The models are named according to the weeks involved (e.g., Model W1, Model W6). MAPEs for Model W3, Model W4, Model W5, and Model W6 are obtained as 11.8%, 6.8%, 7.2%, and 8.1%, respectively. The lowest preday error occurs in the 4-week data model with sliding window approach.

Key words: Decision making, decision support system, demand forecasting, energy management, load management, natural gas, predictive models, regression analysis

1. Introduction

Natural gas wholesale companies make daily gas consumption forecasts for their customers. Those customers are mostly high-consuming industries or gas distribution companies, which largely have household consumers. Unlike industrial steady consumption, household consumptions are affected by climatic alterations. As is known, heating-oriented consumptions increase in cold weather. Relatively, more climatic variations make forecasting difficult and increase the error rate. Subject to the error rate, fines occur. Here, as the rate rises, penalties also increase exponentially. To avoid penalties, several mathematical models are developed to predict natural gas consumption and to decrease the error rate. However, during the prediction period, there is another factor having as much importance as the introduction of the mathematical model. One step before the prediction,

*Correspondence: akpinar@sakarya.edu.tr

preparing data accurately is essential. Here, in order to make estimation from the data, optimal data set size determination is a critical point.

Several natural gas prediction studies have been reported in the literature so far. These may differ depending on the used methodology, mathematical model, and forecasting period. Sarak and Satman used the heating degree-day method to determine the natural gas consumption by residential heating in Turkey. They found in total 14.92 Gm^3 consumption if all the residences used natural gas for heating [1]. Brown et al. used a feedforward neural network for predicting daily natural gas consumption. They reduced the residual root mean square errors by more than half for the linear regression-based model [2]. Gil and Deferrari predicted gas consumption in short and intermediate ranges. In the short range case, they used temperature with the heating degree-day method and found an equation that used annual consumption and variation in time. They also obtained a prediction of 10% uncertainty in consumption for 1 to 5 days [3]. Akpinar and Yumusak predicted daily household natural gas consumptions while removing the cycle effect. They used multiple linear regression (MLR) with 14.38% mean absolute percent error (MAPE) and autoregressive integrated moving average (ARIMA) with 8.48% MAPE [4,5]. Potocnik et al. forecasted the natural gas consumption and found the risky days of consumptions. Other energy markets could also use that model with minor modifications [6]. Brabec et al. described and predicted natural gas consumptions with a nonlinear mixed effects model and compared it with the autoregressive integrated moving average with exogenous variables (ARIMAX) and autoregressive with exogenous variables (ARX) approaches. Sixty-two customers were used for measuring the prediction performance. They found that the standard deviation of their model was 41.6 m^3, which was 47.6 m^3 in the ARIMAX approach [7]. Aydinalp-Koksal and Ugursal studied the conditional demand analysis (CDA) method to model the residential consumption at national level. CDA and characterization of residential consumptions were compared with the neural network and engineering-based model. Socioeconomic factors were included where possible [8]. Sabo et al. forecasted hourly temperature-related natural gas consumption. In Osijek, Croatia, consumptions were tested with given models as well. Implicit and explicit temperature dependence of natural gas consumption was analyzed. The least square method was used to find the optimal parameters of model function. They used past-day data in their model [9]. In 2013, Catalina et al. used the global heat loss coefficient, the south equivalent surface, and the difference between the indoor set point temperature and the sol-air temperature to find a building's heat consumption with multiple dynamic simulations. They used a MLR model to predict the building's heat consumption and the model results were verified with the obtained data from 17 blocks of flats [10].

Prior to the prediction step, determination of data size is also essential. In the literature various studies have been researched on estimating optimal data length. One of the popular techniques that gives adequate and usable results is the sliding window streaming model [11–14]. Gembris et al. studied reference vector analysis in functional magnetic resonance imaging (fMRI). They presented an algorithm to compute correlation coefficients between fMRI time-series and reference time-series using a sliding window [11]. Lee et al. proposed a new sliding window filtering algorithm for incremental mining of association rules in database transaction and evaluate its performance [12]. Luiz et al., in order to measure delay and sampling errors, explored a multisized sliding window workload prediction method for dynamic power management [13]. Suzuki et al., for developing prediction accuracy, introduced a sliding window technique for vector regression. They reduced calculation time [14].

As seen in the literature review, in order to increase the prediction accuracy, different algorithms and methodologies could be used. The aim of this paper is essentially reducing forecasting errors of household

natural gas consumption by using the advantages of the sliding window approach. To achieve this goal, a sliding window technique is used with MLRs to select the most suitable data set sizes. Those two methods have not been applied together in residential demand forecasting before even if the sliding window enables strong estimation with processing the best data set size, as stated in [14]. To accomplish this task, two scenarios are prepared.

2. Method and theory

The relation of two variables is measured with the correlation coefficient (r_{xy}). This coefficient shows the power of relation and its value can be between zero and one, indicating weak and strong relations, respectively. The correlation coefficient [15–18] is given in Eq. (1):

$$r_{XY} = \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) \Bigg/ \sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}, \tag{1}$$

where X and Y are variables, n is the number of data, and $\bar{X}$ and $\bar{Y}$ are averages of X and Y.

If the correlations between variables are close to one, simple linear regression is used to make an equation with variables. The linear regression equation is:

$$Y_i = k_0 + k_1 X_i + e_i, \qquad \dot{Y} = k_0 + k_1 X, \tag{2}$$

where k_0 and k_1 are known as coefficients, X_i is known as an independent variable, Y_i is known as a dependent variable, and e_i is known as error. The predicted value of the dependent variable is shown as $\dot{Y}_i$ and the regression equation does not contain error. Thus, error can be found with e_i. A least square (S) is used for minimizing the sum of e_i to calculate a more accurate prediction.

$$e_i = Y_i - \dot{Y}_i \qquad S = \sum_{i=1}^{n} e_i^2 \tag{3}$$

With the coefficients that make a minimum S, the linear regression equation is found.

2.1. Multiple linear regression

Different from linear regression, MLR as shown in Eq. (4) is used for two or more independent variables [15–17].

$$\dot{Y}_i = k_0 + k_1 X_{i1} + k_2 X_{i2} + \ \ldots \ + k_{p-1} X_{i,p-1} \tag{4}$$

Here, $\dot{Y}_i$ is the estimation of Y_i and $X_{i1} \ \ldots X_{i,p-1}$ are explanatory variables, while $k_0 \ \ldots k_{p-1}$ are known as coefficients. If the number of variables is more than two, a multiple correlation coefficient (MCC) is used [8,9], as in Eq. (5).

$$r_{Y\acute{Y}} = \sqrt{\sum_{i=1}^{n} (\acute{Y}_i - \bar{Y})^2 \Bigg/ \sum_{i=1}^{n} (Y_i - \bar{Y})^2} \tag{5}$$

Here, $\bar{Y}$ is the mean of sum Y_i and $r_{Y\dot{Y}}$ is the MCC.

2.2. Signification of variables

Significant difference from zero of regression coefficients can be determined with individual test parameters. In the MLR equation, $k_0 \ \ldots \ k_{p-1}$ coefficients are the estimated parameters of each independent variable. A significant value is called a t-value and calculated with an estimated parameter divided by standard error. The t-value changes with different degrees of freedom. Degrees of freedom are calculated finding the difference between the number of data steps and the number of independent variables. The t-values' probabilities can be found in the t-table. In this study critical probability is taken as 10%.

2.3. Measuring errors

The MAPE is used for measuring forecast results. Comparing results is over the same step and forecasts are measured as an absolute percentage as in Eq. (6).

$$MAPE = 1/n \sum_{i=1}^{n} \left| \left(\acute{Y}_i - Y \right) \Big/ Y_i \cdot 100\% \right| \tag{6}$$

The side of errors is important in some cases. In such cases, mean percent error (MPE) should be used, as given in Eq. (7). The MPE is the same as MAPE except it does not have absolute value.

$$MPE = 1/n \sum_{i=1}^{n} \left(\acute{Y}_i - Y \right) \Big/ Y_i \cdot 100\% \tag{7}$$

3. Preparation and methodology

3.1. The data

The data processed in this study were provided by the Adapazarı Natural Gas Distribution Company (Turkish acronym: AGDAŞ). The company basically dispenses natural gas for Sakarya Province. The data set holds daily data for 2012 and 43 days from 2011. In total, data from 409 days are studied in the data set consisting of household natural gas consumptions, meteorological data, customer numbers, and holidays. Meteorological data include minimum, maximum, and 1-day lagged average temperature (hereafter referred to as the "lag1 temperature") and minimum and maximum humidity values. Weekends, religious holidays, and national holidays are represented by a dummy variable called "holiday". Two different variables are used for the amount of subscribers. The first variable shows the number of subscribers while the second one is the independent unit amount (IUA). The variable IUA is created for defining building sizes. For 200 m^2 and larger areas, IUA is set to 1. For each additional 100 m^2, 1 is added to the IUA value. For instance, for a 200 m^2 area the IUA would be 1, while for a 370 m^2 area the IUA value would be 3.

3.2. Preliminary analysis

During preliminary analysis, 8 variables explained above are used and MLR is applied to the data set. The relationship between the variables in the data set is examined as correlations. According to the MLR results shown in Table 1, minimum temperature and maximum humidity values are meaningless, so they are removed from the data set. Likewise, in the correlation table (Table 2), there exists a multicollinearity between number of subscribers and IUA. Thus, IUA is removed from the data set. In order to preserve the nature and the

naïveté of the data, only the data mentioned above are removed. The rest of the data are kept unchanged. For instance, extreme values are not eliminated.

Table 1. Probability results of MLR to data set.

Independent variable	t-value	Probability > |t|
Maximum temperature	–9.44	< 0.0001
Minimum temperature	–0.82	0.4146
Minimum humidity	–3.89	0.0001
Maximum humidity	0.36	0.7163
Number of subscribers	–7	< 0.0001
IUA	6.85	< 0.0001
Lag1 mean temperature	–4.83	< 0.0001

Table 2. Correlation results of meaningful independent variables after MLR.

Variable	Holidays	Maximum temperature	Minimum humidity	Number of subscribers	lag 1 mean temperature	Maximum temperature
Holidays	1					
Maximum temperature	–0.0758	1				
Minimum humidity	–0.0034	0.605	1			
Number of subscribers	–0.039	–0.0695	–0.0056	1		
IUA	0.0377	0.0774	0.0048	–0.9971	1	
lag1 mean temperature	0.107	–0.4595	–0.0931	–0.1492	0.1504	1

Eventually, an independent variable data set is composed of minimum humidity, maximum temperature, lag1 mean temperature, holidays, and number of subscribers. The frequencies of the independent variables are shown in Figure 1. As shown, 88% of the minimum humidity is in the range of 25%–74%, while 61% of the maximum temperature is over 20 °C, which is different from the distribution of lag1 mean temperature with 63% for over 14.5 °C. On the other hand, holidays take place 31% of the year. The number of subscribers is a bit different from other distributions, in between 80.500 and 84.999, which is 50% of the variables.

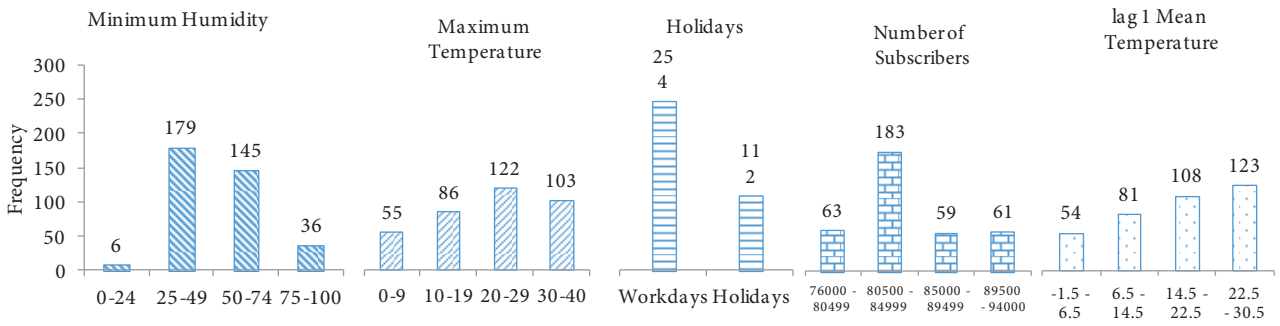

Figure 1. Distribution of independent variables.

3.3. Framework/model outline

In this research, day-ahead forecasting is done. Basically, the day-ahead forecasting method is a prediction approach, which occurs in the next day and the most up-to-date data of it are those of one day before the

forecasting day (Figure 2). Here, "forecasting day" separates the forecasted day from the days before the forecasted day. Thus, data would be placed in 2 categories. Realized consumptions and independent variables are in the first and the forecasted day is in the second. MLR equations are obtained from the values in the first category and consumption predictions are estimated using the independent variables in the second.

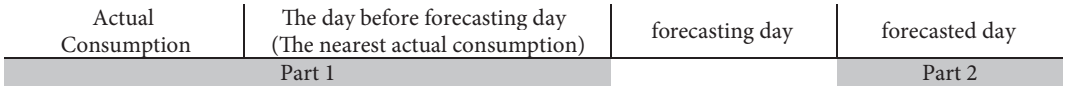

Figure 2. Forecasting step.

As mentioned in the introduction, the sliding window technique has two types based on window size. Those are fixed-sized and multisized windows. Basically, fixed-sized windows with different widths are used in this work. Here the most recent data are appended to the data set while the least recent data in the window are thrown out. The other data remain in the window. Eventually, the attached data form a new data set. Thus, the first part adjusts continuously by time and this change is handled by the sliding window. In Sections 3.4.1 and 3.4.2 implementation of the sliding window are explained more in detail.

3.4. Scenarios

Multiple experiments are formed for day-ahead forecasting. This work mainly includes 2 distinct scenarios. The first scenario has "growing and moving data sets" while the second one has "moving data sets in different sizes". The first forecasted day in the two scenarios is 01.01.2012. The forecasting for this date is done one day before, on 31.12.2011.

3.4.1. Growing and sliding window data sets – scenario 1

In this practice 2 different models, called Model A and Model W6, are applied. Both models use 6-week data until 31.12.2011. Model A has a growing data set, while Model W6 has an unstable moving data set obtained from the sliding window technique. Here "W" refers to the data being gathered weekly and the subsequent number represents the total number of weeks. At the beginning both models have 6-week data (42 days). As the forecasted day gradually progresses, the number of data in Model A increases, while the number of data in Model W6 remains the same. Assuming that 15 January 2012 will be forecasted on 14 January 2012, Model A carries 56-day data. Here, Model W6 stays the same with 42-day data. The first data for Model A start from 19.11.2011, and for Model W6 they start from 03.12.2011. The main goal of the scenario here is examining the error increase or decrease, as the number of days within the data set grows.

3.4.2. Sliding window data sets in different sizes – scenario 2

In order to form a moving data set, this practice uses a sliding window method with 6 distinct models. The date data are processed in the models are considered in a weekly manner. The models are named in "Model Wx" form. Here, "x" shows the actual number of the week for past data. This number varies between 1 and 6 weeks. The first estimation date starts on 01.01.2012, then Model W1 starts on 24.12.2011, Model W2 on 17.12.2011, Model W3 on 10.12.2011, Model W4 on 13.12.2011, Model W5 26.11.2011, and Model W6 on 19.11.2011. For instance, assume that on 22 February, 23 February will be predicted. In this example, the last date in the data set would be 21.02.2012 while the starting dates for W1, W2, W3, W4, W5, and W6 would be 14.02.2012, 07.02.2012, 31.01.2012, 24.01.2012, 17.01.2012, and 10.01.2012, respectively.

4. Results

In natural gas consumption, customers behave differently. As an example, household natural gas consumption varies depending on climate. In 2012, summer gas consumptions were between 4×10^4 m^3 and 10×10^4 m^3 (Figure 3). These numbers go up between 10×10^4 m^3 and 30×10^4 m^3 in seasonal transitions like fall or spring. For seasonal transitions spring outweighs fall, while the 3×10^5 m^3 and 8×10^5 m^3 range is mostly seen for winter consumptions. Moreover, consumptions between 8×10^5 m^3 and 11×10^5 m^3 are only for winter.

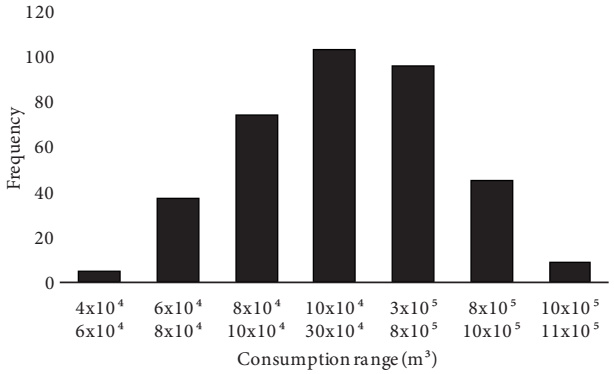

Figure 3. Graph of the frequency for consumption ranges.

Figure 4 shows daily consumptions, daily average temperature, and weekends/weekdays. The black line on the left y-axis represents average temperature in Celsius, the red dotted line on the right y-axis represents daily consumptions, and blue lines show weekends on the graph. In winter with average temperature decreasing, consumption increases proportionally. Likewise, daily consumptions gradually reduce with temperature rises during summer. The impact of weekends on daily consumption can be seen especially clearly on the graph during the summer season.

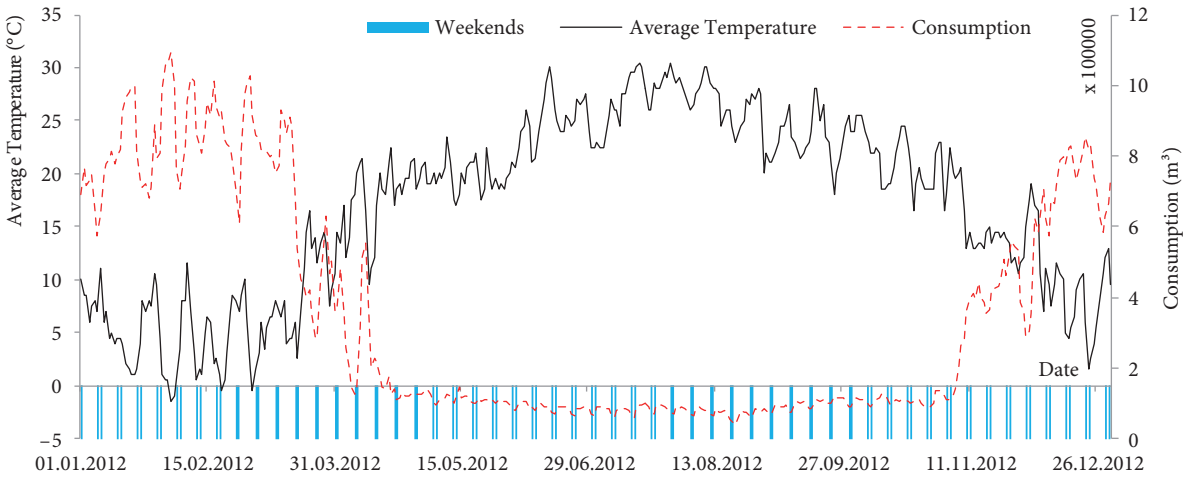

Figure 4. Daily consumption and average temperature with weekends/weekdays.

The consumption values stated before are used for predicting day-ahead consumptions in this study. Each scenario is explored in different subsections. For implementing sliding window functionality, processing data,

computing MLR equations, and gathering performance measurements the Base SAS Engine software, which has its own programming language, is used.

4.1. Results of scenario 1

In this scenario, Model A, which has a growing data set as daily new data are added, and Model W6, which has 6 weeks of data before the prediction date with a sliding window approach, are compared. Here, for every new incoming value, the MLR equation is generated again. Therefore, different predictions would be done for each new data set. For 2 models of this scenario, 712 regression equations are formed in total.

As depicted, meanings of parameters in the equation would change for each new incoming datum (Figure 5). If probabilities of those variables are observed, it is seen that Model A has more meaningfulness than the other (Table 3). The number of probabilities section in Table 3, holiday, maximum temperature, and lag1 mean temperature parameters are formed in Model A and they all have less than 1% probability in all 366 MLR equations. However, this situation is not true for Model W6. Also in the same table, for the percent of probabilities section, if a 10% probability value is accepted as critical, variables other than minimum humidity have more than 50% meaningfulness for both models.

Table 3. Model A and Model W6 predictors' probability comparison table.

Predictors	Holiday		Maximum temperature		Minimum humidity		Number of subscribers		lag 1 mean temperature	
Number of probabilities	Model A	Model W6	Model A	Model W6	Model A	Model W6	Model A	Model W6	Model A	Model W6
Under 1%	366	270	366	183	206	7	111	176	366	178
Under 5%	366	308	366	213	237	43	155	214	366	220
Under 10%	366	326	366	235	281	68	183	230	366	232
Under 50%	366	360	366	300	348	209	262	304	366	299
% of probabilities										
Under 1%	100%	74%	100%	50%	56%	2%	30%	48%	100%	49%
Under 5%	100%	84%	100%	58%	65%	12%	42%	58%	100%	60%
Under 10%	100%	89%	100%	64%	77%	19%	50%	63%	100%	63%
Under 50%	100%	98%	100%	82%	95%	57%	72%	83%	100%	82%

Among the predictions, there are always some unpredictable consumptions that exist. These unpredicted consumptions are called "residual". If Model A and Model W6 are evaluated from a residual view point, the models differentiate after May (Figure 6). Residual distinctness reaches maximum in the summer season for both models. Model W6 has less residual error in summer than fall. Since Model A has whole data, it has the impact of having independent variables in fall. If residual distribution is analyzed, it can be seen that there are close to normal distribution (Figure 6). Having less residual errors in Model W6 than Model A (Figure 7) impacted the MAPE, as well (Figure 8).

The error rates that cannot be seen in the residual graph are obvious in the MAPE graph. In fall, high consumptions cause high residuals. In the same direction, in summer low consumptions cause low residuals. From this view point, in order to rate prediction day errors as percentiles, the MAPE is used. Absolute expression in MAPE shows error directions in the same direction. Although predictions are made with 10% error rate until April, in April and May this rate rises.

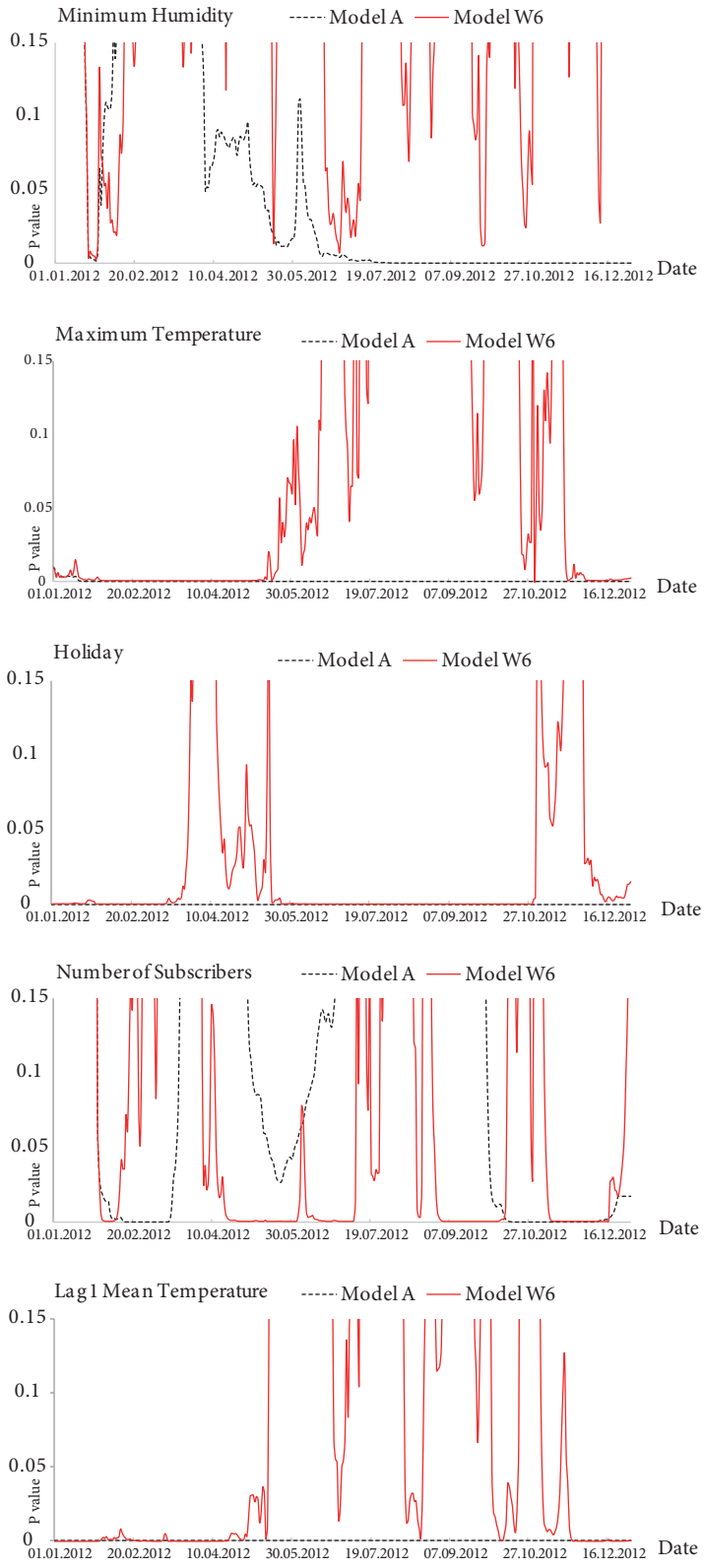

Figure 5. P-values of independent variables in first scenarios.

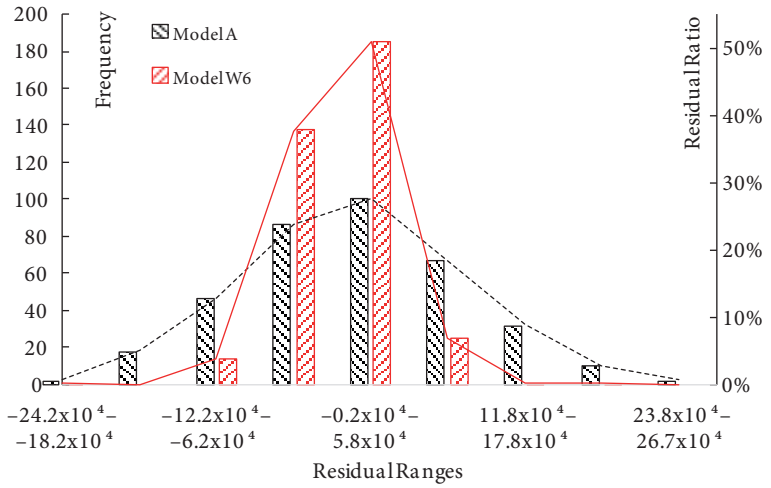

Figure 6. Residual distribution in first scenarios.

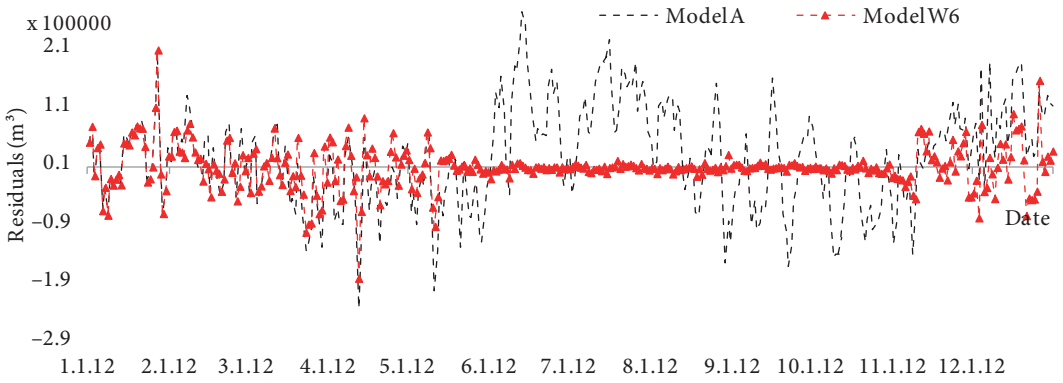

Figure 7. Residuals by date in first scenarios.

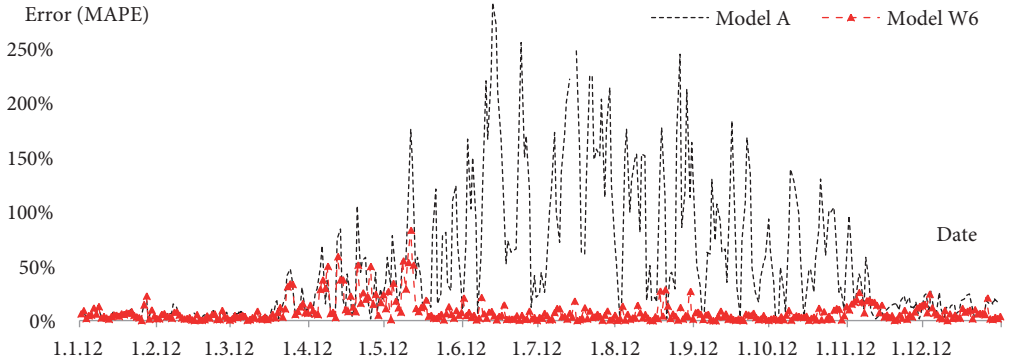

Figure 8. MAPE distribution in first scenarios.

Errors in Model A grow more than in Model W6. Moreover, Model A is affected in summer by the variables in fall, such that the error rate goes up to 300% in summer. If the two models are compared, Model W6 has 8.09% error rate, while Model A has 54.49%. This also means that Model W6 gives more accurate results than the other.

4.2. Results of scenario 2

As seen in the first scenario, an increase in the data set size gives worse predictions. Also, unlike Model A, Model W6 gives better results. This inference leads us to investigate window size effect in the sliding window technique. In other words, the impact of data set size on prediction is studied.

In the second scenario, as a part of the sliding window method 6 different window sizes are chosen. Sizes vary between 7 days (1 week) and 42 days (6 weeks). Six models are applied for each week. Here, models are named as "w" to show that data are weekly collected. The next numerical value stands for the number of weeks.

A total of 2196 MLR equations are used for models of the scenario. Table 4 presents all models where 366 MLR equations are formed and the number of 10% or less critical probability values. Since Model W1 and Model W2 have the lowest values for all parameters, they are ignored. Obviously, meaningful parameter prediction increases as the number of data grows.

Table 4. Number of significant predictors for second scenario models in 10% probability.

Predictors	Holiday	Maximum temperature	Minimum humidity	Number of subscribers	lag 1 mean temperature
Model W1	64	50	36	43	46
Model W2	266	156	37	137	133
Model W3	302	185	46	166	194
Model W4	304	227	58	178	218
Model W5	316	236	75	213	233
Model W6	326	235	68	230	232

In this research, 3 weeks are studied as critical and results are shown in prediction models of 3–6 weeks (Figure 9). It is seen that the variables of maximum temperature and lag1 mean temperature are meaningful in winter month consumptions. The meaningfulness of the minimum humidity parameter generally remains low. Number of subscribers is meaningfully used where temperature variables such as maximum temperature and lag1 mean temperature are insufficient. On the other hand, holidays are usually a meaningful variable. Although in spring and fall, the holiday parameter goes outside of the probability range, they are still the most meaningful variables.

In Figure 10 residual results are presented. Depending on consumptions, there are high residuals until June. In summer months and October low residuals increase in the graph. In seasonal transitions such as April, May, and October, consumption differentials between consumptions and predictions are equal to consumption and prediction differences in winter. Negative and positive residuals in the consumption residual table show that randomness is provided.

Figure 11 presents consumption predictions with MAPE. Here, the highest error is 90% with Model W6. The second highest error is 80% with Model W3. Also in the MAPE graph, Model W3 and Model W6 make predictions with higher error rates than 2 other models. This demonstrates the impact of data set size on prediction.

In Figure 12a residual distributions are presented. All models in the scenario make predictions in similar directions with normal distribution of errors. For error frequencies of consumption predictions, Model W3 and Model W6 make predictions with high residuals. The lowest residual ranges are -6.2×10^4 m^3 to -0.2×10^4 m^3 and -0.2×10^4 m^3 to 5.8×10^4 m^3. Numbers of prediction errors for Model W4, Model W5, and Model W6 are similar.

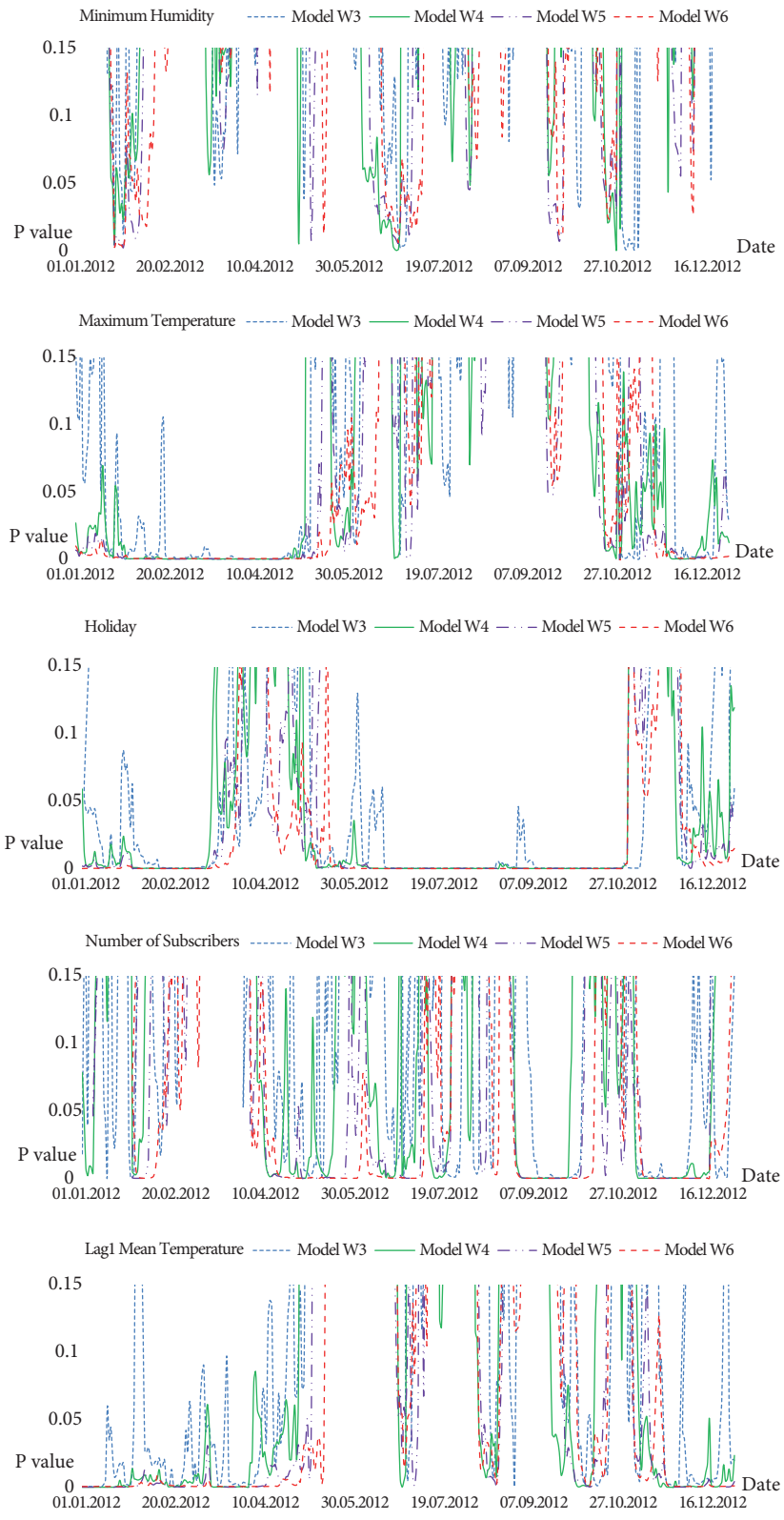

Figure 9. P-values of independent variables in second scenarios.

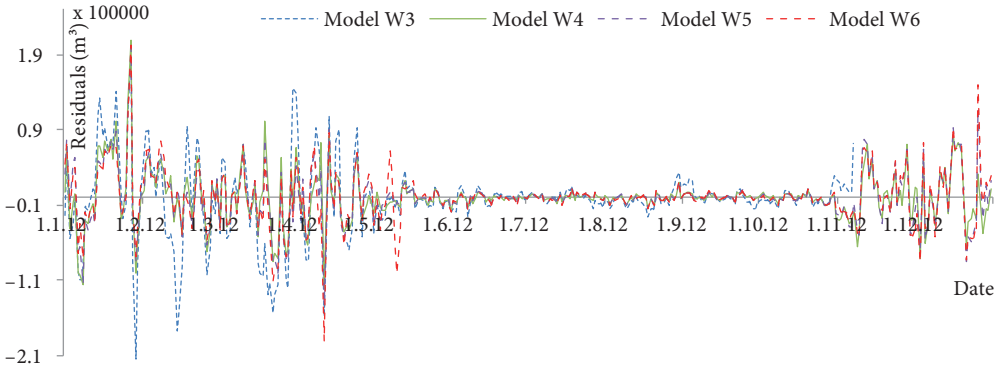

Figure 10. Residuals by date in second scenarios.

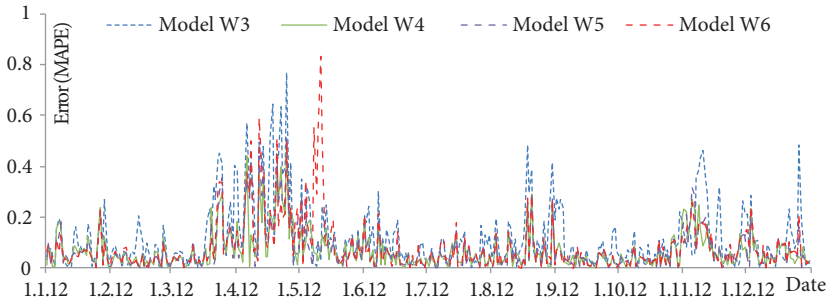

Figure 11. MAPE distribution in second scenarios.

However, there are some cases that the residual distribution graph is not able to predict. For instance, forecasting done in winter has a very low residual, so that percentile error would be low as well. The same situation occurs in summer, as well. Since prediction in summer has very high residual, it also has high percentile error. However, in winter months, where consumption increases, for the same residual amounts in summer, the predicted day percentile consumption error is less. In order to deduct this, not only the residual graph but also the MPE graph should be observed. In addition to the MAPE graph, the MPE graph shows negative percentile errors (inadequate predictions) (Figure 12b). As seen, the MPE graph gives similar results to the residual distribution graph. However, Model W4 in the MPE graph has better performance then Model W5. In the residual distribution graph this situation is opposite, with performance for Model W5. If the MAPE is used for error rate, the MAPEs for Model W3, Model W4, Model W5, and Model W6 are 11.8%, 6.8%, 7.2%, and 8.1%, respectively.

As a general evaluation of the study (Table 5), Model A in the first scenario has 2 times a higher error rate then Model W1, which has the highest error rate among the 6 models in the second scenario. In the first part of the table, the number of predictions higher than the related MAPE rate is shown in model basis. The middle part of the table demonstrates the numbers as percentile rates. The last row of the table indicates each model's own mean absolute percent errors. Except for Model A and Model W1, error rates vary between 6.8% and 14.1%.

5. Conclusions

In this study, for predicting natural gas consumptions, a sliding window method using multiple linear equations and a traditional method are compared for various data set sizes. Before the models are formed, first, MLR is applied to all data and meaningless variables are ignored. In the second phase of data preparation, a correlation

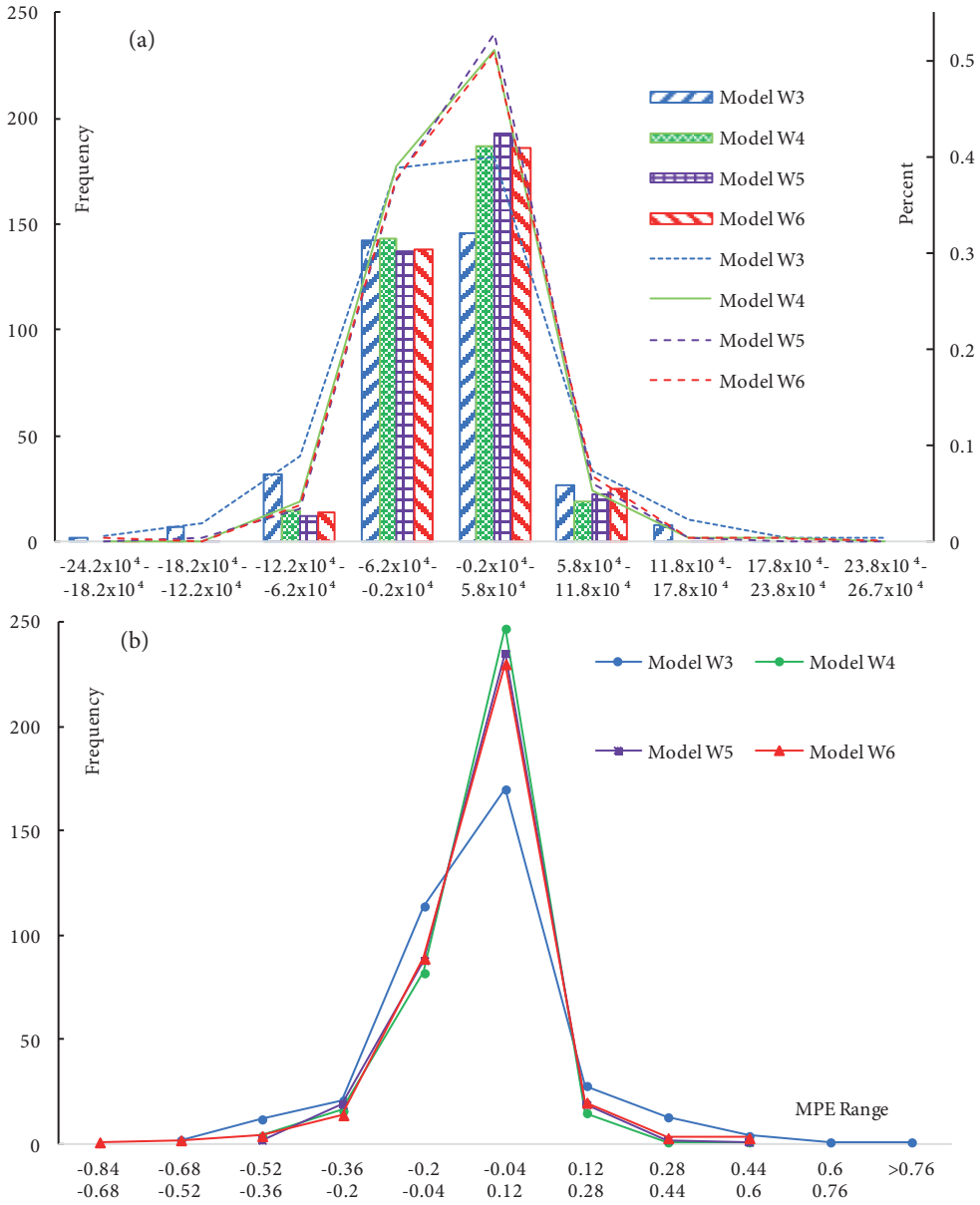

Figure 12. Results for second scenarios: a) Residual distribution, b) MPE distribution.

Table 5. Model-based MAPE, number of days with MAPE and ratios.

	Model W1	Model W2	Model W3	Model W4	Model W5	Model W6	Model A
Up to 5%	295	244	236	164	169	176	302
Up to 10%	231	163	141	72	80	85	252
Up to 15%	179	110	91	45	46	49	220
Up to 20%	139	76	61	27	28	33	197
Percent of up to 5%	80.6%	66.7%	64.5%	44.8%	46.2%	48.1%	82.5%
Percent of up to 10%	63.1%	44.5%	38.5%	19.7%	21.9%	23.2%	68.9%
Percent of up to 15%	48.9%	30.1%	24.9%	12.3%	12.6%	13.4%	60.1%
Percent of up to 20%	38.0%	20.8%	16.7%	7.4%	7.7%	9.0%	53.8%
MAPE	26.3%	14.1%	11.8%	6.8%	7.2%	8.1%	54.5%

table is created and variables having collinearity are discarded. After this step 2 distinct scenarios are developed. In the first scenario, one of the data sets, called Model A, continuously extends. The other data set, called Model W6, dynamically moves with the sliding window technique and carries 6 weeks of data. For the second scenario, 6 different models are generated and named after the number of weeks that they include. Error rates in the 2 models as observed from high to low are Model A, Model W1, Model W2, Model W3, Model W6, Model W5, and Model W4. The exact rates are 54.5%, 26.3%, 14.1% 11.8%, 8.1%, 7.2%, and 6.8%, respectively. The lowest error in this study is obtained with Model W4.

For future work, logarithmic and exponential transitions will be applied for data sets and later predictions will be done with the sliding window technique. Instead of MLR, as a theoretical method, curve fitting, learning algorithms like artificial neural networks, time series analysis like ARIMA, or heuristic techniques like genetic algorithms can be used.

Acknowledgments

The first author would like to thank Kevser Ovaz and the Adapazarı Natural Gas Distribution Company for their support.

References

[1] Sarak H, Sakman V. The degree-day method to estimate the residential heating natural gas consumption in Turkey: a case study. Energy 2003; 28: 929-939.

[2] Brown RH, Kharouf P, Feng X, Piessems P, Nestof D. Development of feed-forward network models to predict gas consumption. In: IEEE International Conference on Neural Networks; 28 June–2 July 1994; Orlando, FL, USA. New York, NY, USA: IEEE. pp. 802-805.

[3] Gill S, Deferrar, J. Generalized model of prediction of natural gas consumption. J Energ Resour-ASME 2004; 126: 90-98.

[4] Akpinar M, Yumusak N. Forecasting household natural gas consumption with ARIMA model: a case study of removing cycle. In: IEEE International Conference on Application of Information and Communication Technologies; 23–25 October 2013; Baku, Azerbaijan. New York, NY, USA. pp. 319-324.

[5] Akpinar M, Yumusak N. Estimating household natural gas consumption with multiple regression: effect of cycle. In: IEEE International Conference on Electronics, Computer and Computation; 7–9 November 2013; Ankara, Turkey. New York, NY, USA. pp. 188-191.

[6] Potocnik P, Thaler P, Govekar E, Grabec I, Poredos A. Forecasting risks of natural gas consumption in Slovenia. Energ Policy 2007; 35: 4271-4282.

[7] Brabec M, Konár O, Pelikán E, Malý M. A nonlinear mixed effects model for the prediction of natural gas consumption by individual customers. Int J Forecasting 2008; 24: 659-678.

[8] Aydinalp-Koksal M, Ugursal VI. Comparison of neural network, conditional demand analysis, and engineering approaches for modeling end-use energy consumption in the residential sector. Appl Energ 2008; 85: 271-296.

[9] Sabo K, Scitovski R, Vazler I, Zekic′-Sušac M. Mathematical models of natural gas consumption. Energ Convers Manage 2011; 52: 1721-1727.

[10] Catalina T, Iordache V, Caracaleanu B. Multiple regression model for fast prediction of the heating energy demand. Energ Buildings 2013; 57: 302-312.

[11] Gembris D, Taylor JG, Schor S, Frings W, Suter D, Posse S. Functional magnetic resonance imaging in real time (FIRE): sliding-window correlation analysis and reference-vector optimization. Magnet Reson Med 2000; 43: 259-268.

[12] Lee CH, Lin CR, Chen MS. Sliding-window filtering: an efficient algorithm for incremental mining. In: Tenth International Conference on Information and Knowledge Management; 5–10 November 2001; Atlanta, GA, USA. New York, NY, USA: ACM. pp. 263-270.

[13] Luiz SO, Perkusich A, Lima AMN. Multisize sliding window in workload estimation for dynamic power management. IEEE T Comput 2001; 59: 1625-1639.

[14] Suzuki Y, Ibayashi H, Kaneda Y, Mineno H. Proposal to sliding window-based support vector regression. Procedia Computer Science 2014; 35: 1615-1624.

[15] Makridakis S, Wheelwright SC, Hyndman RJ. Forecasting Methods and Applications. 3rd ed. New York, NY, USA: Wiley, 1998.

[16] Mason RL, Gunst RF, Hess JL. Statistical Design and Analysis of Experiments with Applications to Engineering and Science. 2nd ed. New York, NY, USA: Wiley, 2003.

[17] Rosenthal JA. Statistics and Data Interpretation for Social Work. 1st ed. New York, NY, USA: Springer, 2012.

[18] Davis BJ. Statistics Using SAS Enterprise Guide. 1st ed. Cary, NC, USA: SAS Institute, 2007.