Turk J Elec Eng & Comp Sci, Vol.20, No.5, 2012, © TÜBİTAK
doi:10.3906/elk-1101-1029

Forecasting natural gas consumption in İstanbul using neural networks and multivariate time series methods

Ömer Fahrettin DEMİREL[1], **Selim ZAİM**[2,*], **Ahmet ÇALIŞKAN**[3], **Pınar ÖZUYAR**[4]

[1]*Department of Industrial Engineering, Fatih University, Büyükçekmece,*
34500 İstanbul-TURKEY
e-mail: odemirel@fatih.edu.tr

[2]*Faculty of Economics and Administrative Sciences, Fatih University,*
Büyükçekmece, 34500 İstanbul-TURKEY
e-mail: szaim@fatih.edu.tr

[3]*Department of Economics, Fatih University, Büyükçekmece, 34500 İstanbul-TURKEY*
e-mail: ahmetcaliskan@fatih.edu.tr

[4]*Center for Energy, Environment, and Economy, Özyeğin University,*
Altunizade, 34662 İstanbul-TURKEY
e-mail: pinar.ozuyar@ozyegin.edu.tr

Received: 19.01.2011

Abstract

The fast changes and developments in the world's economy have substantially increased energy consumption. Consequently, energy planning has become more critical and important. Forecasting is one of the main tools utilized in energy planning. Recently developed computational techniques such as genetic algorithms have led to easily produced and accurate forecasts. In this paper, a natural gas consumption forecasting methodology is developed and implemented with state-of-the-art techniques. We show that our forecasts are quite close to real consumption values. Accurate forecasting of natural gas consumption is extremely critical as the majority of purchasing agreements made are based on predictions. As a result, if the forecasts are not done correctly, either unused natural gas amounts must be paid or there will be shortages of natural gas in the planning periods.

Key Words: *Forecasting, neural networks, natural gas, time series*

1. Introduction

Energy is considered to be a significant input in economic development and a prime factor in the generation of wealth. It is also essential to social development and improved quality of life. Taking the above indicators as fundamental requirements, Turkey's national energy policies are designed to provide the required energy on a timely, reliable, cost-effective, environmentally friendly, and high-quality basis so as to serve as the driving

[*] Corresponding author: Faculty of Economics and Administrative Sciences, Fatih University, Büyükçekmece, 34500 İstanbul-TURKEY

force of development and social progress [1,2]. With such intent, the term “energy security” is rising higher on the agenda of policymakers and city administrators, not just in Turkey, but also in the European Union and the Balkan countries.

Understanding its unique position, Turkey strives to be the Eurasian energy corridor between eastern suppliers and western consumers. Hence, Turkey is expected to attract foreign investments in natural gas in the coming decades [1,2]. Due to its strategic position and geographic location between European markets and Central Asian and Middle Eastern producing countries, Turkey provides, in this context, a most convenient passage [3]. Thus, it is situated in a strategically advantageous position in terms of the Eurasian natural gas market. Being located between Europe and the energy-rich countries of Central Asia, it can be an energy corridor between these 2 regions. It can import gas from a number of countries and diversify its sources. This situation may also provide motivation for a competitive gas market. The number of suppliers is one of the factors affecting the market price competitiveness [2].

Worldwide, most of the increase in the demand for natural gas is expected to come from developing countries, where gas consumption is expected to grow from 0.5 tcm in 1999 to 4.5 tcm in 2020. Following the economic crises worldwide, on the assumption that the world economy began to recover in 2010, primary gas demand is projected to rebound, growing on average by 2.5% per year between 2010 and 2015. Gas demand in the OECD countries is estimated to increase through 2030, as coal- and fuel-oil-fired generators are replaced by plants using renewable and nuclear power [2]. The use of natural gas in OECD countries is projected to grow by 2.4% annually, compared to a more modest rate of increase of 1.1% in oil consumption. This would account for 49% of the projected increase in total energy consumption of these countries [1].

In Turkey, however, a 4.3% increase in the annual average in real GDP for the last 7 years has been realized. Analyses indicate that Turkish economy is expected to experience an annual real GDP growth rate of 6.7% during 2011-2017. Furthermore, according to data from the United Nations Conference on Trade and Development, Turkey ranked as the 15th most attractive foreign direct investment destination for 2008-2010.

Turkey’s current annual total primary energy supply (TPES) of 99 Mtoe is expected to increase by 87% [4]. However, Turkey, which is still extensively dependent on energy imports as it was in the past, can only meet 28% of its energy demand through its own resources. The shares dependent on foreign supplies of oil and natural gas are, respectively, 93% and 97%, which are much higher. While the crude oil and oil products cost the country US$8.6 billion in 2004, this figure increased to $12.4 billion in 2005. Parallel to this, while the cost of the natural gas was $4.4 billion in 2004, it increased to $7.1 billion in 2005. Including coal and electrical energy, the net importation cost reached $18.6 billion, with an increase of 30% in 2005 relative to 2004. This means that Turkey spends 25% of its income on imported energy [5,6].

Natural gas has met a major part of Turkey’s rapidly growing energy needs, rising from hardly 6% of the TPES in 1990 to 31% in 2008. Currently, Turkey imports 98% of its gas needs. Despite Turkey’s intensified efforts in exploration, domestic gas production is only around 1 bcm/year. The remaining recoverable natural gas reserves are predicted to be 6 bcm. Power generation was the largest gas user in 2008, accounting for 55% of total demand. Households consumed 22% of all gas, industry 11%, services 10%, and other sectors 2%. Gas demand almost doubles in winter when gas use in the residential and power sectors is at its highest. The residential sector is the main contributor to the growing seasonality of natural gas demand.

Given that Turkey’s natural gas consumption depends heavily on imports and that natural gas purchase contracts have been signed in accordance with base annual purchase, exporting countries can deliver the contracted gas anytime within the year. The contracts do not take into consideration the fluctuations that would occur within the year. In such a volatile environment, suppliers can put ceilings on imports for economic and political reasons. There are 2 feasible alternatives to remedy this uncertainty: constructing storage facilities and diversifying sources. Nonetheless, flexibility is very limited in the short term; most gas users are not practically amenable to utilization of alternative fuels. This fact is accentuated by the lack of abundant and adequate natural gas storage capacity in Turkey. It is very important to forecast short-term gas consumption in order to optimize strategic transmission, storage, and distribution plans and the need for relevant necessary investments [3,4,7].

Since the early 1970s, various studies of energy demand have been undertaken using a range of estimation methods. In most of these studies, the purpose has been to measure the impact of economic activity and energy prices on energy demand, such as estimating income and price elasticity, which are of the utmost importance in forecasting energy demand. Evidence shows long-term income elasticity around unity, or slightly above, and price elasticity is typically found to be rather small [8].

Although there are a number of studies on the forecasting of long-term natural gas demand in Turkey, short-term studies are scarce. This paper evaluates the real industry implications of the existing forecasting methods and applies neural networks and multivariate time series methods to predict natural gas consumption for a leading Turkish company operating in the energy sector.

2. Background information about research methodology

2.1. OLS regression

The ordinary least squares (OLS) regression model was employed as a benchmark model upon which the performances of the more advanced autoregressive and moving average (ARMA) and artificial neural network (ANN) models were compared. In a time series context, the OLS regression model takes the form of:

$$y_t = \beta X_t + \varepsilon_t. \tag{1}$$

Here, left-hand-side variable y is the dependent variable and right-hand-side vector X includes independent variables. Normally, vector X includes variables that are considered to be factors that influence dependent variable y. Within the context of our time series approach in this paper, however, X also includes lags of dependent variable y. The particular OLS specification employed in this paper was discovered after the model identification process conducted in Section 4.

2.2. ARMAX model

Autoregressive time series models can be structured as univariate or multivariate models. In a univariate specification called the autoregressive integrated moving average (ARIMA), future forecasts of a variable are based on the historical values of the same variable. ARIMA specification takes the following general form:

$$y_t = a_0 + a_1 y_{t-1} + ... + a_p y_{t-p} + \varepsilon_t + b_1\varepsilon_{t-1} + ... + b_q\varepsilon_{t-q}. \tag{2}$$

Here, there are potentially p autoregressive (AR) and q moving average (MA) terms. p and q are integers and are called orders of the model. a_i (i = 0,1,2, ... p) and b_i (i = 1,2,... q) are parameters to be estimated. Disturbances ε_t are assumed to be independent and identically distributed (iid) and normally distributed with mean zero and varianceσ^2.

In many cases, in addition to the lags of the dependent variable, it is useful to include independent factors in the specification that are considered to influence the dependent variable. Since natural gas consumption is forecasted in this paper, it is natural to include other independent variables such as temperature, price of natural gas, and number of gas consumers. Fortunately, the statistics package that was used allows the specification of a structural model with AR and/or MA disturbances. This type of a specification is called an ARMAX model. Formally, an ARMAX model can be expressed as a structural equation and an equation to specify ARMA disturbances, respectively [9]:

$$y_t = X_t \beta + \mu_t, \tag{3}$$

$$\mu_t = a_1 \mu_{t-1} + ... + a_p \mu_{t-p} + \varepsilon_t + b_1 \varepsilon_{t-1} + ... + b_q \varepsilon_{t-q}. \tag{4}$$

Disturbances of the structural equation μ_t are allowed to follow an ARMA (p, q) process. Likewise, ε_t values are assumed to be iid and normally distributed with mean zero and varianceσ^2. Vector X_t includes both independent factors and possibly lags of the dependent variable. This specification is flexible enough to exploit all available historical information content for forecasting future values of the dependent variable.

2.3. ANN models

ANNs are analytic techniques modeled on the learning processes of the human cognitive system and the neurological functions of the brain. Recently, there has been considerable interest in the development of ANNs for solving a wide range of problems from different fields [10-12]. Neural networks are distributed information processing systems composed of many simple computational elements interacting across weighted connections. Inspired by the architecture of the human brain, neural networks exhibit certain features such as the ability to learn complex patterns of information and generalize the learned information. Neural networks are simply parameterized nonlinear functions that can be fitted to data for prediction purposes [13].

ANNs can be classified into several categories based on supervised and unsupervised learning methods and feedforward and feedback recall architectures. Multilayered feedforward networks use a supervised learning method and feedforward architecture. A backpropagation neural network is one of the most frequently utilized neural network techniques for classification and prediction [14,15].

The main appeal of neural networks is their flexibility in approximating a wide range of functional relationships between inputs and outputs. Indeed, sufficiently complex neural networks are able to approximate arbitrary functions arbitrarily well. One of the most interesting properties of neural networks is their ability to work and forecast even on the basis of incomplete, noisy, and fuzzy data. Furthermore, they do not require a priori hypotheses and do not impose any functional form between inputs and outputs. For this reason, neural networks are quite practical to use in cases where knowledge of the functional form relating inputs and outputs is lacking, or when a prior assumption about such a relationship should be avoided [16].

The success of ANN models depends on properly selected parameters such as the number of nodes (neurons) and layers, the nonlinear function used in the nodes, the learning algorithm, the initial weights of the inputs and layers, and the number of epochs for which the model is iterated. The structure of the ANN model shown in Figure 1 consists of an input layer, a hidden layer, and an output layer. The neurons are connected by weights, depicted as lines in Figure 1.

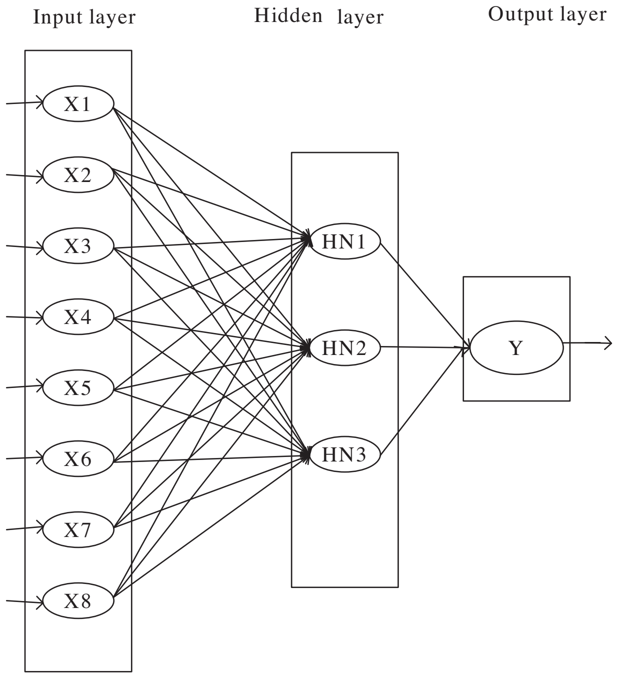

Figure 1. Structure of ANN model.

In ANN methodology, the sample data are often divided into 2 main subsamples, which are called training and test sets. During the training process, the neural network learns the relationship between the output and input criteria, while in the testing process, the test set is used to assess the performance of the model.

During the learning stage, the initial weights of the model are trained. Because of the strong nonlinear correlation between evaluation criteria, 3 nonlinear evaluation models are taken into consideration: the back propagation algorithm, NeuroShell's TurboProp, and genetic training methods [17]. They carry out supervised learning of neural network weights using training data as inputs and a known output minimizing the mean square error (MSE). During the training process, the neural network learns the relationship between output and input nodes. The input nodes are the previous lagged observations while the output provides the forecast for the future value. Hidden nodes with appropriate nonlinear transfer functions are used to process the information received by the input nodes. The network arcs connect the processing units to the neurons. Weight values are assigned to the links between the inputs and outputs neurons. The inputs of each neuron from other neurons are aggregated. The net value represents a weighted combination of the neuron inputs. The hierarchy is coded in a hierarchical neural network, where each neuron corresponds to a criterion [18].

3. The proposed forecasting methodology

The individual steps of the study procedure are shown in Figure 2. The principal purposes of this procedure are to find the most appropriate method for forecasting the gas consumption in the city of İstanbul, and to select the most important evaluation criteria to improve the forecasting performance and generate recommendations for the upper management of the company.

Step 1: In the first stage, data are collected, examined, and preprocessed. The available data are converted into scale of 0 to 1 through the preprocessing step by using:

$$NV = \frac{Y_t - Y_{\min}}{Y_{\max} - Y_{\min}}, \tag{5}$$

where NV is the normalized value, Y_t refers to the actual consumption value in time period t, and $Y_{\max}$ and $Y_{\min}$ represent the maximum and minimum consumption values of the dataset.

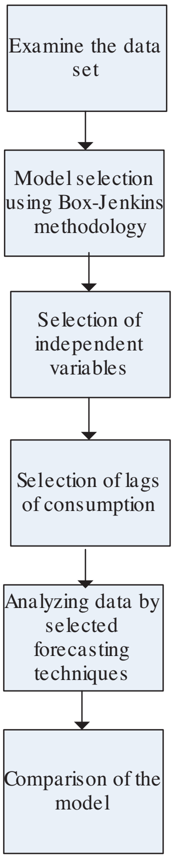
Examine the data set
Model selection using Box-Jenkins methodology
Selection of independent variables
Selection of lags of consumption
Analyzing data by selected forecasting techniques
Comparison of the model

Figure 2. Proposed steps of forecasting methodology.

Step 2: In this step, Box-Jenkins methodology is utilized to select the most appropriate model in the time series.

Step 3: Independent variables are selected for further analysis.

Step 4: Lags of consumption are selected using Box-Jenkins methodology.

Step 5: State-of-the-art forecasting techniques such as traditional time series methods and neural network approaches are evaluated and the best-performing model specifications are selected.

Step 6: The accuracy of the forecasting methods are compared by using 3 different performance measures. These are root mean squared error (RMSE), mean absolute deviation (MAD), and mean absolute percentage error (MAPE).

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2}{n}} \tag{6}$$

$$MAD = \frac{\sum_{t=1}^{n}|Y_t - \hat{Y}_t}{n} \tag{7}$$

$$MAPE = \frac{\sum_{t=1}^{n} \frac{(Y_t - \hat{Y}_t)}{Y_t}}{n} \tag{8}$$

Here, Y_t refers to the actual value and $\hat{Y}_t$ refers to the predicted value at time t, and n is the number of observations in the test dataset. After selecting the best model based on the performance measures mentioned above, the most important variables affecting the natural gas consumption were determined.

4. Application of the proposed forecasting methodology

The objective of this study was to develop a general model that will help to predict natural gas consumption for a major Turkish natural gas distributor, İstanbul Gas Distribution Industry and Trade Joint Stock Company (İGDAŞ), which is the largest natural gas distributor in Turkey.

The steps of the forecasting methodology are as follows.

Step 1: Investigate the characteristics of the natural gas consumption data.

The daily natural gas consumption data is plotted in Figure 3. The dataset covers the time period from 1 January 2004 through 31 December 2009. Consumption appears to show a strong seasonality depending on daily temperature. This is related to the impact of the share of natural gas used for residential and commercial heating in İstanbul.

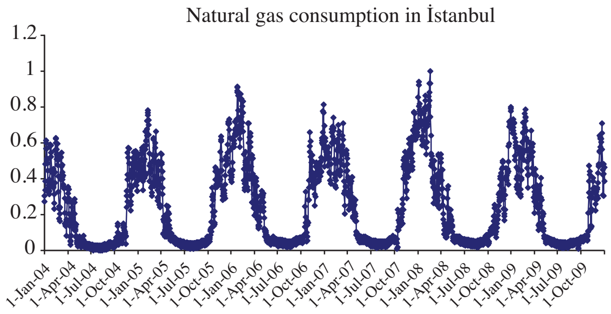

Figure 3. Daily natural gas consumption in İstanbul.

Daily gas consumption and all independent variables used in the analysis were normalized to the 0-1 range. This type of linear normalization does not influence the results of the OLS and ARMAX models since these are linear models. ANN algorithms are designed to pick nonlinear patterns in the data. In ANN models, normalization makes it easy to compare weights of links between neurons. Normalization also seems to be a common practice in the literature of ANN forecasting.

Step 2: The Box-Jenkins methodology of model selection.

Box-Jenkins [19] methodology is frequently used in choosing the best model in time series analyses [20]. The stages of Box-Jenkins methodology are model identification, parameter estimation, and diagnostic checking. At the model identification stage, one needs to check whether the time series are stationary or not. Formally, for a series to be "covariance stationary," its mean, variance, and autocorrelations should be finite and should not depend on time [21]. If the analyst runs OLS or ARIMA estimation on variables that are nonstationary, statistically significant relationships between variables are observed erroneously. This fallacy is called "spurious regression" [21-23].[1] To avoid this fallacy, we test the natural gas consumption data for stationarity below.

If time series data are not stationary, they are usually said to follow a unit root process. Unit root processes are a generalization of random walk processes with serially correlated errors [21]. Daily natural gas consumption data were initially tested for stationarity using the Dickey-Fuller and Philips-Perron tests. Both tests rejected the unit root at a 1% significance level. The augmented Dickey-Fuller (ADF) test was also used for 25 lags of the daily consumption series. The ADF test rejected the unit root in 20 of 25 lags of consumption at 5%, and the remaining lags were rejected at 10% significance.[2] These results reveal that consumption series do not follow a unit root process.

Both autocorrelation and partial autocorrelation plots of the daily consumption values were checked. The autocorrelations died out slowly, and partial autocorrelations of the first lag were quite high while the other lags were very small. This autocorrelation structure suggested that we should use an AR term of order one (AR(1)) in our ARMAX model [23]. Combined with unit root test results, we concluded that the consumption series do not follow a unit root process, but today's consumption level is possibly largely determined by historical consumption levels. This property of the data strongly justified employing time series models for forecasting consumption.

Step 3: Selection of independent variables.

Careful inspection of the consumption data in Figure 3 helped us determine which variables to include in our model specification. First of all, there is high seasonality in consumption, and it appears that daily consumption heavily depends on the weather conditions and daily temperature. This is because the natural gas distributed by İGDAŞ is primarily used for residential and commercial heating.[3] Therefore, daily average temperature, measured in degrees Celsius, was included as an independent variable. As the temperature drops, consumption is expected to increase, and vice versa. A negative coefficient for the temperature variable was thus expected. Potentially, the relationship between temperature and consumption may be nonlinear. In order to capture the second-order (or higher) effects of temperature, squared temperature (or possibly a higher-order term) was included as another independent variable.

Economically, demand for gas should be related to its price. Historical subscriber price information from İGDAŞ[4] was obtained. The subscriber price was expressed in terms of Turkish lira (TL)/m^3. A plot of the price series in Figure 4 indicates that price does not fluctuate frequently; sometimes it stays constant for months. This is quite natural since İGDAŞ is a natural monopoly and subscriber prices are heavily regulated by the Energy Market Regulatory Authority of Turkey. Although there have been attempts to create and strengthen competition in the natural gas distribution market in Turkey since 2001, progress has been sluggish [24]. Since price is heavily regulated and does not appear to respond to fluctuations in demand, there seems to be little problem with taking price as exogenous. The law of demand states that there should be a negative relationship between the price and the quantity of gas demanded. Hence, a negative effect of price on the daily consumption level is expected.

In this study, the number of gas subscribers was also controlled. The number of İGDAŞ subscribers has rapidly increased from around 2.3 million at the beginning of 2004 to close to 4.2 million at the end of

[1] Suppose 2 processes, x_t and y_t, both follow independent random walks such that: $x_t = x_{t-1} + \varepsilon_t$ and $y_s = y_{s-1} + \delta_s$, where $E(\varepsilon_t, \delta_s) = 0$ for all t and s. If we run y_t on x_t by OLS, then we find high R^2 values and a statistically significant effect of x on y [21-23].

[2] Results are available upon request from the authors.

[3] For detailed information about how much natural gas was used for various purposes in Turkey, see [21].

[4] There is another price listed by İGDAŞ, namely the end consumer price. However, since the 2 prices are very closely correlated with each other, we chose to include the subscriber price in the analysis.

2009. Figure 5 presents the number of subscribers over the 5-year period. Such a strong growth in the number of consumers made it necessary for us to account for this factor in demand estimation. Naturally, as more subscribers demand gas, the total consumption is expected to increase.

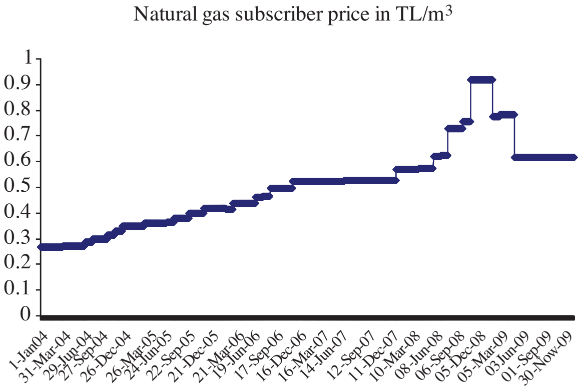

Figure 4. Price of natural gas.

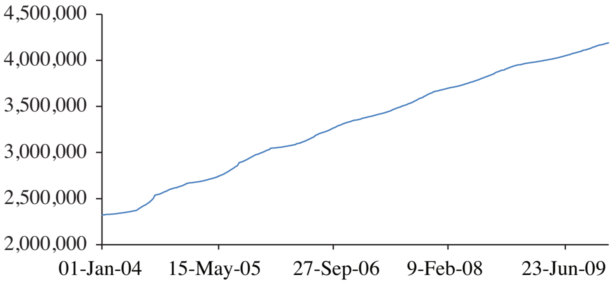

Figure 5. Number of gas subscribers.

Step 4: Selection of lags of consumption.

Lags of consumption were chosen iteratively based on Box-Jenkins methodology. First of all, it is obvious that the first lag (yesterday's consumption) contains considerable prediction power for today's consumption level. As a first step, we included all independent variables mentioned above and the first lag of consumption on the right-hand side of our ARMAX specification. In addition, disturbances of the structural equation were allowed to follow an AR(1) process. Formally, the initial model (Model I) is expressed as:

$$cons_t = \beta_0 + \beta_1 cons_{t-1} + \beta_2 temp_t + \beta_3 tempsq_t + \beta_4 price_t + \beta_5 numcons_t + \mu_t, \tag{9}$$

$$\mu_t = a_1 \mu_{t-1} + \varepsilon_t \tag{10}$$

In the above equation, the terms *cons*, *temp*, and *tempsq* represent consumption, temperature, and squared temperature, respectively. *Price* is the subscriber price, and *numcons* represents the number of consumers. Subscript t denotes day t, where t = 1,..., 2192 for the full sample.

The above model was estimated and plotted autocorrelations of the residuals[5] were obtained from this estimation. Residual autocorrelations of Model I are presented in Figure 6.

If the model specification is satisfactory, residuals should not exhibit large autocorrelations. This is because, according to the assumptions of the OLS and ARMAX models, disturbances[6] (denoted as ε_t in Eqs. (1), (2), (4), and (10)) need to be independent of each other. Residual autocorrelations estimated from Model I were all less than 0.15 in absolute value except at lags of multiples of 7 days. At lags of 7, 14, 21, 28, and so on, autocorrelations exhibited spikes that exceeded 0.2 in magnitude. This autocorrelation structure suggests that there are weekly effects that we need to consider in the demand estimation. When further lags of up to a few months and 1 year are considered, the weekly pattern of spikes of the autocorrelation continues until the end of the year. This suggests that the gas consumption level weeks or months ago includes useful information for today's consumption. Beyond purely statistical concerns and from a practical perspective, if the results reveal that the weeks or months ahead of a lagged demand are powerful predictors for today's consumption level, this

[5] "Residual" here refers to the difference between the actual consumption value and the value predicted by our model.

[6] Residuals are empirical, consistent estimates of disturbances. The term "disturbance" refers to the theoretical error. See [25].

information must be used to increase our forecasting accuracy. Lags to be included in the model were chosen as independent variables based on a trial-and-error procedure. ARMAX regressions were run, first adding up to 6 weeks of lags and then, after that, lags for every month up to 1 year. The lags that were statistically significant in the regressions (at least 10% significance based on z-statistics) were kept and insignificant lags were dropped. After a number of iterations, it was found that the following lags of consumption were significant: lags 1, 7, 14, 21, 28, 196 (28 weeks), 280 (40 weeks), 336 (48 weeks), 343 (49 weeks), 350 (50 weeks), 357 (51 weeks), and 364 (52 weeks).

The final model used in this paper consisted of 4 independent variables and 12 lags of today's consumption. The independent variables were temperature, squared temperature, price, and the number of consumers. The lags were the 12 lags of consumption mentioned above. The formal model is given below.

$$\begin{aligned} cons_t = {} & \beta_0 + \beta_1 cons_{t-1} + \beta_2 cons_{t-7} + \beta_3 cons_{t-14} + \beta_4 cons_{t-21} + \beta_5 cons_{t-28} + \beta_6 cons_{t-196} \\ & + \beta_7 cons_{t-280} + \beta_8 cons_{t-336} + \beta_9 cons_{t-343} + \beta_{10} cons_{t-350} + \beta_{11} cons_{t-357} + \beta_{12} cons_{t-364} \\ & + \beta_{13} temp_t + \beta_{14} tempsq_t + \beta_{15} price_t + \beta_{16} numcons_t + \mu_t \end{aligned} \tag{11}$$

$$\mu_t = a_1 \mu_{t-1} + \varepsilon_t \tag{12}$$

It is to be noted that Eq. (12) specifies an AR(1) disturbance structure and is only available in the ARMAX model, not in the OLS model. The OLS model consists of only Eq. (11). Residual autocorrelations of the above ARMAX model are presented in Figure 7. Autocorrelations of lags of up to 360 days were mostly within the [–0.05, +0.05] range and mostly inside the 95% confidence band.

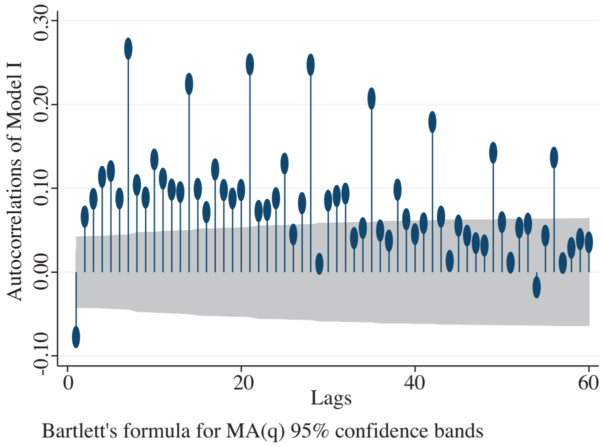

Figure 6. Autocorrelations of Model I residuals: 60 lags.

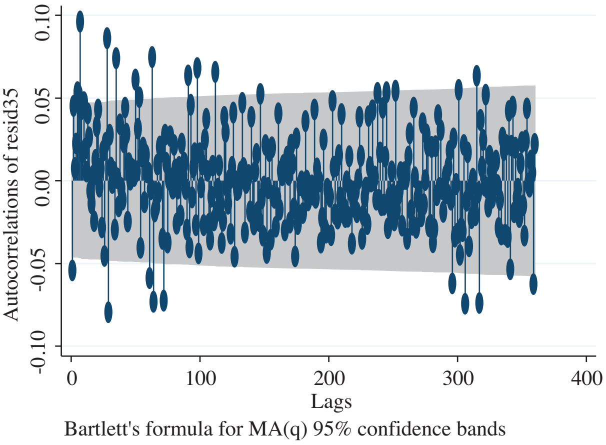

Figure 7. Autocorrelations of final ARMAX model residuals.

Step 5: Analysis of the data.

In this stage, state-of-the-art forecasting techniques such as traditional time series methods and neural network approaches were used to analyze the data. We started with the results of the OLS regression, based on Eq. (11). As mentioned earlier, for all estimation techniques used in this paper, the model was initially trained using data from 1 January 2004 to 31 December 2008; this training sample consisted of 1827 days. The forecasting performance of each model was then evaluated by conducting out-of-sample tests on the 365 days of 2009, called the test sample.

As mentioned above, ANNs consist of interconnections of a number of neurons. In this study, the multilayer perceptron type of neural network model with different learning algorithms was examined. The structure of the proposed model is given in Figure 1. The performance of ANN models depends on some parameters such as the number of nodes (neurons) and layers, the nonlinear function used in the nodes, the learning algorithm, and initial weights of the inputs and output layers. In this study, the neural network model consisted of an input layer, a hidden layer, and an output layer. Since there are 16 input variables, which are the main measures of natural gas consumption, 16 nodes were placed in the input layer. The number of nodes in the hidden layer was decided after testing the different models. The MSE decreased for both the training data and the test data when 9 nodes were used in the hidden layer. After this point, overtraining occurred within the model. The daily natural gas consumption was selected as an output variable.

In the first neural network model, the hyperbolic tangent transfer function was used in the hidden layer, and a linear transfer function was used in the output layer. The Levenberg-Marquardt backpropagation algorithm was used to train the ANN model as the learning algorithm. The optimization algorithm was selected as a conjugate gradient algorithm.

The neural network model in the second method was developed using the NeuroShell Predictor by Ward Systems Group, Inc. The NeuroShell Predictor has 2 different training approaches, neural and genetic. The number of hidden neurons, which determines how well the network can learn patterns, was left at the default number set by the software. The default number of hidden neurons for this network is calculated as $^1/_2$ (number of inputs + number of outputs) + square root of the training patterns. In the neural training approach, the activation function of the hidden layer was used as the Gaussian, hyperbolic tangent, and Gaussian complement functions, which were recommended by the software. NeuroShell's TurboProp training method was employed for the model. This training method trains faster and is not sensitive to learning and momentum rates [17].

Furthermore, this study used an alternative approach to learning that is called the genetic algorithm (GA). The GA has 4 operations: initialization, selection, crossover, and mutation. In the initialization stage, randomly selecting a population, called chromosomes, of possible solutions is the starting point of the search. After the initialization stage, each individual in the initial generation is evaluated using a fitness function to obtain the best solution. This process is called selection. The GA takes 2 fit individuals and mates them (crossover). Individuals may also change through mutation operation. The processes of crossover and mutation are referred to as reproduction. When a GA works well, the population converges to a good solution of the optimization problem. In fact, this solution will be a global optimum or close to a global optimum [16].

The results of OLS regressions are presented in Table 1. With OLS regression, lags 21, 28, and 336 were insignificant. These terms were therefore dropped in the second regression.

The results show that 98% of the variation in the consumption series can be explained by using 9 lags of consumption and 4 independent variables with the OLS specification. Based on t-statistics and in order of importance, the most important variables for forecasting consumption were yesterday's consumption, temperature, squared temperature, number of consumers, price, and consumption levels 1 year ago, 50 weeks ago, 2 weeks ago, 28 weeks ago, and so on.

As expected, both temperature and price had significant negative effects on consumption. The effect of temperature is not linear; instead, it is a concave relationship. As temperature falls, gas consumption increases with a decreasing rate. The fact that price and consumption are inversely related is a natural consequence of the law of demand.

It is an important finding that historical levels of consumption at various lags have great forecasting power for today's consumption. Consumption levels as distant as 52, 51, 50, and 49 weeks ago are especially useful predictors for today's demand. This finding helps decision makers substantially in estimating future demand.

Table 1. Results of OLS regression.

	Dependent variable: consumption on day t		
		(i)	(ii)
$cons_{t-1}$	coeff.	0.500***	0.501***
	t-stat	37.8	38.1
$cons_{t-7}$	coeff.	0.025**	0.027***
	t-stat	2.50	2.71
$cons_{t-14}$	coeff.	0.022**	0.025***
	t-stat	2.22	2.88
$cons_{t-21}$	coeff.	0.004	
	t-stat	0.36	
$cons_{t-28}$	coeff.	0.007	
	t-stat	0.80	
$cons_{t-196}$	coeff.	0.024***	0.023***
	t-stat	2.98	2.85
$cons_{t-280}$	coeff.	0.014**	0.012**
	t-stat	2.46	2.43
$cons_{t-336}$	coeff.	0.001	
	t-stat	0.12	
$cons_{t-343}$	coeff.	0.021**	0.021**
	t-stat	2.15	2.33
$cons_{t-350}$	coeff.	0.036***	0.036***
	t-stat	3.85	3.88
$cons_{t-357}$	coeff.	0.024**	0.025***
	t-stat	2.49	2.63
$cons_{t-364}$	coeff.	0.050***	0.051***
	t-stat	5.14	5.37
$temp_t$	coeff.	–1.25***	–1.25***
	t-stat	–35.5	–35.6
$tempsq_t$	coeff.	0.769***	0.769***
	t-stat	27.3	27.4
$price_t$	coeff.	–0.133***	–0.135***
	t-stat	–7.57	–7.75
$numcons_t$	coeff.	0.094***	0.096***
	t-stat	9.04	9.30
$constant$	coeff.	0.513***	0.516***
	t-stat	36.11	36.88
# of observations		1463	1463
R-squared		0.98	0.98
RMSE		0.0415	0.0417
MAD		0.0601	0.0626
MAPE		0.3687	0.3753

*, **, and *** indicate 10%, 5%, and 1% significance levels, respectively.

RMSE, MAD, and MAPE statistics were calculated from the test sample, which was 2009 data. In other words, they are out-of-sample test statistics.

Table 2. Results of ARMAX regression.

	Dependent variable: consumption on day t	
$cons_{t-1}$	coeff.	0.189***
	z-stat	13.28
$cons_{t-7}$	coeff.	0.054***
	z-stat	4.39
$cons_{t-14}$	coeff.	0.060***
	z-stat	4.90
$cons_{t-21}$	coeff.	0.050
	z-stat	4.28
$cons_{t-28}$	coeff.	0.034***
	z-stat	2.72
$cons_{t-196}$	coeff.	0.041**
	z-stat	2.02
$cons_{t-280}$	coeff.	0.035***
	z-stat	2.88
$cons_{t-336}$	coeff.	0.021*
	z-stat	1.68
$cons_{t-343}$	coeff.	0.031**
	z-stat	2.41
$cons_{t-350}$	coeff.	0.056***
	z-stat	4.39
$cons_{t-357}$	coeff.	0.049***
	z-stat	3.91
$cons_{t-364}$	coeff.	0.099***
	z-stat	8.12
$temp_t$	coeff.	–1.45***
	z-stat	–36.27
$tempsq_t$	coeff.	0.873***
	z-stat	19.14
$price_t$	coeff.	–0.213***
	z-stat	–4.93
$numcons_t$	coeff.	0.122***
	z-stat	4.69
AR(1) coeff. a_1	coeff.	0.69***
	z-stat	34.1
constant	coeff.	0.618***
	z-stat	35.36
# of observations		1463
Log likelihood		3045
RMSE		0.0370
MAD		0.0323
MAPE		0.1984

*, ** and *** indicate 10%, 5%, and 1% significance levels, respectively.

RMSE, MAD, and MAPE statistics were calculated from the test sample, which was 2009 data. In other words, they are out-of-sample test statistics.

The results of the ARMAX model are presented in Table 2. The ARMAX model was based on Eqs. (11) and (12).

As expected, the ARMAX model had much better forecasting performance than the OLS model according to all 3 performance criteria. Based on z-statistics, the order of importance of variables changed with the ARMAX model. Temperature and AR(1) coefficient a_1 became the most important variables. They were followed by squared temperature, yesterday's consumption, price of gas, number of consumers, and consumption 2 weeks ago, last week, 3 weeks ago, and other lags of consumption. All variables were significant at the 5% level, except lag of consumption 336, which was significant at 10%.

From both models, daily temperature stands out as an important determinant because its coefficient is larger than the other significant regressors. The ARMAX model also confirms that the relation between temperature and consumption is nonlinear. This is one reason why ANN models are used as alternatives to time series methods. In the next section, the results of both methods in terms of forecasting performance are compared.

Step 6: Comparison of the models.

In this step, 2 time series models and 3 ANN models were compared in terms of forecasting performance. Forecasting performance was measured using 3 alternative error measures, the RMSE of Eq. (6), MAD of Eq. (7), and MAPE of Eq. (8). Table 3 compares the 5 models in considering the test dataset. The results indicated that the ANN model with an error backpropagation algorithm model outperformed the other forecasting models in terms of RMSE and MAPE, but with respect to MAD performance measurement, ARMAX provided the best results. RMSE and MAPE performance measurement values of the ANN backpropagation method were 0.0347 and 0.1833, respectively, whereas the MAD score for the ANN backpropagation method was 0.0376, slightly higher than that of the ARMAX model. Except for the ANN model with the GA, all of the remaining forecasting models provided similar results.

Table 3. Comparison of models.

Model	RMSE	MAD	MAPE
ARMAX	0.0370	0.0323	0.1984
OLS	0.0415	0.0601	0.3687
NeuroShell TurboProp learning algorithm	0.0550	0.0823	0.4584
NeuroShell GeneHunter learning algorithm	0.0517	0.0971	0.2800
Backpropagation/gradient descent algorithm	0.0347	0.0376	0.1833

Table 4 displays the importance of all 16 variables used in each of the 5 forecasting approaches. As shown, temperature, temperature squared, and the first lag of daily consumption were identified as the first 3 leading criteria of daily gas consumption in the ANN models with backpropagation, TurboProp, the multiple regression model, and the ARMAX model. In the ARMAX model, the AR(1) term represents the shock element in today's consumption level. On the other hand, according to the ANN model with the GA approach, temperature was found to be the most important criterion and natural gas price was identified as the second most important criterion.

Table 4. Independent variable importance.

	OLS	ARMAX	NeuroShell TurboProp algorithm	NeuroShell GeneHunter algorithm	algorithm Backpropagation
	% importance	% importance	% importance	% importance	% importance
L1.ncons	16.9	4.6	12.1	11.1	11.8
L7.ncons	0.9	1.3	3.0	0.4	2.7
L14.ncons	0.8	1.5	0.1	5.4	2.2
L21.ncons	0.0	1.2	4.2	5.7	3.1
L28.ncons	0.0	0.8	3.0	4.6	2.8
L196.ncons	0.8	1.0	0.5	7.5	2.8
L280.ncons	0.4	0.9	0.1	7.9	0.9
L336.ncons	0.0	0.5	1.3	0.8	1.1
L343.ncons	0.7	0.8	0.4	3.8	1.7
L350.ncons	1.2	1.4	1.1	2.1	2.5
L357.ncons	0.8	1.2	3.8	4.9	2.1
L364.ncons	1.7	2.4	4.8	1.1	2.1
ntemp	42.1	35.7	31.6	17.6	37.6
ntempsq	25.9	21.5	27.5	5.0	23.3
subpri	4.5	5.2	0.7	15.0	1.6
nconsumer	3.2	3.0	5.8	7.2	1.7
AR(1)		17.0			

5. Managerial implications and conclusions

This study has dealt with one of the most important issues in natural gas consumption management, providing better decisions for natural gas prediction in İstanbul using appropriate quantitative approaches such as multiple regression, the ARMAX model, and neural networks. This research concluded that the neural network model with backpropagation outperforms multiple regression, neural network NeuroShell, the neural network model with the GA, and the ARMAX model for natural gas forecasting. ARMAX and ANNs with backpropagation models indicated that temperature and the first lag of consumption were the most important factors for natural gas prediction. On the other hand, the suggestion made by an ANN with the GA model, which had the lowest performance scores in terms of MAD, MAPE, and RMSE, contradicted the ARMAX and ANN with backpropagation models. As a result, temperature, natural gas price, and the first lag of consumption were found to have the highest impact on natural gas consumption.

As mentioned above, there are various factors that determine annual gas consumption, including seasonal variation, temperature, new construction of buildings, new agreements with other suppliers, income, and changes in consumer demand. Therefore, it is very difficult to make accurate predictions of İstanbul's natural gas demand. Two of these factors have the highest impact on natural gas consumption in Turkey and specifically in İstanbul. These are temperature and natural gas price. These 2 factors are very important for effective natural gas management for a country like Turkey, which is dependent on foreign sources for 98% of its energy needs [26].

Although consumer income is a very important factor that affects natural gas consumption in the literature, it was not considered in this study. In this paper, we used and forecasted daily natural gas consumption. The unit of time period was a day, and our aim was to forecast daily consumption values in the year 2009 based on historical data. Empirically, it was impossible to find daily GDP data to use in our model. The highest-frequency income data we could obtain was quarterly data. If we had used quarterly data in the model, there would have been no variation in income throughout 3 months or, on average, 91 days. For the training period of 2004-2008, there would have only been 20 different values of income, and for the year 2009, which we were forecasting, there would have only been 4 different values. When an independent variable exhibits very low variation, optimization techniques of maximum likelihood estimation of the ARMAX model run into difficulties and cannot find a maximum of the likelihood function due to the close-nonstationarity of the income variable. In fact, we tried to run the ARIMA model by adding quarterly real GDP values, but the model could not produce a solution due to insufficient variation in the real GDP. Therefore, from a practical point of view, we decided not to include the income variable. In fact, since we were already using a large number of lags of the dependent variable, we assumed that the additional contribution of low-frequency income data would not significantly improve the forecasts.

As is known, the price of natural gas increased by almost 80% in 2008, and this price change still had an effect on natural gas consumption in 2010. According to data from the state-owned Turkish Pipeline Corporation (BOTAŞ), Turkey only consumed 17.7 bcm of natural gas in the period from January 2010 through July 2010. This indicates that Turkey's actual natural gas consumption was less than the expected natural gas consumption demand by approximately 2.5 bcm in the first 7 months of 2010. This suggests that Turkey may have paid a huge bill by the end of 2010 for unused natural gas consumption that may have totaled up to a remarkable $1.5 billion. Turkey paid $704 million and $600 million for unused natural gas in the years 2008 and 2009, respectively.

This study also indicated that although natural gas consumption can be forecasted in the short run, for long-term planning and unexpected externalities, Turkey needs gas storage facilities to save money when actual natural gas consumption in İstanbul is less than the predicted consumption. This will reduce İstanbul's and, on a wider scale, Turkey's natural gas bill against the adverse effects of agreements with suppliers that are based on estimated total volume gas consumption.

References

[1] M.A. Kiliç, "Turkey's natural gas necessity, consumption and future perspectives", Energy Policy, Vol. 34, pp. 1928-1934, 2006.

[2] M. Tunç, Ü. Çamdali, C. Parmaksizoğlu, "Comparison of Turkey's electrical energy consumption and production with some European countries and optimization of future electrical power supply investments in Turkey", Energy Policy, Vol. 34, pp. 50-59, 2006.

[3] T. Çetin, F. Oguz, "The reform in the Turkish natural gas market: a critical evaluation", Energy Policy, Vol. 35, pp. 3856-3867, 2007.

[4] International Energy Agency, Energy Policies of IEA Countries - Turkey- 2009 Review, 2010. Available at http://www.iea.org/publications/free_new_Desc.asp?PUBS_ID=2276.

[5] M.A. Ozgur, "Review of Turkey's renewable energy potential", Renewable Energy, Vol. 33, pp. 2345-2356, 2008.

[6] I. Yüksel, "Energy production and sustainable energy policies in Turkey", Renewable Energy, Vol. 35, pp. 1469-1476, 2010.

[7] International Energy Agency, World Energy Outlook 2010, 2010. Available at http://www.iea.org/Textbase/npsum/weo2010sum.pdf.

[8] E. Erdogdu, "A review of Turkish natural gas distribution market", Renewable and Sustainable Energy Reviews, Vol. 14, pp. 806-813, 2010.

[9] StataCorp, Stata Statistical Software: Release 11, College Station, Texas, StataCorp LP, 2009.

[10] N. Türker, F. Güneş, T. Yıldırım, "Artificial neural design of microstrip antennas", Turkish Journal of Electrical Engineering and Computer Sciences, Vol. 14, pp. 445-453, 2006.

[11] M. Carcenac, "An implicit surface modeling technique based on a modular neural network architecture", Turkish Journal of Electrical Engineering and Computer Sciences, Vol. 12, pp. 11-26, 2004.

[12] İ. Dalkıran, K. Danışman, "Artificial neural network based chaotic generator for cryptology", Turkish Journal of Electrical Engineering and Computer Sciences, Vol. 18, pp. 225-240, 2010.

[13] S. Haykin, Neural Networks: A Comprehensive Foundation, New Jersey, Prentice Hall, 1998.

[14] R. Kizilaslan, B. Karlik, "Combination of neural networks forecasters for monthly natural gas consumption prediction", Neural Network World, Vol. 19, pp. 191-199, 2009.

[15] B. Karlık, "Differentiating type of muscle movement via AR modeling and neural network classification", Turkish Journal of Electrical Engineering and Computer Sciences, Vol. 7, pp. 45-52, 1999.

[16] M. Krishnaswamy, P. Krishnan, "PM – Power and Machinery: nozzle wear rate prediction using regression and neural network", Biosystems Engineering, Vol. 82, pp. 53-64, 2002.

[17] Ward Systems Group, NeuroShell Easy Predictor Instructions, Frederick, Maryland, Ward Systems Group Inc., 2008.

[18] M. Khashei, M. Bijari, "An artificial neural network (p, d, q) model for timeseries forecasting", Expert Systems with Application, Vol. 37, pp. 479-489, 2010.

[19] G.E.P. Box, G.M. Jenkins, Time Series Analysis: Forecasting and Control, San Francisco, Holden-Day, 1970.

[20] G.E.P. Box, G.M. Jenkins, G.C. Reinsel, Time Series Analysis: Forecasting and Control, 3rd ed., Englewood Cliffs, New Jersey, Prentice Hall, 1994.

[21] J.H. Cochrane, Time Series for Macroeconomics and Finance [online]. Available at http://faculty.chicagobooth.edu/john.cochrane/research/papers/time_series_book.pdf.

[22] C. Granger, P. Newbold. "Spurious regressions in econometrics", Journal of Econometrics, Vol. 2, pp. 111-120, 1974.

[23] W. Enders, Applied Econometric Time Series. New York, Wiley, 1995.

[24] E. Erdogdu, "Natural gas demand in Turkey", Applied Energy, Vol. 87, pp. 211-219, 2010.

[25] B.H. Baltagi, Econometrics, 4th ed., Berlin, Springer-Verlag, 2008.

[26] H. Zaim, "Knowledge management implementation in IZGAZ", Journal of Economic and Social Research, Vol. 8, pp. 1-25, 2006.