ISSN 1330-3651
UDC/UDK 665.612 : 004.032.26

PREDICTING NATURAL GAS CONSUMPTION BY NEURAL NETWORKS

Zlatko Tonković, Marijana Zekić-Sušac, Marija Somolanji

Preliminary notes

The aim of the paper is to create a prediction model of natural gas consumption on a regional level by using neural networks, and to analyze the results in order to improve prediction accuracy in further research. The output variable consisted of the next-day gas consumption in hourly intervals, while the input space included previous-day consumption in addition to exogenous variables. After conducting a feature selection procedure, two neural network algorithms were trained and tested: the multilayer perceptron and the radial basis function network with different activation functions. The dataset consisted of real historical data of a Croatian gas distributor. The best neural network model is selected on the basis of the mean absolute percentage error obtained on the test sample. The results were analyzed, and some critical hours and days were identified. Guidelines were reported that could be valuable to both researchers and practitioners in this area.

Keywords: *natural gas consumption, neural networks, multilayer perceptron, radial basis function network, fuzzy variable*

Predviđanje potrošnje prirodnog plina pomoću neuronskih mreža

Prethodno priopćenje

Cilj rada je kreirati prediktivni model potrošnje prirodnog plina na regionalnoj razini koristeći neuronske mreže, kao i analizirati rezultate s ciljem unapređivanja točnosti predviđanja u budućim istraživanjima. Izlazna varijabla sastojala se od potrošnje prirodnog plina sljedećeg dana u satnim intervalima, dok je ulazni prostor varijabli uključivao potrošnju prethodnog dana, te dodatne egzogene varijable. Nakon procedure selekcije značajnih varijabli, trenirana i testirana su dva algoritma neuronskih mreža: višeslojni perceptron i mreža s radijalnom funkcijom koristeći različite aktivacijske funkcije. Skup podataka sastojao se od stvarnih povijesnih podataka jednog hrvatskog distributera plina. Na temelju srednje apsolutne postotne greške dobivene na testnom uzorku izabran je najbolji model neuronske mreže. Rezultati su analizirani, i identificirani su kritični sati i dani. Iznesene su određene smjernice koje mogu biti korisne za istraživače i praktičare u ovom području.

Ključne riječi: *potrošnja prirodnog plina, neuronske mreže, višeslojni perceptron, mreža s radijalno zasnovanom funkcijom, fuzzy varijabla*

1 Uvod Introduction

Natural gas is one of the mostly used energy resources and its consumption is increasing daily considering economic and social development, as well as civilized life in general. The proportion of natural gas in total world energy consumption has been constantly increasing during the last 30 years and currently reaches around 25 % of primary energy, while in some countries, such as Russia, exceeds 50 % [1]. The growth trend of natural gas consumption is 1 to 3 % per year [1], which is partially caused by intense exhaustion, unstable oil market and a very high demand to reduce environmental pollution. Natural gas is a product of nature in its purest form. The main constituent of natural gas is methane (80 - 98 %), the share of nitrogen is negligible, and the rest are mainly hydrocarbons. Since the share of carbon in natural gas is significantly lower than the share of hydrogen, natural gas after burning - compared with other forms of energy - pollutes environment the least, therefore is an important energy resource from the ecological point of view. The distribution of natural gas in each country is subjected to a license issued by the appropriate ministry. For example, there are 38 licensed natural gas distributors in Croatia, and all of them purchase natural gas from a common national supplier.

The predictions of natural gas consumption are being conducted on different geographical levels: national, regional, and local. This paper is focused on creating a regional prediction model that could be used by distributors, but can be easily adjusted to suppliers' needs. An accurate prediction of gas consumption is needed due to the fact that distributors are required (by their suppliers) to nominate the amount of natural gas they will need for the next day. There is a regulated tolerance interval and a penalty system if the real consumption exceeds the tolerance interval of a nominated amount. The purchase of non-sufficient amount of gas implies unsatisfied customers, which is not acceptable due to some big issues it could cause, such as customer complaints, cancelation of business agreement with a distributor, and other economic and social issues. Since an over purchase implies large costs for a distributor due to the penalty system, an accurate prediction is important due to financial and customer satisfaction reasons.

The objectives of this paper are (1) to create an efficient neural network model on a Croatian dataset that will predict the next-day hourly consumption of natural gas based on the previous-day hourly consumption and a set of exogenous variables as predictors, and (2) to analyze the model behavior at different days in the test set in order to find some guidelines for future model redesign that will improve the accuracy of prediction. Exogenous variables included meteorological data such as temperature prognoses for every 6 hours of the next day, wind velocity and direction prognoses for every 6 hours of the next day, *Day type* (working day, holiday, and day after holiday), *Day of the week*, and *Season* detection in the form of a fuzzy variable. Two neural network algorithms were tested: the multilayer perceptron and the radial basis function network. The experiment was conducted for a gas distribution metering and control unit (MCU) of a Croatian gas distributor in the north-east region of Croatia. The datasets covered daily observations from January 01, 2008 to March 31, 2009. The best neural network model was identified for each of the distribution points, and the results were analyzed in the sense of the prediction accuracy and the selection of important predictors. The a-posteriori analysis revealed

some guidelines for possible future research that will improve the accuracy of prediction.

The paper is structured as following: the next section provides a review of previous research in this area, the methodology of neural networks used in this research is described in the third section. Section 4 presents data and modeling strategy, while the results and conclusion with guidelines for further research are given at the end of the paper.

2
Review of previous research
Pregled dosadašnjih istraživanja

Analysis of previous research in the area of energy consumption (gas or electricity) reveals that various deterministic and stochastic models have been applied to describe and forecast the natural-gas consumption [2]. The authors mostly used statistical forecasting models, such as autoregressive moving average (ARMA), cycle analysis, or multiple regression, while recent papers show that neural networks (NNs), as a nonparametric and nonlinear method, produce a successful prediction and have some advantages over statistical methods [3]. Past load and weather data were generally used as the network inputs, while forecasted load values represented the outputs.

Darbellay and Slama [3] forecasted short term demand for electricity in Czech Republic by using neural networks and ARMA model. In the data-preprocessing phase, they suggest a three-stage procedure to decide whether the problem is nonlinear or not. First, they propose a nonlinear measure of statistical dependence, then they analyse the linear and the nonlinear autocorrelation functions of the electric consumption, and third, they compare the predictions of nonlinear models (artificial NNs) with linear models (of the ARMA type). They found that forecasting the short-term evolution of the Czech electric load is primarily a linear problem. However, by comparing the predictions of nonlinear models (artificial neural networks) with linear models (ARMA), it was shown that although neural networks do not outperform linear ARMA models in sense of prediction accuracy, there are certain conditions under which neural networks could be superior to linear models. Neural networks do not require differencing in input data, and are able to integrate more information and thus produce better forecasts, for example with hourly load data and daily temperatures. They used the normalized mean square error (NMSE) and the mean absolute percentage error (MAPE) as measures of model successfulness. NMSE ranged from 0,8 to 3,8 % and MAPE ranged from 1,0 to 3,0 % by ARIMA, and similar by NNs. The models were created for the horizons of 1, 12, 24, and 36 hours.

Beccali et al. [4] predicted daily electric load of a suburban area of the town Palermo in Italy. They used a self-organized Kohonen unsupervised network in a preprocessing phase in order to identify clusters of data, and then created a two layered feed forward neural network for prediction purposes, trained with the back propagation algorithm. The input variables included 24-hours weather data (hourly dry bulb temperatures, relative humidity, global solar radiation) along with historical load data available from 2001 to 2003. The output consisted of 24 units representing hourly simultaneous load forecast for the day concerned. The authors do not report the average error on the test data set, but only the percentage error obtained for one forecasting day, which was in average 1,97 %, while the maximum error was 3,81 %. Their paper shows that neural networks have potential to serve as a useful instrument in tackling short-term load forecasting problems and could become a precious decision supporting tool in energy planning, especially for periods in which the influence of weather conditions on electric consumption is certainly overwhelming, but very difficult to evaluate precisely.

Thaler et al. [5] used the radial basis neural network algorithm to build a model for energy consumption in the gas distribution system in Slovenia. Prediction is performed by a conditional average estimator based upon known prototype patterns and given future values of environmental variables. They used genetic algorithms to determine the relevance of these variables. Prediction error amounts to a few percents of the actual consumption. Besides calculating the prediction error, the authors estimated the probability distribution of prediction for the one-day time interval. This distribution can be used to estimate the risk of energy demand beyond a certain prescribed value. They also propose a cost function that includes operation and control costs of a distribution system as well as penalties related to excess energy demand. The probability that actually observed consumption will surpass a certain prescribed value of a maximal allowed consumption (MAC) at each predicted value is also suggested describing the risk of excess energy demand by clients. Gelo [6] investigated a multivariate model of monthly gas consumption of residential customers in Croatia. The author used the total and average monthly consumption as output variables, while the input variables were the average monthly temperature, the price of natural gas for residents, and the average salary of residents. The multivariate regression analysis was used as a methodological basis, and it was found that natural gas consumption primarily depends on the average monthly temperature, while the impact of other input predictors is less significant. It was also shown that the model that uses the average monthly gas consumption at its output is better fitted with the data than the model that uses the total monthly gas consumption.

Potocnik et al. [7] use a statistics-based machine forecasting model to predict future consumption of natural gas in Slovenia in 2005 and 2006. They use previous consumption, past weather data, weather forecast, and some additional parameters, such as seasonal effects and nominations as input variables. At the output, they predicted the future daily and weekly gas consumption. Their model produced the average percentage error of 2,8 % on the whole sample. The analysis also showed a strong correlation among the error of the temperature forecast and the error of the output variable. In their newer paper, Potocnik et al. [8] investigated a short-term forecasting model of Slovenian gas consumption more deeply, and used the daily forecast for the next gas consumption day, delivered in hourly time intervals. They analyzed gas consumption cycles (yearly, weekly, daily) and revealed some patterns which were incorporated into the forecasting model. The following forecasting model is proposed: (1) two separate sub models for the winter and summer seasons; (2) input variables including past consumption data, weather data, weather forecasts and basic cycle indexes; (3) a hierarchical forecasting structure: a daily model as the basis, with the hourly forecast obtained by modeling the relative daily profile. Their approach showed that the model fitted well with the data, and obtained a validation error of

0,026 on Slovenian dataset. They also extended their research on using forecasting errors to build an economic model, which defines critical forecasting error levels that can turn the cash flow into a positive or a negative one. The economic incentive model and the forecasting model served as a basis for a risk model that estimated the risk associated with critical error levels. The risk model analyzed the forecasting error in the context of various influential parameters, such as seasonal data, month, day of the week, and temperature [9].

It can be concluded from the previous research that authors mainly used statistical regression analysis, statistical forecast methods, and neural networks as methodological basis for predicting and forecasting natural gas consumption. Models created for different countries and regions vary according to the selection of input variables, time horizons observed, and the accuracy of prediction. Since meteorological data, such as temperature forecast, as well as previous consumption, were found to be significant predictors of gas consumption in most papers, those findings were used as guidelines in this paper to create a model on a Croatian dataset that will incorporate some previous findings and add some new information to gas consumption modeling.

3
Neural network methodology
Metodologija neuronskih mreža

Artificial NNs were used in this research due to their numerous advantages, such as nonlinearity, adaptivness, and high degree of robustness [10]. As a non-parametric method, NNs have the ability to overcome the proportionality and linearity constraints imposed by parametric methods in prognoses [10]. Potocnik et al. [9] showed that NNs as a non-parametric method have the ability to perform at least as well as statistical methods in predicting gas consumption. However, the lack of standardized paradigms that can determine the efficiency of certain NN algorithms and architectures in particular problem domains is emphasized by many authors [11]. Therefore, we test and compare the performance of two NN algorithms: multilayer perceptron and the radial basis function.

MLP is a general purpose feed forward network, and one of the most frequently used NN algorithms. In order to optimize the error function it uses the classical back propagation algorithm based on deterministic gradient descent algorithm originally developed by Paul Werbos in 1974, extended by Rumelhart, Hinton, Williams (in [12]). Since the main disadvantage of the back propagation algorithm is the danger of local minima, the conjugate gradient algorithm is also tested in order to overcome this limitation [12]. Conjugate gradient is combined with the classical back propagation such that back propagation is used in first 100 epochs, while the conjugate gradient is used in the next 500 epochs. The standard delta rule was used for learning; the learning rate was dynamically optimized during the learning process (ranged from 0,08 to 0,01, while the momentum ranged from 0,3 to 0,1). Overtraining is avoided by a cross-validation process which alternatively trains and tests the network (using a separate test sample) until the performance of the network on the test sample does not improve for n number of attempts (n=10). The maximum number of epochs in our experiments was set to 500. The optimal number of hidden units is determined in a cascading procedure, which gradually increases the number of hidden units during the training phase, starting from a minimum to a maximum value. The minimum value was set to 1 and the maximum was set to 50 in all experiments with the multilayer perceptron network. The MLP algorithm was tested by using two activation functions in its hidden layer: sigmoid and hyperbolic tangent. The results of the models using each of the activation functions were reported. After the best network is selected, it is tested on a new validation sample to determine its generalization ability.

RBFN is based on a clustering procedure for computing distances between each input vector and a center, represented by the radial unit. The ability of RBFN with one hidden layer to approximate any nonlinear function is proved by Park and Sandberg (in [13]). Michelli (in [13]) showed how this network can produce an interpolating surface which passes through all the pairs of the training set. RBFN algorithm uses Euclidean distance and Gaussian transfer function in the hidden (or pattern) layer which maps the output of the distance function according to formula:

$$f(x) = \varphi \cdot \left(\| x - c \| \right) = \mathrm{e}^{\left(\frac{I_k^2}{\sigma_k^2} \right)} \quad (1)$$

where x is an input vector, c is the center determined by a clustering algorithm, and parameter σ is determined by the nearest neighbor technique. The number of hidden units in our experiments is set to approximately one-half of the size of the training sample (160).

In order to describe the topology of neural network models, it is important to notice that the period of one gas consumption day in the dataset covered the 24-hour time interval from 6,00 am (day i) to 6,00 am (day i+1). Due to current technical conditions, Croatian gas distributors usually have available data on real gas consumption of the preceding day (day i-1), and are obliged to predict the gas consumption of the next day (day i+1). Therefore, the term "forecast day" in our models denotes day i+1, while the data for the current day (day i) will not be available at the expected time of the model usage. The output variable consisted of the 24-hour gas consumption of the forecast day, and can be noted as $Y(i+1, t_j)$, where i is the observed day, t_j is the hour within the observed day, and j=1,2,...,24. The input layer of the NN architecture consisted of 43 input variables including the hourly *gas consumption* of the preceding day $Y(i-1, t_j)$, *Day type* of the forecast day (i+1), *Day of the week* of the forecast day (i+1), *Month*, *Season detection* as fuzzy variable of the forecast day (i+1), and meteorological forecast data of the day (i+1): *temperature*, *wind direction*, and *wind velocity*, available for every 6 hours (k=1,7,13,18,24). The output layer consisted of 24 neurons representing hourly gas consumption for each hour of the forecast day (day i+1) from 6,00 am to 6,00 am. Categorical input variables *Day of the week*, *Day type*, and *Month* were represented binary in the NN model, while all the other input variables were continuous. Variable values were normalized before running NN algorithms. The number of hidden units in MLP networks varied from 1 to 50, and was optimized in a cascading procedure during the network training phase. The network topology (built according to the slightly modified approach of Beccali et al. [4] which gave similar NN architecture for daily urban

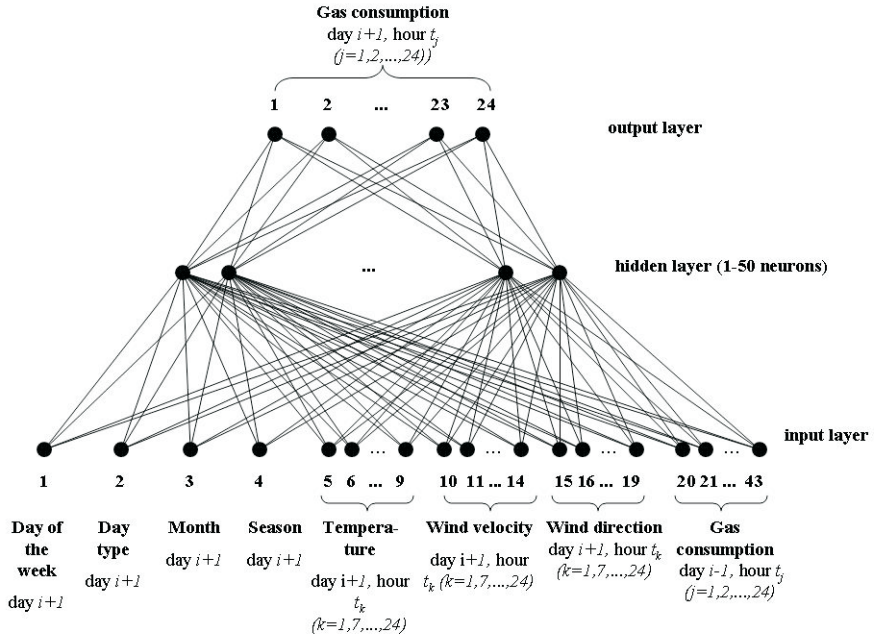

Figure 1 *Topology of neural network models*
Slika 1. *Topologija modela neuronske mreže*

electric load forecast, but with different input and output variables) is presented in Figure 1.

The modeling strategy of neural networks included varying the NN algorithm, activation function in the hidden layer, and network topology. The accuracy of prediction was measured by the NMSE and MAPE error. Although the NMSE and MSE are the standard measures of performance for all regressive-type NN algorithms, some others suggest using measures that will be more useful from the practitioner's point of view [12], such as MAPE, also used by [3]. MAPE is computed by the formula:

$$MAPE = \frac{1}{N} \cdot \sum_{t=1}^{N} \frac{|x_t - \hat{x}_t|}{\hat{x}_t} \qquad (2)$$

where N is the number of observations in the test set, t is the time, x_t is the output computed by the neural network model in the observed time t, while $\hat{x}_t$ is the real output in time t. In our experiments, MAPE is computed for each hourly interval in the test set, and the average daily MAPE (as average of hourly MAPEs) is computed on the test set and used as the criterion for selecting the best NN model. The best NN model is further analyzed regarding the hourly variance. NN models were run by using the Statistica 8.0 software.

4
Data description and preprocessing
Opis i priprema podataka
4.1
Data sample and variables
Uzorak podataka i varijable

The experiments were conducted on a dataset containing real historical data of gas consumption at a distribution point of a Croatian second-large natural gas distributor, as well as historical data of exogenous variables. The distributor used in this research supplies natural gas to 66 486 customers in the north-east region of Croatia through a network of gas pipelines of 2 133 km, delivering annually more than 166 million cubic meters of natural gas. Customers are divided into two basic groups: residential (who use natural gas for their own needs in their homes) with its share of 54 % in total natural gas consumption and commercial (any customer other than residential) with its share of 46 % in total natural gas consumption [14].

The hourly data were available for the period from January 01, 2008 to March 3, 2009. Due to the fact that the distributor is obliged to nominate hourly gas consumption for the next day from 6,00 am to 6,00 am (day i+1), the basic time interval used for the observation was one gas consumption day, while the output and some input variables were observed in hourly intervals within a day. Therefore, the size of the total dataset was 454 days. All 43 input variables together with their statistical descriptive analysis were presented in Table 1.

It was assumed that all input variables have a certain impact to the output. *Month* was included as an input variable in order to analyze the difference of gas consumption on each month. The reason for using the *Day of the week* as the input variable was to investigate the difference in gas consumption at each day of the week, and separately on holidays. The variable *Day type* further emphasizes the difference among working days, weekends, and working days that follow immediately after weekends. It was previously reported [9] that *temperature* is a relevent variable for gas consumption. Other input variables, such as *wind velocity* and *wind direction*, were used as available meteorological data for which it was assumed that also have some impact to the output. In order to analyze seasonal effect, the graph of hourly gas consumption during the whole observed period is presented in Figure 2.

Figure 2 shows fluctuations in gas consumption through the whole observed period, and a very high increase of consumption at the beginning of September 2008. Due to the fact that the NNs do not require to remove seasonality in data before modeling as statistical forecasting methods do, data preprocessing phase in this experiment did not include

Table 1 *Input variables and their descriptive statistics*
Tablica 1. *Ulazne varijable i njihova opisna statistika*

Variable no.	Variable description and coding	Descriptive statistics
1	Month (1-12)	1=13,66 %, 2=12,56 %, 3=13,22 %, 4=6,61 %, 5=6,83 %, 6=6,61 %, 7=6,83 %, 8=6,83 %, 9=6,61 %, 10=6,83 %, 11=6,61 %, 12=6,83 %
2	Season detection (fuzzy variable)	mean=0,655, stdev=0,390
3	Day type (1="working day", 2="holiday", 3="day after holiday"	1=52,86 %, 2=31,06 %, 3=16,08 %
4	Day of the week (1="Monday", 2="Tuesday", 3="Wednesday", 4="Thursday", 5="Friday", 6="Saturday", 7="Sunday", 8="Holiday")	1=13,88 %, 2=13,88 %, 3=13,66 %, 4=13,88 %, 5=14,10 %, 6=14,32 %, 7=13,66 %, 8=2,64 %
5	Temperature at 6:00 of day *i*+1	mean=6,330, stdev=7,866
6	Temperature at 12:00 of day *i*+1	mean=13,482, stdev=10,281
7	Temperature at 18:00 of day *i*+1	mean=12,307, stdev=10,331
8	Temperature at 0:00 of day *i*+1	mean=7,646, stdev=7,919
9	Temperature at 6:00 of day *i*+1	mean=6,362, stdev=7,828
10	Wind direction at 6:00 of day *i*+1	mean=18,295, stdev=8,542
11	Wind direction at 12:00 of day *i*+1	mean=18,485, stdev=8,633
12	Wind direction at 18:00 of day *i*+1	mean=16,794, stdev=9,235
13	Wind direction at 0:00 of day *i*+1	mean=18,573, stdev=8,607
14	Wind direction at 6:00 of day *i*+1	mean=18,295, stdev=8,542
15	Wind velocity at 6:00 of day *i*+1	mean=1,916, stdev=1,139
16	Wind velocity at 12:00 of day *i*+1	mean=2,707, stdev=1,463
17	Wind velocity at 18:00 of day *i*+1	mean=2,108, stdev=1,110
18	Wind velocity at 0:00 of day *i*+1	mean=1,907, stdev=1,084
19	Wind velocity at 6:00 of day *i*+1	mean=1,918, stdev=1,138
20	Gas consumption at 6:00 of day *i*-1	mean=2285,143, stdev=769,983
21	Gas consumption at 7:00 of day *i*-1	mean=2717,115, stdev=950,272
22	Gas consumption at 8:00 of day *i*-1	mean=3063,764, stdev=1126,436
23	Gas consumption at 9:00 of day *i*-1	mean=3168,852, stdev=1212,430
24	Gas consumption at 10:00 of day *i*-1	mean=3138,530, stdev=1187,399
25	Gas consumption at 11:00 of day *i*-1	mean=3109,024, stdev=1139,841
26	Gas consumption at 12:00 of day *i*-1	mean=3080,543, stdev=1109,643
27	Gas consumption at 13:00 of day *i*-1	mean=3013,839, stdev=1098,713
28	Gas consumption at 14:00 of day *i*-1	mean=2929,084, stdev=1121,653
29	Gas consumption at 15:00 of day *i*-1	mean=2870,720, stdev=1147,543
30	Gas consumption at 16:00 of day *i*-1	mean=2886,252, stdev=1189,814
31	Gas consumption at 17:00 of day *i*-1	mean=2940,413, stdev=1243,063
32	Gas consumption at 18:00 of day *i*-1	mean=2974,128, stdev=1244,229
33	Gas consumption at 19:00 of day *i*-1	mean=2996,845, stdev=1228,100
34	Gas consumption at 20:00 of day *i*-1	mean=3027,925, stdev=1195,966
35	Gas consumption at 21:00 of day *i*-1	mean=3018,804, stdev=1111,676
36	Gas consumption at 22:00 of day *i*-1	mean=2841,071, stdev=968,002
37	Gas consumption at 23:00 of day *i*-1	mean=2508,455, stdev=789,529
38	Gas consumption at 00:00 of day *i*	mean=2226,700, stdev=670,892
39	Gas consumption at 01:00 of day *i*	mean=2030,552, stdev=627,731
40	Gas consumption at 02:00 of day *i*	mean=1955,219, stdev=621,325
41	Gas consumption at 03:00 of day *i*	mean=1906,989, stdev=613,909
42	Gas consumption at 04:00 of day *i*	mean=1909,389, stdev=632,972
43	Gas consumption at 05:00 of day *i*	mean=2006,079, stdev=668,463

data transformation in the form of moving averages. However, in order to include season detection to the neural network model, a fuzzy input variable is created with fuzzy values representing whether or not a certain day falls into the season of large gas consumption. Since the dataset contained the time period of only one year, the peeks and zero-points of the seasonal variable were determined by a supplier's research, which provided the information about the dates of the highest gas consumption, and the dates of the lowest gas consumption. By that previous research, it was determined that the top of the season is in months November, December, January and February, while the lowest consumption is in July and August. Therefore, fuzzy values in these months were 0 and 1 accordingly. The fuzzy values in other months of observation were calculated according to the linear function. Degrees of membership in interval (0,1) were obtained for these observations. The graph of fuzzy membership function of variable *Season* is presented in Figure 3.

Although neural network methodology does not require input variables to be mutually independent as the multiple regression does, Pearson correlation coefficients were computed in order to investigate possible linear relationships among variables. As expected, the variable *Month* is very highly correlated to *temperature* (at each hour). There is a high correlation (r>0,7) between *Season*

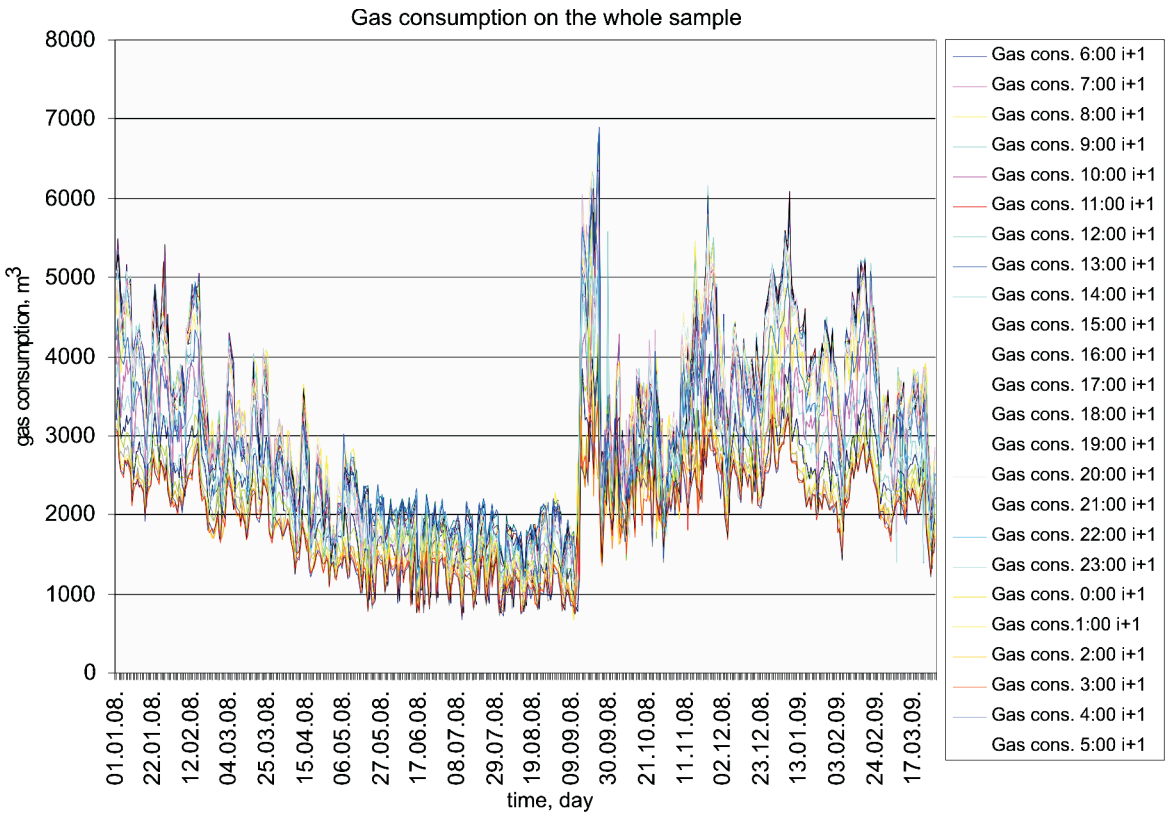

Figure 2 *Hourly gas consumption for the whole observed period*
Slika 2. *Satna potrošnja prirodnog plina za čitavo promatrano razdoblje*

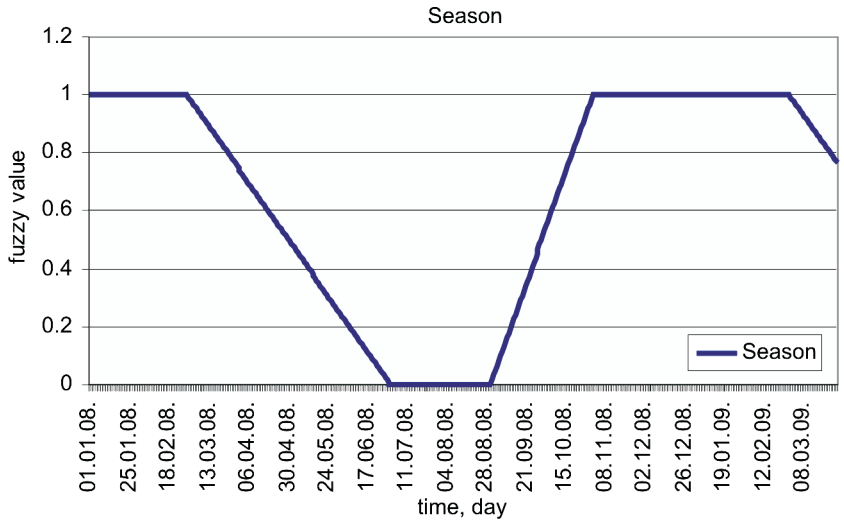

Figure 3 *Graph of fuzzy membership function for the variable Season*
Slika 3. *Graf fuzzy funkcije za varijablu Season*

and *Temperature*, as well as between *Season* and *Previous-day gas consumption*. *Type of the day* and *Day of the week* are mutually highly correlated, but are not correlated to any other input variable. Correlation coefficients also show a high relationship among *Temperature* and *Previous-day gas consumption* (r>0,7). Other input variables are not mutually significantly correlated.

4.2
Variable selection
Odabir varijabli

In cases with a large number of candidate predictors one of the variable selection procedures is usually conducted in order to reduce the number of variables. Building models with smaller number of variables reduces the time of data collection and computation during the model usage. Some of those procedures could be applied before running NN, such as correlation analysis, and principal component analysis (PCA) (Becalli), while the others are usually applied during the process of running NN, such as forward or backward selection procedures, pruning input units with low weights, sensitivity-based pruning after the training process, and genetic algorithms [15]. For regression-type problems, variables could be selected based on standard correlation coefficients, which measure linear relationships among variables. Since NNs do not assume linear or even monotone relationship between the predictors and the dependent variable, a more general detection of relationship is required. Procedure implemented in this paper is based

on the ratio of the between category variance to within category variance (of the dependent variable) for intervals of predictor variables determined depending on their nature (continuous vs. categorical). For continuous predictors, the procedure divides the range of values in each predictor into k intervals (k=10 in our experiments) to "fine-tune" the sensitivity of the algorithm to different types of monotone and/or non-monotone relationships. The importance of each predictor is represented by the F and p value.

The feature selection procedure conducted on the initial set of 43 input variables presented in Table 1 revealed that F and p-values show a significant impact (F>3,84 and p<0,01) for all input variables except *Day type* and *Day of the week*. These results served as a guideline for modeling strategy of NNs, which included (1) testing models with all available variables, and (2) models without two variables that were found non-significant for the output.

4.3
Sampling procedure for the neural network model
Postupak podjele uzorka u modelu neuronske mreže

For the purpose of neural network training, cross-validation and final testing, the whole sample was divided into three subsamples, such that the training set contained 70 % of data, the cross-validation set consisted of 10 % of data, while the rest or 20 % of data was used for final testing. In time series prediction, random sampling is not appropriate, since the data should be trained on earlier period, and tested on newer period of time [12]. The number and percentage of cases in each subsample, as well as the time period covered are presented in Table 2.

Table 2 *Sampling procedure used in neural network modeling*
Tablica 2. *Uzorkovanje korišteno kod modeliranja neuronskih mreža*

Subsample	Number of cases	%	Time interval
Train	318	70	January 01, 2008 – November 13, 2008
Cross-validation	45	10	November 14, 2008 – December 28, 2008
Test	91	20	December 28, 2008 – March 29, 2009
Total	454	100	January 01, 2008 – March 29, 2009

While the train and cross-validation samples were used during the training phase of the NN models, the test sample was left out at this stage of modeling and used only for final testing of all NN architectures in order to evaluate and compare the efficiency of NN models, i.e. for best model selection.

5
Results
Rezultati
5.1.
Selecting the best neural network model
Izbor najboljeg modela neuronske mreže

In order to find the most successful NN model for predicting gas consumption, the modeling strategy included following steps:

(1) Model 1 - train and test the model with all available input space:
  (a) by varying NN algorithm (multilayer perceptron and radial-basis function),
  (b) by varying activation function (logistic and tangent hyperbolic),
(2) Model 2 - train and test the model without predictor variables that were not found significant in the feature selection procedure - steps (1a) and (1b) repeated here,
(3) Select the best model based on MAPE computed on the holdout test set,
(4) Analyze the MAPE of daily and hourly time intervals to find guidelines for further model redesign.

All NN architectures conducted in steps 1 and 2 were trained, cross-validated (in order to find the best number of hidden units and learning time), and finally tested on the hold-out test set.

The NMSE and MAPE performance measures obtained on the test set were reported and compared.

When all 43 available variables were included in the NN (Model 1), the hourly MAPEs are produced for each hour of all gas consumption days included in the test sample, and average daily MAPE and NMSE are also reported. The results are presented in Table 3.

Table 3 *Neural network test results of Model 1 on the test sample*
Tablica 3. *Rezultati neuronskih mreža za Model 1 na testnom uzorku*

Hour	MLP NN, logistic activation function	MLP NN, tanh activation function	RBF network, Gaussian activation function
6:00	12,0973	11,2636	22,9579
7:00	12,5118	12,4286	25,5848
8:00	11,1834	11,3383	27,3683
9:00	9,1101	8,4676	29,4671
10:00	8,5540	8,4709	29,8481
11:00	9,0818	9,4317	29,2761
12:00	11,1070	12,1880	29,3381
13:00	11,5944	12,8454	30,0115
14:00	12,1193	13,1727	31,3232
15:00	12,4877	13,3626**	32,4046
16:00	11,5876	11,8703	32,5505
17:00	9,8240	10,0161	32,5873
18:00	10,3025	10,2994	33,3162
19:00	9,2519	9,3142	32,1785
20:00	8,1418	8,2792	30,2611
21:00	8,8785	8,4168	28,0602
22:00	8,2614	8,0982	25,8514
23:00	8,0547	7,4144	23,5560
0:00	7,6743	7,2504*	22,8446
1:00	8,7501	7,9557	24,2736
2:00	9,4273	8,5434	23,9108
3:00	10,0642	8,8725	23,8196
4:00	10,6012	9,9531	23,9530
5:00	11,1856	10,8187	23,2249
Average MAPE	10,0772	10,0030	27,8320
NMSE	0,0877	0,0865	0,7717
* - minimum hourly MAPE of the best NN model ** - maximum hourly MAPE of the best NN model			

It can be seen from Table 3 that MLP algorithm using tangent hyperbolic activation function produced the lowest MAPE of 10,00 %, while the NMSE of this architecture was

0,0865. The network was trained with the maximum number of 50 hidden units, and in a cascading procedure resulted with 14 final hidden units. When the same algorithm is trained by the logistic function, the average MAPE was slightly higher (10,08 %), as well as the NMSE which is 0,0877. The highest NMSE and MAPE (i.e. the worst performance of Model 1) was obtained by the radial-basis function network (MAPE=27,83 %, NMSE=0,7717) which was trained with 160 neurons in its pattern. When hourly MAPEs of the best network in Table 3 (MLP network with tangent hyperbolic function) are compared, it can be noticed that the minimum MAPE is obtained for gas consumption at 0:00 (MAPE is 7,25 %), while the same network has the lowest accuracy at 15:00. It is interesting to observe that the other two NN architectures in Table 3 are also the most accurate at 0:00, while the hour with maximum MAPE of those two architectures vary.

In case of Model 2 – when the variable *Day of the week* and *Day type* were not used due to their insignificant *F* and *p* values obtained by the feature selection procedure - the NN architectures produced the results presented in Table 4.

Table 4 *Neural network test results of Model 2 on the test sample*
Tablica 4. *Rezultati neuronskih mreža za Model 2 na testnom uzorku*

Hour	MLP NN, logistic activation function	MLP NN, tanh activation function	RBF network, Gaussian activation function
6:00	12,2533	8,9352	12,8707
7:00	11,9639	12,0745	15,5904
8:00	10,2067	10,8180	16,0046
9:00	8,9534	8,5666	16,8500
10:00	8,6951	7,4501	18,4237
11:00	9,4247	10,0681	19,5348
12:00	11,2443	11,7695	21,5557
13:00	11,5131	11,4815	22,4542
14:00	12,3430	12,1483**	23,0550
15:00	12,7964	11,8559	23,6215
16:00	12,1900	11,9195	22,8907
17:00	11,1213	9,3032	21,4275
18:00	11,3497	9,8534	21,1259
19:00	10,4802	8,8919	19,0054
20:00	8,8510	8,1161	17,1900
21:00	8,6899	7,6188	17,0058
22:00	8,4961	7,3940	16,5700
23:00	7,6581	6,7294*	14,8518
0:00	7,8667	7,2085	13,1407
1:00	9,3165	7,4385	13,8641
2:00	9,3060	8,1191	13,4549
3:00	11,0218	8,3043	13,4536
4:00	12,2232	8,8472	13,5902
5:00	11,2723	9,8104	12,9205
Average MAPE	10,3849	9,3634	17,5188
NMSE	0,0969	0,0808	15,5904
* - minimum hourly MAPE of the best NN model ** - maximum hourly MAPE of the best NN model			

Table 4 shows that the MLP algorithm tested on Model 2 with tangent hyperbolic activation function produced better results than the other two architectures tested on this model. The NMSE of this network obtained on the test sample was 0,0808, while the MAPE was 9,36 %. The same algorithm using logistic activation function yielded MAPE of 10,38 %, while the radial-basis function network was again the worst in performance producing the MAPE of 17,52 %. The best network of Model 2 consisted of only 3 hidden neurons as the result of a cascading procedure. The same network shows the highest accuracy at 23:00 hours (MAPE is 6,73 %) while its lowest accuracy occurs at 14:00 (12,15 %).

When the results of Model 1 and Model 2 were compared (see Table 3 and Table 4), it can be seen that the lowest MAPE of all NN architectures is produced by the MLP algorithm with tangent hyperbolic function with the average MAPE on the test set of 9,36 %. It is obtained by Model 2, showing that the feature selection procedure improved the model accuracy. In order to further analyze the results of this best model, its performance is observed in more details in the following section.

5.2.
Analysis of the best neural network model results
Analiza rezultata najboljeg modela neuronske mreže

For a gas distributor it is important to observe the performance of the best NN model in hourly intervals, also in different days at week. Figure 4 presents the real average gas consumption (*Y real*) and the consumption predicted by the NN model (*Y computed*) in hourly time intervals averaged by the number of observations in the test sample.

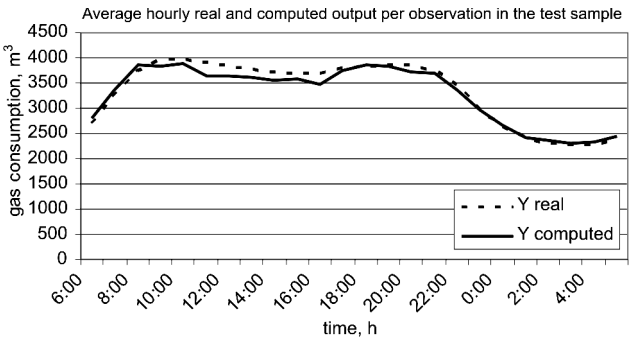

Figure 4 *Average real and computed hourly gas consumption on the test sample,*
Slika 4. *Prosječna stvarna i izračunata satna potrošnja plina na testnom uzorku*

It can be seen from Figure 4 that the NN model is, with some deviations, able to closely follow the line of the real average hourly gas consumption. Deviations are especially observable during the working hours.

Figures 5a and 5b enable a closer look into the movement and some critical points of MAPE through the test sample.

The hourly MAPE averaged by the number of observations in the test sample graphically presented in 2D-view in Figure 5a shows that the period from 21:00 to 1:00 and exactly at 10:00 hours has the lowest MAPE (bellow 8 %) revealing that the consumption at late night period of day is the most predictable one. Figure 4 shows a decrease of gas consumption at that period of time, which could be expected due to the fact that it is the end of working day for most residents and companies. The most problematic period of time is the period with MAPE higher than 10 %, which is observed from 7:00 to 8:00 and from 12:00 to 16:00. A specific characteristic of the first problematic period is the beginning of the working day of most companies in that region, and it is the time of day when residents usually start the heating in their homes. Deviations of gas consumption at

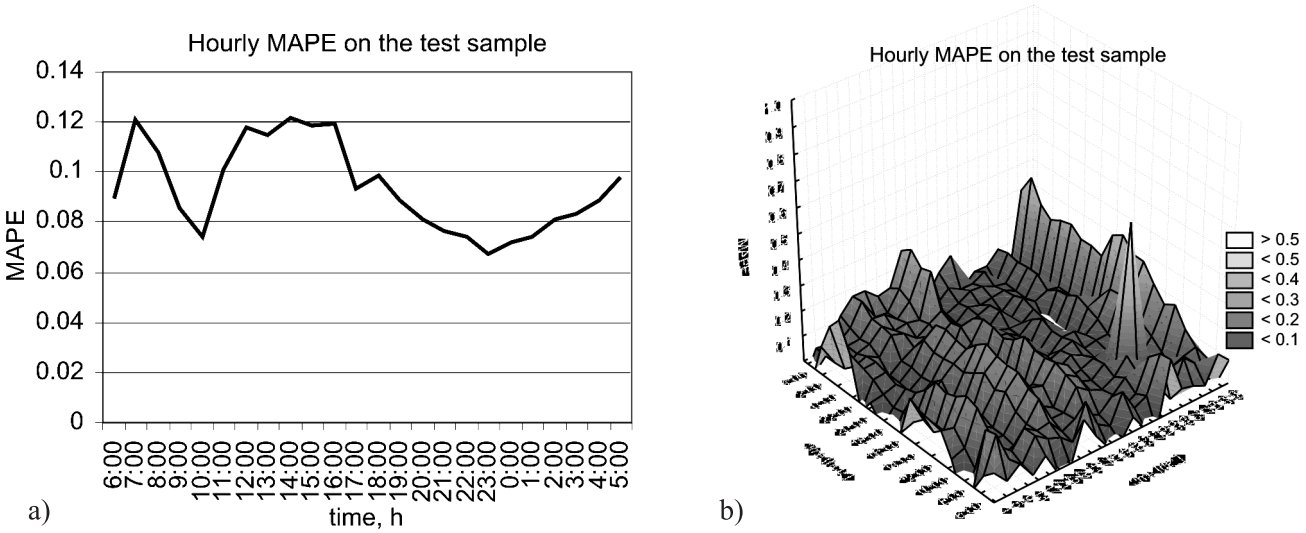

Figure 5 *2D graph of the hourly MAPE on the test sample (a) and 3D graph of the hourly MAPE on the test sample (b)*
Slika 5. *2D graf satne MAPE greške na testnom uzorku (a) i 3D graf satne MAPE greške na testnom uzorku (b)*

that time of the day should be investigated in more details, since it is not explainable by the variations in day type (working day, weekend holiday) because the feature selection procedure excluded the Day type and Day at week as relevant variables. 3D view of MAPE presented in Figure 4b shows that there are some extremely high peeks observed by the end of the test sample period (in March, 2009). One of them shows a high MAPE error of 117 % at 13:00 hours on March 22, 2009. If we look at the data of real gas consumption at that time, a high fall of consumption from 3056 cubic meters to 1389 cubic meters of gas occurred at that time. Such situation is not usual; it could be caused by some damage on the distribution system or other factors. In order to more closely analyze that period of time, the graph of real and computed total daily gas consumption on the test sample is presented in Figure 6, and the graph of MAPE on the same sample is shown in Figure 7.

It can be seen from Figure 6 that in some observations of the test set the NN model manages to closely predict the real daily consumption, while the deviations were present especially in observations where the line of the real consumption fluctuates rapidly.

Figure 7 shows that the average MAPE line varies regarding the date in the test sample. The lowest observed daily MAPE is 3,06 % obtained on March 18, 2008, while the highest MAPE is also observed in the same month, where the maximum MAPE on the test sample was 40,02 %. The above indicates that there is an unusual behavior of gas consumption in a certain period of the test sample that should be investigated in more details in further research. Some possible causes of such behavior should be searched for, and one of the solutions is to include longer period of time into the sample in order to determine if such deviations are caused by seasonal effects, or they could be treated as outliers.

The above results indicate that: (1) the best NN model based on 41 input variables, using MLP algorithm with the tangent hyperbolic activation function, is able to predict the next-day hourly gas consumption with the average MAPE of 9,36 % obtained on the test sample, (2) the largest MAPE (above 10 %) is obtained at working hours from 11:00 to 16:00, indicating that those hours should be investigated in more details, perhaps by developing a separate model for this problematic period of day, (3) the analysis of MAPE in each day of the test sample indicates that there are problematic days with large fluctuations at the end of the test sample, which could not be captured by the variable "Season" that indicates a seasonal effect. The results obtained in this paper are not directly comparable to other authors' results described in section 2, due to the fact that other authors use different modeling strategy, time horizons, or a different set of input variables. The topology of the NN model is most comparable to Beccalli et al. [4] who reported only the MAPE of one forecasting day of electricity demand, which

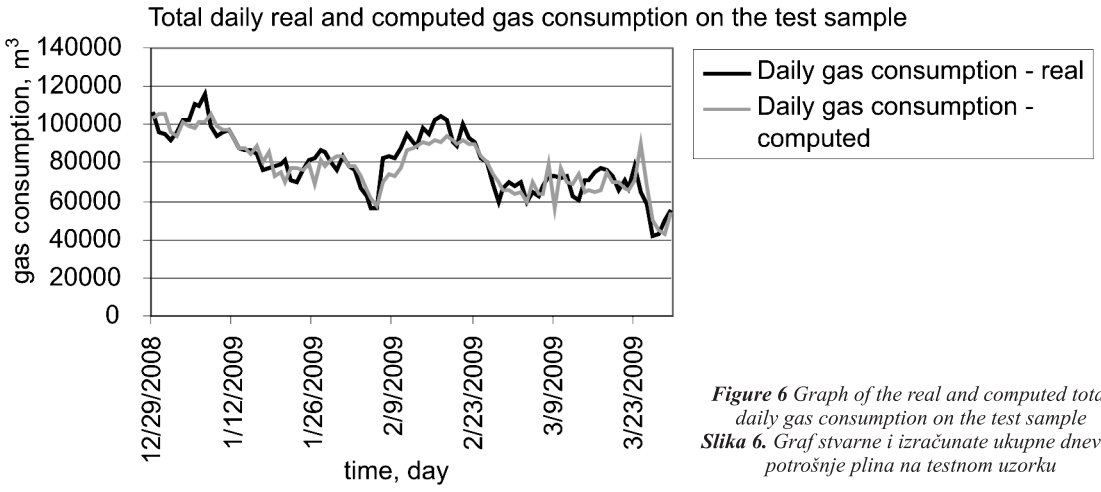

Figure 6 *Graph of the real and computed total daily gas consumption on the test sample*
Slika 6. *Graf stvarne i izračunate ukupne dnevne potrošnje plina na testnom uzorku*

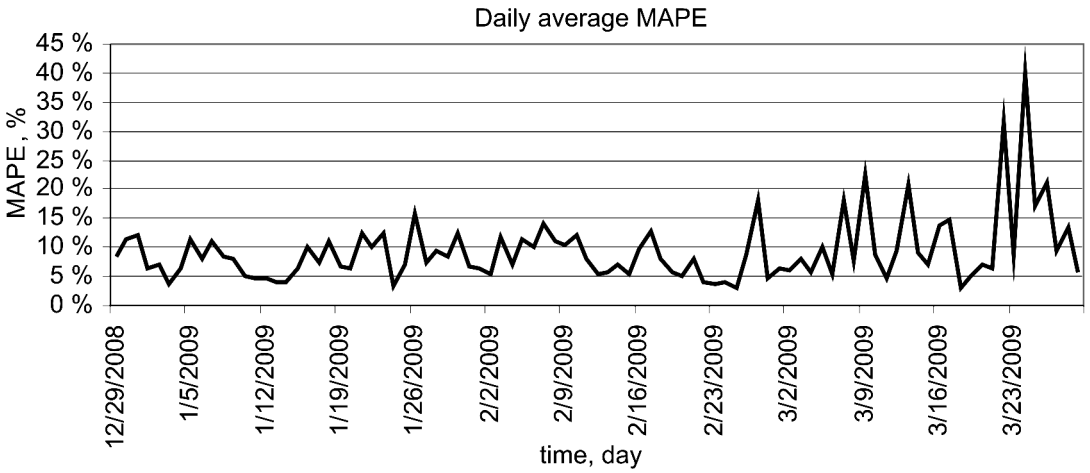

Figure 7 *Graph of the daily average MAPE of the best NN model obtained on the test set*
Slika 7. *Graf prosječne dnevne MAPE greške najboljeg modela NM dobivenog na testnom uzorku*

was in average 1,97 %. Other authors, such as Darbellay and Slama [3], also report lower errors of electricity demand prediction, while the research of Potocnik et al. [9] modeled gas consumption by using different error measures and longer period of time in the total sample. Since this is a preliminary research, the improvement in model accuracy is to be expected.

In order to improve the prediction accuracy, we plan to enlarge the dataset to capture at least the period of two years so that seasonal effects could be investigated in more details. It should be determined if certain rapid fluctuations belong to outliers, or to a pattern that should be discovered and included in the model as an additional input variable that will indicate expected sudden rise or a fall of the output. Regarding this problem, the suggestion of Beccali et al. [4] should also be considered which includes supervised learning algorithm to identify clusters in the data previously to NN modeling. Furthermore, designing separate models dedicated to problematic hours could enable the neural network to more accurately capture the gas consumption in those hours.

6 Conclusion Zaključak

The paper investigates the prediction of natural gas consumption on a regional level by using neural networks. Two neural network models were created to predict the next-day hourly consumption of natural gas based on the previous-day hourly consumption and a set of exogenous variables as predictors. A feature selection procedure extracted meteorological data (temperature prognoses, wind velocity, wind direction), season detection fuzzy variable, and previous-day hourly consumptions as important ones, while the *Day type* and the *Day of the week* were found irrelevant for the output. Two neural network algorithms were tested: the multilayer perceptron and the radial basis function network with different activation functions. The best neural network model is selected on the basis of the mean absolute percentage error obtained on the test sample. The smallest error (9,36 %) is produced by the multi layer perceptron algorithm. Its results were analyzed in terms of critical periods where the error is above 10 %. Some critical hours within a day, as well as problematic days within the test sample were identified. Since this is a preliminary research, a more thorough research activity is planned that will include larger dataset to enable more detailed investigation of seasonal effects, patterns and outliers. It would be also valuable to design separate models dedicated to critical hours and days, and different time horizons (2, 3, or more days). Further research should also include other methodologies besides neural networks, such as ARMA forecasting technique, multiple regression, stochastic methods, and others. Those methods require more data transformation than neural networks in the data-preprocessing phase, but the comparison of prediction accuracy provided by different methods will be worthwhile.

In order to add some applicability to the prediction model, an economic component should be also added to calculate expected gain or loss for a distributor in different time intervals, for example day, week, month, and year. Integrating the model into a decision support system of a distributor will enable managers to improve the quality of their operational and strategic decisions.

The model could be implemented by using XML technology for network deployment and by automating the process of incremental data load and results reporting by a web-based application that will also enable periodical network retraining and continuous usage. Although developed for predicting regional gas distribution, with minor modifications in the selection of input variables, the model could be easily implemented in wider area of other energy resources, or at other geographical levels, such as local or national.

7
References
Reference

[1] Šunić, M. Regulatori tlaka plina i regulacijske stanice, Energetika marketing, Zagreb, 2001.
[2] Gutierrez, R.; Nafidi, A.; Gutierrez Sanchez, R. Forecasting total natural-gas consumption in Spain by using the stochastic Gompertz innovation diffusion model, Applied Energy, 80 (2005), pp. 115–124.
[3] Darbellay, G. A.; Slama, M. Forecasting the short-term demand for electricity - Do neural networks stand a better chance?, International Journal of Forecasting, 16 (2000), pp. 71–83.
[4] Beccali, M.; Cellura, M.; Lo Brano, V.; Marvuglia, A. Forecasting daily urban electric load profiles using artificial neural networks, Energy Conversion and Management, 45 (2004), pp. 2879–2900.
[5] Thaler, M.; Grabec, I.; Poredos, A. Prediction of energy consumption and risk of excess demand in a distribution system, Physica A 355 (2005), pp. 46–53.
[6] Gelo, T. Ekonometrijsko modeliranje potražnje za plinom (Econometric modelling of the gas demand), Ekonomski pregled (Economic Review), 57, 1-2(2006), pp. 80-96.
[7] Potocnik, P.; Thaler, M.; Govekar, E.; Grabec, I.; Poredoš, A. Prakticni vidiki strojnega napovedovanja odjema plina (Practical aspects of machine forecasting of natural gas consumption), Mednarodno srecanje daljinske energetike in IX. strokovno posvetovanje SDDE, Portoroz, 10-12.04.2006.
[8] Potocnik, P.; Govekar, E.; Grabec, I. Short-term natural gas consumption forecasting, Proceedings of the 16th IASTED International Conference Applied Simulation and Modeling, August 29-31, 2007, Palma de Mallorca, Spain, pp. 353-357.
[9] Potocnik, P.; Thaler, M.; Govekar, E.; Grabec, I..; Poredoš, A. Forecasting risks of natural gas consumption in Slovenia, Energy Policy, 35 (2007), pp. 4271–4282.
[10] Detienne, K. B.; Detienne, D. H.; Joshi, S. A. Neural networks as statistical tools for business researchers. *Organizational Research Methods*, 2003;
6. doi. 10.1177/1094428103251907, pp. 236-265.
[11] Li, E. Y. Artificial neural networks and their business applications. Information & Management, 27(1994), pp. 303-313.
[12] Masters, T. Advanced algorithms for neural networks. New York. John Wiley & Sons; 1995.
[13] Karayiannis, N. B.; Weigun, G. M. Growing Radial Basis Neural Networks. Merging Supervised nad Unsupervised Learning with Network Growth Techniques. IEEE Transactions on Neural Networks, 8, 6(1997), pp. 1492-1505.
[14] HEP Plin Ltd., About us (O nama), http://www.hep.hr/plin/onama/default.aspx, 28.08.2009.
[15] Witten, I. H.; Frank, E. Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann Publishers, San Francisco, 2000.

Authors' addresses
Adrese autora

dr. sc. Zlatko Tonković
Croatian Electrical Company (cr. HEP)
HEP - Plin Ltd., HR-31000 Osijek, Croatia
zlatko.tonkovic@hep.hr

prof.dr.sc. Marijana Zekić-Sušac
University of Josip Juraj Strossmayer in Osijek
Faculty of Economics in Osijek, HR-31000 Osijek, Croatia
marijana@efos.hr

Marija Somolanji
Croatian Electrical Company (cr. HEP)
HEP - Plin Ltd., HR-31000 Osijek, Croatia
marija.somolanji@hep.hr


CALL FOR PAPERS
Fifth International Conference on Waste Management and the Environment
WASTE MANAGEMENT 2010
12 - 14 July 2010
Tallinn, Estonia
ORGANISED BY:
Wessex Institute of Technology, UK
Nagoya University, Japan
SPONSORED BY:
WIT Transactions on Ecology and the Environment
WESSEX INSTITUTE OF TECHNOLOGY
Advancing International Knowledge Transfer
www.wessex.ac.uk