Journal of Electrical Engineering and Information Technologies, Vol. **5,** No. 2, pp. 135–142 (2020)
Article 179 In print: ISSN 2545–4250
Received: March 25, 2020 On line: ISSN 2545–4269
Accepted: April 24, 2020 UDC: 621.316.1:339.133]:519.246.8(497.7)

Original scientific paper

SHORT-TERM LOAD FORECASTING USING TIME SERIES ANALYSIS: A CASE STUDY FOR THE REPUBLIC OF NORTH MACEDONIA

Ana Kotevska, Nevenka Kiteva Rogleva

Faculty of Electrical Engineering and Information Technologies,
"Ss. Cyril and Methodius" University in Skopje,
Rugjer Boškovik 18, PO Box 574, 1000 Skopje, N. Macedonia
ana.kote@hotmail.com

A b s t r a c t: Accurate load forecasting models are essential for normal operation and schedule planning in utility company. This paper presents the development of short-term load forecasting models for the Republic of North Macedonia and gives a comparison review of various models for load forecasting. These models use time series analysis such as the Autoregressive Integrated Moving Average model and the Seasonal Autoregressive Integrated Moving Average with Explanatory Variable model. Time series approach is one of the most used methods for short-term load forecasting. Models were designed and implemented in Python. The results were evaluated by the Mean Absolute Percentage Error of 0.5% for the forecasted day.

Key words: autoregressive integrated moving average (ARIMA); autocorrelation function (ACF); mean absolute percentage error (MAPE); partial autocorrelation function (PACF); seasonal autoregressive integrated moving average with explanatory variable (SARIMAX)

КРАТКОРОЧНА ПРОГНОЗА НА ПОТРОШУВАЧКА НА ЕЛЕКТРИЧНА ЕНЕРГИЈА СО КОРИСТЕЊЕ НА ВРЕМЕНСКИ СЕРИИ: СТУДИЈА ОД СЛУЧАЈ ЗА РЕПУБЛИКА СЕВЕРНА МАКЕДОНИЈА

А п с т р а к т: Точните модели за прогнозирање на потрошувачката на електрична енергија се од суштинско значење за нормално работење и планирање на оптоварувањето кај компаниите кои се занимаваат со снабдување и тргување со електрична енергија. Овој труд вклучува развој на модели за краткорочна прогноза на потрошувачката на електрична енергија во Република Северна Македонија, како и споредување на различни модели. Овие модели користат временски серии како што се авторегресивен интегриран движечки просек и сезонски авторегресивен интегриран движечки просек со модел на егзогени променливи. Моделите се изработени во Питон. Резултатите се оценети со средна апсолутна процентуална грешка од 0,5% за прогнозираниот ден.

Клучни зборови: авторегресивен интегриран движечки просек; функцијата на автокорелација; средна апсолутна процентуална грешка: делумната функција на автокорелација; сезонски авторегресивен интегриран движечки просек со модел на егзогени променливи

1. INTRODUCTION

Load forecasting is crucial for the effective and efficient operation of any power system. It is important and helpful not just for the operation phase of the electricity industry, but also from a financial point of view. Forecasting supports energy planners in understanding the influence of some uncertain and uncontrollable factors such as weather conditions, electricity prices, economic activity, an increase in the number of consumers, as well as the price of other power sources on energy consumption.

Long-term load forecasting is fundamental for strategic planning, construction of new generations, and develops the power supply and delivery system;

and short-term load forecasting is used for balancing electricity supply. Short-term load forecasting is difficult because it may depend on the load from previous days and also on the type of that day (in terms of the working day, weekend or holiday) in the previous weeks, even in the previous years. It is difficult to model the relation between the load and the external factors, such as time variation, holidays, etc. [1, 2].

The main focuses of this paper are the models for short-term load forecasting. Setting of a model usually is subject to numerous variations like available data used as inputs, the time frame, the time resolution (minutely to annually), the scale, etc. Some of the most widely used methods for short-term prediction are the Autoregressive Integrated Moving Average (ARIMA) and the Autoregressive Moving Average (ARMA). Time series approach is based on the assumption that in absence of major disruption events, an event in the future will be related to past events and can be expressed via models developed from historical data. Their accuracy depends on the sufficient and accurate past data, given that some past trends can provide inaccurate results and disrupt the series [3].

For this kind of load forecast, the election of the model has a crucial role, along with the input parameters and their transformation. Also, it is important to have knowledge of statistics, so that the gained parameter values can be understood and consequently make modifications in order to improve the model. ARIMA model is used for load prediction for Distribution System Operator (DSO) of the Republic of North Macedonia. DSO's electricity historical data for 2 years (2014 and 2015) were implemented in the model. Data from 2014 till September 2015 was used for model development and then these models have been tested on the rest of 2015 load data. These models were designed and implemented in Python.

The rest of the paper is organized as follows: Section 2 briefly discusses the electricity load forecasting, while section 3I presents the methods used to load forecasting, and in section 4 the results are discussed. Section 5 concludes the paper.

2. ELECTRICAL LOAD FORECASTING

Unlike other material products, the power electricity has a unique feature, it cannot be stored in bulk and it should be produced as soon as it is demanded. Energy suppliers use methods to forecast power consumption in order improve operating and maintaining phase and increase their budget, as well as increasing their efficiency in supplying and distribution of power energy. Load forecasts are used in all segments of the electricity industry, including production, transmission, distribution and retail. Due to the fundamental role of load forecasting in the service of supply operations, incorrect load forecast can result in financial losses or even bankruptcy of the enterprise. An accurate load forecast has a crucial role in operation and further development of electric power system, while an inaccurate forecast can lead to equipment failures. In addition, an accurate load forecast can be helpful in developing a power supply strategy, market research, financial planning and has a crucial role in the forecasting of energy capacity prices and energy prices. Without an optimal load forecast, the power supply services are at risk of excessive or insufficient energy purchases in the market for the day ahead. Although companies can buy or sell energy on the intraday market to correct the inaccuracy of the forecasting, this results in higher prices on the intraday market. The goal of the forecast model is to obtain a forecast with the least error because a reduction in the average error of the load forecast has the potential to save hundreds of millions of dollars [4].

Electricity load forecast is influences by various uncertain and uncontrollable factors such as [5]:

1. Economic factors. The economic situation in specific area can affect the form of consumption. This situation can include the customer type (as residential, commercial and industrial), demographic conditions, industrial activities and population. These conditions mainly affect the long-term load forecast.

2. Time factors. Time factors include seasonal, weekly and holiday effects. The holidays have a big impact on the load curve because they go lower than usual consumption. Here, in most of the cases, the technique for similar days (that involves searching for historical data that has similar characteristics with the forecasted day) is used.

3. Weather factors. Temperature is the most essential extreme factor in load forecasting. The change in temperature can affect the amount of energy needed for winter heating and air conditioning during the summer. Usually, today's consumption is affected by yesterday's temperature, so if the previous day was particularly hot, the consumer can increase the use of air conditioning in the current day. Other weather factors that affect load forecast include humidity, especially in hot and humid areas, rainfall, thunderstorms, cloud cover, wind intensity and daylight.

4. Random disturbances. Large industrial consumers can cause sudden changes in their consumption and are going to lead to the load curve impulses. Also, certain events and conditions can cause sudden changes in load flow diagram, such as popular TV shows, sports events or shutdown of industrial operations.

5. Price factor. In the power markets, the price of electricity can also be an important factor in load forecasting.

6. Other factors. The form of consumption may be different due to geographical conditions, type of consumer, the price of other energy supply sources, etc. From the factors mentioned above usually weather and time factors can be injected as inputs into the Kalman filter, since it is extremely difficult to deal with the complex variables like economic and customer factors or random disturbances.

3. METHODS AND IMPLEMENTATION

This section describes the methods appropriate for short-term forecasting of North Macedonia load flow. Due to the presence of a double seasonal pattern in load demand data, Autoregressive Integrated Moving Average (ARIMA) model and Seasonal Autoregressive Integrated Moving Average with Explanatory Variable (SARIMAX) model are used for the load flow prediction.

The initial step in developing stage of statistical model is the election of the input variables and data processing, which includes cleaning and transforming the data, as well as filling in the missing data. During the investigation we faced with too many problems related to quality of the real data. The most common problem was the lack of values, which may be due to a temporary shutdown of the system, SCADA or meteorological station. These missing readings can be technically "corrected" by filling in the missing data based on some linear extrapolations.

Figure 1 shows the graph of the daily average consumption and daily average temperature for 2014 and 2015 in North Macedonia. The correlation coefficient between the daily average temperature and the daily average consumption is –0.89, which shows that an increase in the daily average temperature would cause a decrease in the daily average consumption by 89%. This inverse ratio between the data can be especially noticed in the summer and winter periods. It can be calculated as follows:

$$\rho_{x,y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{E(X-\mu_X)^2(Y-\mu_Y)^2}}, \qquad (1)$$

where Cov is the covariance; σ_X is the standard deviation of X; σ_Y is the standard deviation of Y; μ_X is the mean of X; μ_Y is the mean of Y; E is the expectation.

Figure 2 shows the consumption of electric power based on the type of day. The consumption profiles for working days are significantly different from the consumption profiles for weekends and holidays.

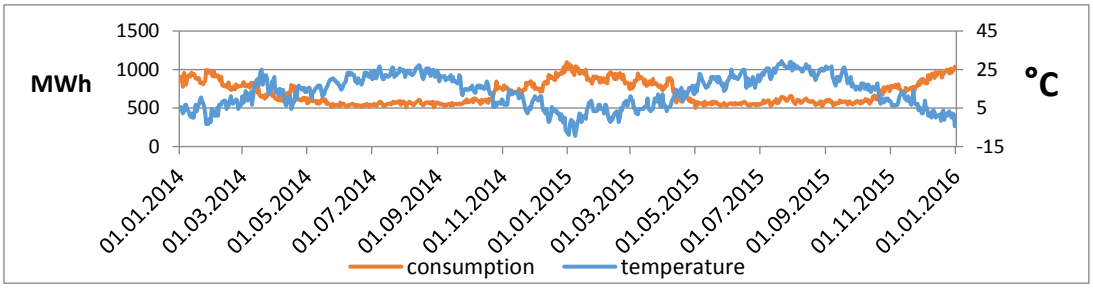

Fig. 1. The graph shows the daily average consumption and the daily average temperature for 2014 and 2015

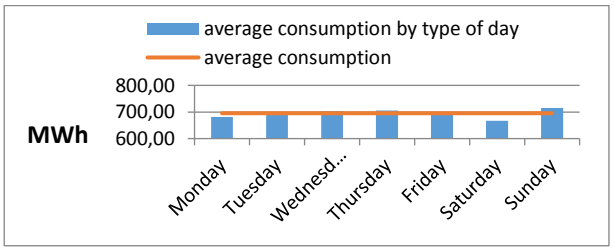

Fig. 2. Average consumption of electricity power by type of day

This means that the average errors in load forecasting for working days are lower than those on weekends and holidays, as load curves for working days are almost identical.

However, Saturday and Sunday can be observed as separate profiles. Saturdays are more close to the holiday load. Sunday is most likely affected by the low prices from the power markets and the low tariff for the retail.

3.1. *ARIMA*

The input series for ARIMA should be stationary, i.e. to have a constant mean value, variance and autocorrelation over time. Therefore, usually, the series firstly should go through a differencing process until it is stationary. In the initial phase of model development, analyzing the autocorrelation and the partial autocorrelation of different series can identify the number of AR and MA components. The name of the model itself gives the key aspects of the model:

AR: Autoregression, a model that uses a dependent relationship between observation and a number of lag observations. It is a linear regression model that uses its own lags as predictors because they work better when predictors are not correlated and are independent of each other.

I: Integration, use of differentiation on raw observations in order to make the time series stationary.

MA: Moving average, a model that uses the relationship between observations and the error of the moving average model applied to lag observations.

Each of these components is explicitly stated in the model as a parameter. The standard notation used is ARIMA (p, d, q), where the parameters are replaced with values to indicate the specific ARIMA model used. The model parameters are defined as follows:

p: The order of AR term. It refers to the number of lags of *Y* to be used as predictors. It can be calculated as follows:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} + \epsilon_t, \quad (2)$$

where $\beta_1, \ldots, \beta_p$ are the parameters of the model; α is a constant; ϵ_t is white noise.

d: The number of differencing required to make the time series stationary. It is the minimum number of differencing needed to make the series stationary. If the time series is already stationary, then $d = 0$.

q: The order of the MA term. It refers to the number of lagged forecast errors that should go into the ARIMA model. It can be calculated as follows:

$$Y_t = \mu + \epsilon_t + \varphi_1 \epsilon_{t-1} + \varphi_2 \epsilon_{t-2} + \cdots + \varphi_q \epsilon_{t-q}, \quad (3)$$

where μ is the mean of the series; $\varphi_1, \ldots, \varphi_q$ are the parameters of the model; $\epsilon_t, \epsilon_{t-1}, \ldots, \epsilon_{t-q}$ are the white noise error terms.

3.2. *Order of differencing (parameter d)*

The right degree of differentiation is the minimum differentiation needed to get an almost stationary series. If the autocorrelations are positive for a large number of lags, then the series needs to be further differentiated. On the other hand, if the autocorrelation for lag 1 is too negative, then the series is overly differentiated. In case it is not possible to decide between two degrees of differentiation, the degree that gives the least standard deviation in the differentiated series is used. To determine the stationary of the time series, the Augmented Dickey-Fuller (ADF) test is used. The purpose of this test is to determine how strongly the time series is defined by the trend. This is a statistical test that uses two hypotheses. The zero hypothesis indicates that the time series is not stationary (it has a certain dependence on time). The alternative hypothesis is that the time series is stationary. The result is interpreted using the p-value of the test. If the p-value is greater than 0.05, the zero hypothesis is not rejected, the series is non-stationary. If the p-value is less or equal to 0.05, the zero hypothesis is rejected, the series is stationary.

If this ADF test is applied to the data for model development, it will show that the time series is not stationary (Table 1), the p-value is 0.358453, i.e. it is greater than 0.05.

T a b l e 1

ADF test on the data for model development

ADF statistic	–1.844808
p-value	0.358453
Number of lags used	44
Number of observations used	17449

The next step is to use differentiation to make the time series stationary. Figure 3 shows the original time series, as well as the use of the first and second degree of differentiation and the corresponding auto-correlations. The time series reaches stationary with both degrees of differentiation. However, with the second degree of differentiation, the lag on the autocorrelation enters in the negative zone fairly quickly, which indicates that the series may have been over-differentiated. Therefore, the first degree of differentiation is used, although the series is not perfectly stationary (weak stationary). Table 2 shows the ADF test on the first degree of

differentiation data, where the p-value is less than 0.05 and the time series is stationary now.

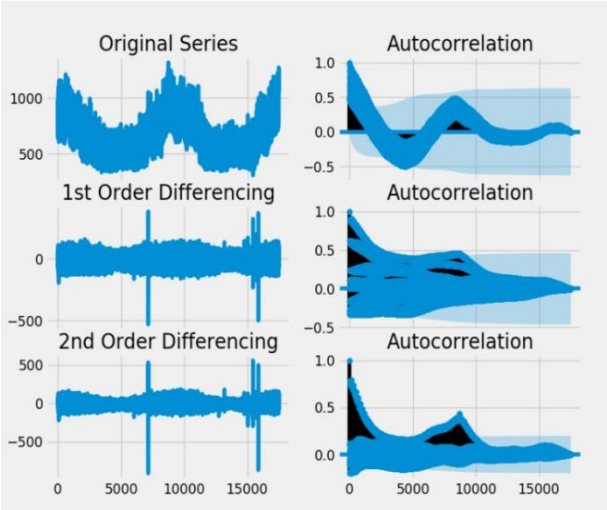

Fig. 3. Order of differencing

T a b l e 2

ADF test on the first degree of differentiation data

ADF statistic	- 35.905130
p-value	0
Number of lags used	44
Number of observations used	17448

3.3. Order of AR term (parameter p)

The required number of lag observations can be obtained using the partial autocorrelation function (PACF). The PACF can be thought of as a correlation between the series and its lag, after removing the effects of the intermediate lags. Any autocorrelation in a stationary series can be corrected by adding enough AR terms. Initially, the AR term is equal to as many lags as possible exceeding the PACF limit. If PACF is used on the first degree differentiation data (Figure 4), then lag 1 and lag 24 are high above the significant limit.

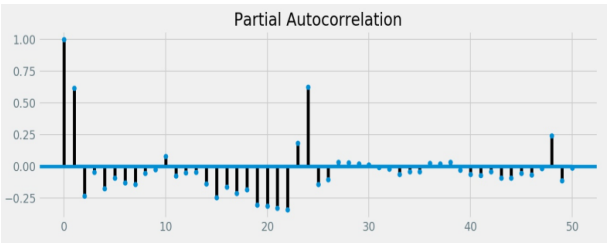

Fig. 4. PACF on the first degree differentiation data

3.4. Order of MA term (parameter q)

Autocorrelation function (ACF) can be used to determine the MA term, where the lag forecast error is technically seen. The ACF shows how many MA terms are needed to remove autocorrelation in a stationary series. If ACF is used on the first degree differentiation data (Figure 5), the lag 1 and lag 2 are high above the significant limit.

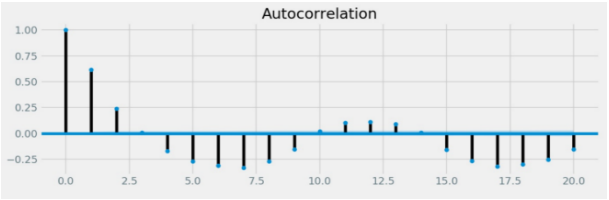

Fig. 5. ACF on the first degree differentiation data

3.5. SARIMAX

The problem with the ARIMA model is that it does not support seasonality. If the time series has seasonal behavior, then it is better to use SARIMA which uses seasonal differentiation. Seasonal differentiation is similar to regular differentiation, but instead of taking away the subsequent terms, the value from the previous season is taken away. The model is presented as follows, SARIMA (p, d, q) × (P, D, Q), where, p, d, q are the trend elements; x is the frequency of time series; P, D, Q are the seasonal elements. The SARIMA model can be improved if an exogenous variable is used. This model is called the SARIMAX model. The only requirement for the use of an exogenous variable is to know the value of the variable during the forecast period.

3.6. MAPE

For the measurement of the forecast accuracy, Mean Absolute Percentage Error (MAPE) is used. MAPE measures the amount of error in terms of percentage. It is calculated as the average of the absolute percentage error. It can be calculated as follows:

$$MAPE = \frac{\sum \frac{|Actual_t - Forecast_t|}{|Actual_t|}}{n} * 100. \qquad (4)$$

4. RESULTS

The models used in this study were developed in Python, using the following libraries:

- statsmodels.graphics.tsaplots for ACF and PACF,

- statsmodels.tsa.stattools for ADF,
- statsmodels.tsa.arima_model for ARIMA
- statsmodels.tsa.statespace.sarimax for SARIMAX.

4.1. *ARIMA model*

Three forecasting models were developed for 09.09.2015:

Model 1: Hourly consumption just from the previous day.

Model 2: Hourly consumption of the type of day from the previous week (for example, if is forecasted Wednesday, the consumption from the previous Wednesday will be used).

Model 3: All the available data till the forecasted day.

Based on the above calculations for the parameters *d* (ADF test), *p* (PACF) and *q* (ACF), several ARIMA models are performed. The one with *P*-value (column '*P* > | *z* |') less than 0.05 for each AR and MA terms and the lowest AIC value is chosen (Figure 6). For Model 1 and Model 2 after the calculations and analysis, ARIMA (1, 1, 0) is used and for Model 3, ARIMA (24, 1, 1) is used. Figure 7 shows the graph that compares the forecast and actual values with the ARIMA models. The calculated MAPE for these models is around 5% which implies that the model is correct about 95% in the forecast for 09.09.2015.

ARIMA Model Results

Dep. Variable:	D.consumption	No. Observations:	17519
Model:	ARIMA(24, 1, 1)	Log Likelihood	-75434.778
Method:	css-mle	S.D. of innovations	17.923
Date:	Mon, 04 Nov 2019	AIC	150923.555
Time:	07:18:11	BIC	151133.373
Sample:	01-01-2014 - 01-01-2016	HQIC	150992.644

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0089	0.066	0.133	0.894	-0.121	0.139
ar.L1.D.consumption	-0.0191	0.007	-2.741	0.006	-0.033	-0.005
ar.L2.D.consumption	-0.1094	0.005	-20.028	0.000	-0.120	-0.099
...						
ar.L23.D.consumption	-0.0337	0.006	-6.009	0.000	-0.045	-0.023
ar.L24.D.consumption	0.7030	0.005	130.790	0.000	0.692	0.714
ma.L1.D.consumption	0.3149	0.009	36.098	0.000	0.298	0.332

Fig. 6. ARIMA (24, 1, 1) model

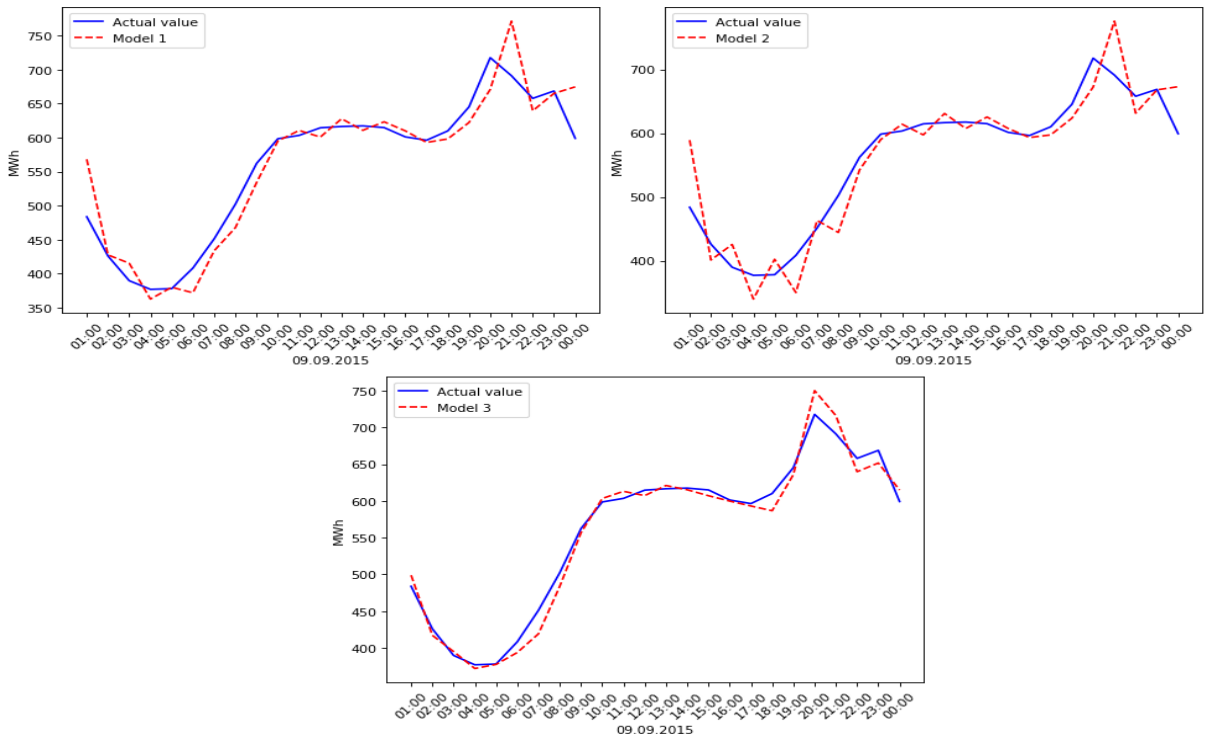

Fig. 7. The graph that compares the forecast and actual values with the ARIMA models

4.2. *SARIMAX model*

As an exogenous variable, the temperature is used because it has shown the highest correlation with consumption. After the calculations and analysis for SARIMAX models, SARIMAX (1, 1, 5) × (1, 1, 1, 24) is used. The following models are developed for 09.09.2015.

Model 4: The training data set starts from 01.09.2014 to 08.09.2015.

Model 5: The training data set starts from 01.01.2014 to 08.09.2015.

The calculated MAPE for these models is around 3.6% which implies that the SARIMAX model is slightly better than the ARIMA model. Figure 8 shows the graph that compares the forecast and actual values with the SARIMAX models.

4.3. *Improvement for ARIMA model*

Improvement of ARIMA model is made by increasing the training data set. Since there are only two years available, the forecasted day is 23. 12. 2015, and all the data till 22. 12. 2015 is used for learning. The model ARIMA (24, 1, 1) is used and the calculated MAPE, in this case, is 0.5%, which means that the forecasted day can be predicted with an accuracy of 99.5%. Figure 9 shows the graph that compares the forecast and actual values with this ARIMA model.

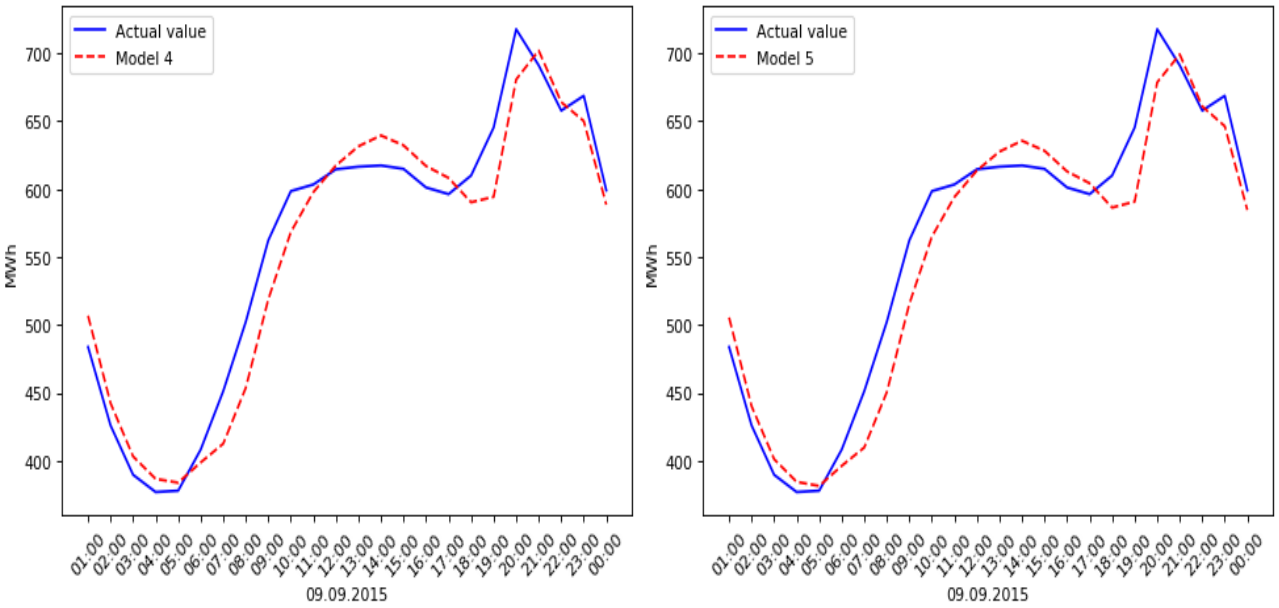

Fig. 8. The graph that compares the forecast and actual values with the SARIMAX (1, 1, 5) × (1, 1, 1, 24) models

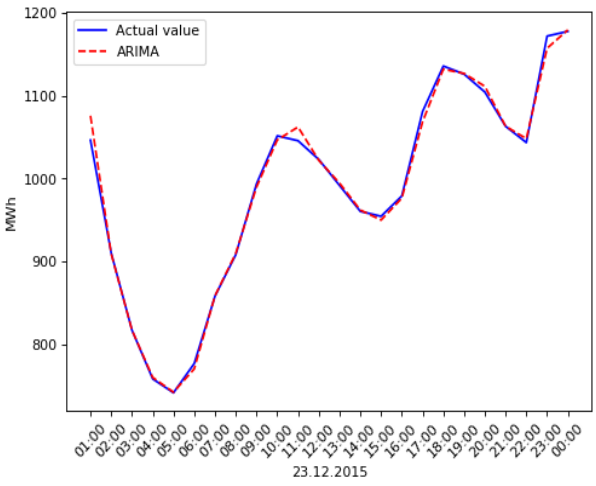

Fig. 9. The graph that compares the forecast and actual values with the ARIMA (24, 1, 1) model

5. CONCLUSION

The objective of this paper was to develop a model for the load demand forecasting which gives a best prediction values with minimum errors so that Control Central planners can estimate what the near demand will be.

The short-term load forecasting using time series analysis has been applied to the load flow data (2014 and 2015) for the Republic of North Macedonia. This paper gives a comparation analysis of ARIMA and SARIMAX models. After simulation and estimation of the parameters and the coefficients for those models, the application of SARIMAX models, using the temperature as an exogenous variable, did not significantly improve the

results. The results obtained from simulation also indicate that the size of the data set has a crucial role in reducing the error measurement for the ARIMA model, having a MAPE of 0.5%. We can conclude that ARIMA model should be used when having historical data of 2 – 3 years this is because the ARIMA models place heavy emphasizes on the recent past rather than the distant past.

Future work will focus on comparing those models with others techniques such as Artificial Neural Network.

REFERENCES

[1] Feinberg, E. A., Genethliou, D.: *Load Forecasting*, Springer, Boston, MA, Online ISBN: 978-0-387-23471-7, Print ISBN: 978-0-387-23470-0.

[2] Mosad Alkhathami1, Introduction to electric load forecasting methods, *Journal of Advanced Electrical and Computer Engineering*, Vol. **2**, No. 1 (2015), Columbia International Publishing,

[3] Isaac Adekunle Samuel, Tolulope Ojewola, Ayokunle Awelewa, Peter Amaize, Short-term load forecasting using the time series and artificial neural network methods, *IOSR Journal of Electrical and Electronics Engineering* (*IOSR-JEEE*), e-ISSN: 2278-1676, p-ISSN: 2320-3331, Volume **11**, Issue 1, Ver. III (Jan. – Feb. 2016), PP 72–81.

[4] Hinman, J., Hickey, E.: *Modeling and Forecasting Short-Term Electricity Load using Regression Analysis*, Illinois State University, Fall, 2009.

[5] Hahn, H., Meyer-Nieberg, S., Pickl, S.: Electric load forecasting methods: Tools for decision making, *European Journal of Operational Research* 199.

[6] Kotevska, A.: *Analysis and Models for Short Term Load Forecasting in the Republic of North Macedonia*, master thesis, Ss. Cyril and Methodius University in Skopje, Faculty of Electrical Engineering and Information Technologies, North Macedonia, 2020,