Short-Term Load Forecasting Using Time Series Analysis: A case study for Republic of North Macedonia

Ana Kotevska, Nevenka Kiteva Rogleva

*Faculty of Electrical Engineering and Information Technologies,*

*Ss. Cyril and Methodius University in Skopje,*

*Rugjer Boshkovik 18, PO Box 574, 1000 Skopje, Macedonia*

ana.kote@hotmail.com

Abstract: Load forecasts are important for energy suppliers. Accurate models for load forecasting are essential to the operation and planning of a utility company. This paper involves the development of short-term load forecasting models for the Republic of North Macedonia and a comparison of various models. 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. 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 на случај за Република Северна Македонија

Апстракт: Прогнозата на потрошувачка на електрична енергија е важна за снабдувачите со електрична енергија. Добрите и точните модели за прогнозирање на потрошувачката на електрична енергија се од суштинско значење за работењето и планирањето на компаниите кои се занимаваат со снабдување и тргување со електрична енергија. Овој труд вклучува развој на модели за краткорочна прогноза на потрошувачка на електрична енергија за Република Северна Македонија 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. 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.

#  INTRODUCTION

Load forecasting is crucial for the effective and efficient operation of any power system. It is important and helpful not just in the operation of the electricity industry, but also from a financial point of view. The load forecast is impacted by many 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. 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’s difficult to model the relation between the load and the external factors, such as time variation, holidays, etc. ([1], [2]).

In this study the models for short-term load forecasting are explained, using time series analysis for the Republic of North Macedonia. For the 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 values of the parameters can be understood and consequently make modifications in order to improve the model. Including data from 2 years (2014 and 2015), the load 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.

Time series approach is one of the most used methods for short-term load forecasting. This method 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].

This paper is organized as follows: Section II briefly discusses the electrical load forecasting, while section III presents the methods used to load forecasting, and in section IV the results are discussed. Section V concludes the paper.

#  ELECTRICAL LOAD FORECASTING

Unlike other material products, the power electricity has a unique feature, which is that it cannot be stored and must be produced in a load that could respond to the demands. Energy suppliers that use methods to forecast consumption will always save on operating, maintaining, and increasing their budget, as well as increasing their efficiency in suppling power energy and distribution. 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 holds a lot of benefits for the electric power system, in the operation and planning process, 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 one percent reduction in the average error of the load forecast has the potential to save hundreds of millions of dollars [4].

The first step in an accurate load forecast is to identify the factors that have an impact on consumption patterns. Some of these factors are [5]:

**1. Economic factors:** The economic situation in an 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. Examples of seasonal that affect the consumption pattern. The holidays have a big impact on the load curve because they go lower than usual consumption. Here, often the technique for similar days that involves searching for historical data that has similar characteristics with the forecasted day is used. These features can include time, date, and even month and year. In this case, the day with similar characteristics will be taken into account in the forecast. This means that the type of day is also used because there are important differences in consumption between working days and weekends. Consumption of different working days may also behave differently. For example, Monday and Friday, as they are close to the weekend, have a structurally different consumption from Tuesday to Thursday. Also, Saturday and Sunday can be observed separately. Holidays are more difficult to be predicted than non-holidays, due to their relative rarity.

**3. Weather factors:** Temperature is the most influential weather 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 Disruptions:** Large industrial consumers can cause sudden changes in their consumption. Also, certain events and conditions can cause sudden changes in consumption, 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.

Temperatures, past load, special events and time have a significant impact on the consumption in any region.

#  METHODS AND IMPLEMENTATION

The methods of short term load forecasting used in this study are the time series, such as Autoregressive Integrated Moving Average (ARIMA) model and Seasonal Autoregressive Integrated Moving Average with Explanatory Variable (SARIMAX) model. This study will show the results obtained from the short term load forecast that was carried out for the next day (24 hours ahead) and the gained results compared with the actual values recorded within the forecasted period. The variables taken are the hourly historical consumption, time and average daily temperature. As a temperature representative in this study, the average daily temperature in Skopje is used [6].

The first step in developing a 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. There are many problems with the quality of real data. The most common problem with data quality is 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.

Figure1 shows the graph of the daily average consumption and daily average temperature for 2014 and 2015. 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.

Figure2 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. 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.

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

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.

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

## 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 must 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 is the order of AR term. It refers to the number of lags of Y to be used as predictors.

d is 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 is the order of the MA term. It refers to the number of lagged forecast errors that should go into the ARIMA model.

## i.i. 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 (10 or more), then the series needs to be further differentiated. On the other hand, if the autocorrelation for lag1 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 (Table1), the p-value is 0.358453, i.e. it is greater than 0.05.

Table 1. Augmented Dickey-Fuller test on data for model development

|  |  |
| --- | --- |
| ADF Statistic | -1.844808 |
| p-value | 0.358453 |
| Number of Lags Used | 44 |
| Number of Observations Used | 17449 |
|  Critical Values |
| 1% | -3.431 |
| 5% | -2.862 |
| 10% | -2.567 |



Figure 3. Order of differencing

The next step is to use differentiation to make the time series stationary. Figure3 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). Table2 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.

Table 2. ADF test on first degree of differentiation data

|  |  |
| --- | --- |
| ADF Statistic | - 35.905130 |
| p-value | 0 |
| Number of Lags Used | 44 |
| Number of Observations Used | 17448 |

## i.ii. 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 (Figure4), then lag 1 and lag 24 are high above the significant limit.



Figure 4. PACF on the first degree differentiation data

## i.iii. 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 (Figure5), the lag 1 and lag 2 are high above the significant limit.



Figure 5. ACF on the first degree differentiation data

##  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) x (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.

## 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:

|  |  |
| --- | --- |
| $$MAPE= \frac{\sum\_{}^{}\frac{|Actual\_{t}- Forecast\_{t}|}{|Actual\_{t}|}}{n}\*100$$ | (1) |

#  RESULTS

## ARIMA model

Based on the above calculations and the analysis of the AIC value and P-value in the ARIMA model, ARIMA (24,1,1) has shown the best performance (Figure6). According to Figure6, the values below the ‘coef’ column are weights of the corresponding AR and MA terms, respectively. The P-value in the column ‘P> | z | ' is less than 0.05 for both AR and MA terms, which means that the model is well chosen. With the hourly consumption, three forecasting models are 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

 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. Figure7 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.

## SARIMAX model

 As an exogenous variable, the temperature is used because it has shown the highest correlation with consumption. After the calculations for SARIMAX model, SARIMAX (1,1,5) x (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. Figure8 shows the graph that compares the forecast and actual values with the SARIMAX models.



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



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



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

## Improvement for ARIMA model

Improvement 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%. Figure9 shows the graph that compares the forecast and actual values with this ARIMA model.

 It can be concluded that with greater training data set, the ARIMA model gives better results, which are closer to the real ones. However, as the data set grows, all the calculations and analyses need to be done again, because the parameters for the ARIMA model can be changed.



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

#  CONCLUSION

The short-term load forecasting using time series analysis has been applied to the load data (2014 and 2015) for the Republic of North Macedonia. This paper gives a comparative study of different models with time series. The SARIMAX models have shown better results than the ARIMA models if they are compared in terms of error measurements with MAPE having 3.6%. The size of the data set has a crucial role since it has shown a significant reduction of the error measurement for the ARIMA model, having a MAPE of 0.5%. This study presents usable and accurate models for short-term load forecasting which is very important for the operation of any power system.

##### REFERENCES

1. Eugene A. Feinberg, Dora Genethliou, “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”, Columbia International Publishing, Journal of Advanced Electrical and Computer Engineering (2015) Vol. 2 No. 1
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. Jennifer Hinman, Emily Hickey, “Modeling and Forecasting Short-Term Electricity Load using Regression Analysis”, Illinois State University, Fall 2009
5. Heiko Hahn, Silja Meyer-Nieberg, Stefan Pickl, “Electric load forecasting methods: Tools for decision making”, European Journal of Operational Research 199
6. Kotevska Ana, “Analysis and Models for Short Term Load Forecasting in Republic of North Macedonia”, Ss. Cyril and Methodius University in Skopje, Faculty of Electrical Engineering and Information Technologies, master thesis