# TREATMENT AND VISUALISATION OF BIG DATA IN POWER SYSTEMS

*Ilir Keka, Betim Çiço*

Faculty of Contemporary Sciences and Technologies,
South East European University (SEEU)
e-mails: ilir_keka@hotmail.com, b.cico@seeu.edu.mk
Macedonia

**Abstract:** Because the energy is difficult to store in a long time then for the electric companies, especially for trading companies and for transmission operator, it is very important to know the network load. The aim of this paper is to find a model of the relationship between electricity load and time for some Substations using a sample of Big Data. This is achieved using Revolution R Enterprise language. Furthermore, Big Data were analyzed to find the distribution of data and to calculate the descriptive statistical parameters for these data. At the end of the paper is made evaluation of this mathematical model.

**Key words:** Big Data, Electric Power, Revolution R Enterprise, Time, Regression model.

## 1. INTRODUCTION

Rapid developments of technology allow companies to record data known as Big Data. Reduced risks, improved decisions and maximize returns of costs are some key benefits during analyzing of Big Data in the wholesale energy market [1]. In [2] is used the k-Means algorithm and KNIME platform to treat Big Data of electricity in order to find time series forecasting. A new system architecture BlueDBM is presented in [3] as a homogeneous cluster of host servers using flash storage and in- processing capability analysis of Big data. Big Data analysis using the heterogeneous mixture learning technology is applied in [4] and as an example of applying of this technology is proposed the prediction of the electricity demand of a building. Deep Learning algorithm is used in [5] to help Big Data Analytics to address important problems like extracting complex patterns from massive volumes of data, semantic indexing or data tagging. In [6] is proposed a concept design for the visualization process of the control systems to the power systems, based on the GUI.

Information Technology and Communication (ICT) has covered all areas of life. It is difficult to imagine the economic, social, educational or industrial development without the use of ICT. Many countries around the world have prioritized investment in ICT, in order to achieve a greater Gross Domestic Production (GDP). ICT is also used in industry, as in Power Systems, especially in the electricity grid automation. When we mention the concept of ICT, in fact, we think the "Internet of things". Therefore, the use of the ICT opens new possibilities for automatic control of the electricity network, automatic maintenance of circuit breakers or switches in a grid and monitoring streaming of electricity data in real time. ICT applied in various companies could be used for making decisions by management. This can be achieved using and processing Big Data.

In our work have been discussed the characteristics and sources of Big Data, too. Moreover, the relationship between the Data Warehouse, OLAP and Data Mining has been presented visually and discussed. After that, the paper describes the tools that are used for the treatment of Big Data, especially for the Revolution R Enterprise tool. The RRE is used to analyze Big Data, to calculate statistical parameters and to visualize the data in order to get valuable information about electricity load and data distribution. We have written scripts in RRE in order to get the relationship between electricity load and time based on the regression model. At the end, we have made the evaluation of the model and conclusions.

## 2. THE FEATURES AND SOURCES OF BIG DATA

While the Big Data are considered as data that increased every moment, with different formats and collected mostly in real time, then the question arises: where does the Big Data come from?

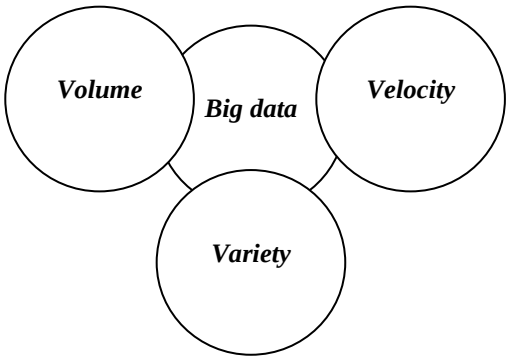



*Fig. 1. Big Data and three Vs*

There are many sources that generate Big Data. On the web, there is social media: Facebook, Twitter, Snapchat, or LinkedIn. Also, there are videos uploaded in YouTube, documents uploaded in the digital libraries or in the Cloud. Other

sources thanks to "Internet of things" are satellites, sensors, mobile telecommunications, RFID, weather observations, energy, bank transactions, location tracking, GPS, web application tracking, etc.

Sensor data from smart electric meters, videos, car sensors, etc., and machine log data: click data, event logs, mobile locations, etc., present the Big Data in the best way, having three characteristics of 3Vs (Volume, Velocity, and Variety) in the highest levels as in Figure 1, above. Also, from public web and social media, there are many collected data with 3Vs. However, the collected data from the archives, and social media in most cases are unstructured data. Nevertheless, business applications and data storage contain structured data.

## 3. DATA WAREHOUSE, OLAP AND DATA MINING

Data Warehouse (DW) is the central repository that stores large amounts of data by integrating the data from different systems. DW stores Big Data and integrates various exponentially growing data based on size, velocity, and variety. The purpose of DW as a central repository of data is to store data and to enable easy access to reporting and analyze. DW integrates data from multiple heterogeneous sources and transforms them into a multidimensional presentation for applications that support decisions [10]. DW has more to do with the preservation of historical data, so it does not preserve the actual data of the transaction. That means it is separated from the operational database known as Online Transaction Processing (OLTP). DW peculiarity is that the data cannot be changed, but they can be stored and accessed through the queries.

The results in OLAP usually appear as matrix OLAP. OLAP tools through multi-dimensional queries that they do in DW obtain an adequate response by allowing us to see the data from different angles. Usually, these queries are made to aggregate the data from multiple tables and in this way it has been earned a result. As an example is to get all the sales of a market for the first quarter and second quarter, and then finding what is the forecast for sales in the third quarter. Data Mining (DM) solutions serve to find new hidden data, such as trends, clustering, behavior, and other models. DM and OLAP look the same, but in fact, they differ from the way of manipulation of the data. Moreover, DM will know more about relationships and patterns in a large number of data.

OLAP tool allows data to be divided into many parts and then to find their resultant. This means that the aggregate operation is used more for OLAP. However, for the DM is used separation or division. An example of use of OLAP is to find the total number of expenditure of electricity in a country, for a certain time interval. While, in this case, DM gives us the factors that impact to the expenditure of this energy. OLAP creates the opportunity of multidimensional analysis through the cubes, called Hypercube by using complicated queries. While DM through its techniques and methods, requires the hidden relationships between

numerous historical data, in order to find any trend or pattern of them. OLAP and DM use data of Data warehouses for research and analysis, but the OLAP deals more with the question "what" or "how", while DM deals with the question "what and why" or "how and why". A better view of the difference between DW, OLAP and DM could be seen from Figure 2.

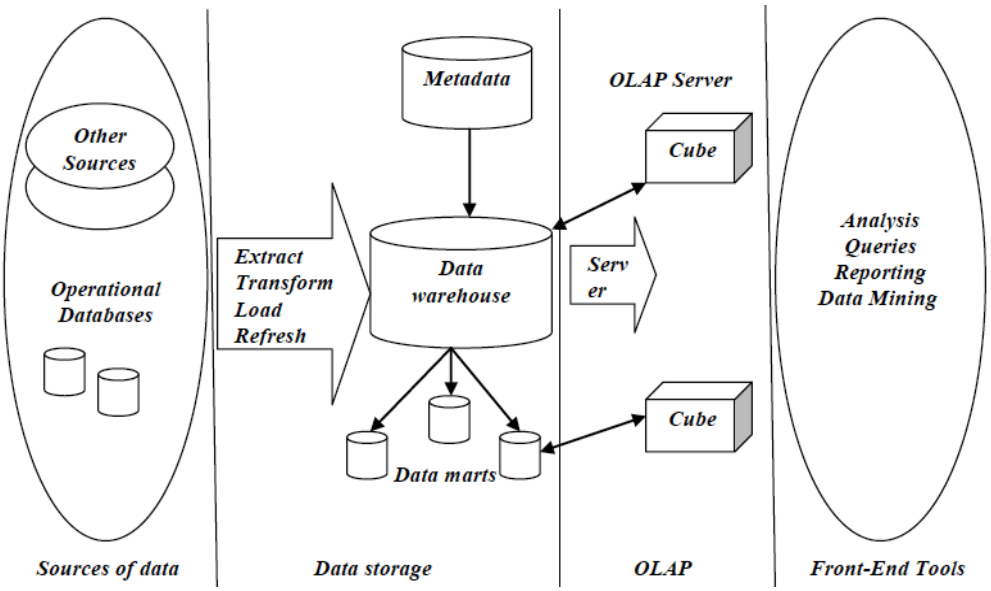



*Fig. 2. Three- Tier architecture*

The figure above shows that there are three tiers or bands, and it presents in which tier, or column is Data Mining, OLAP or Data Warehouse:

The first column shows sources of data, which can be operational databases-OLTP and other digital or media sources. The middle column has to do with the storage of data: Data Warehouse, or Data Marts. In addition, there are relational and multidimensional OLAP presented by the cubes. The last column has to do with the analysis, queries or with report tools and techniques of Data Mining.

## 4. USED DATA

From the Figure 3, it can be seen that an Electric Power System has three components: the Generation (Power Plant, Hydro Plant, Solar Panels, or other forms of producing the electricity), Transmission Operator (TSO) and the Distribution part.

The transformers as parts of Substations are responsible for the conversion of voltage, step-up and step-down or vice versa, depending on the level of

transmission. The distribution component delivers electricity to the end users, or customers, which can be expressed as the load. Most of the systems for transmission of electricity are based on three phase technology, which means using three conductors to carry electricity. This can be expressed as sinusoidal voltage waveforms.

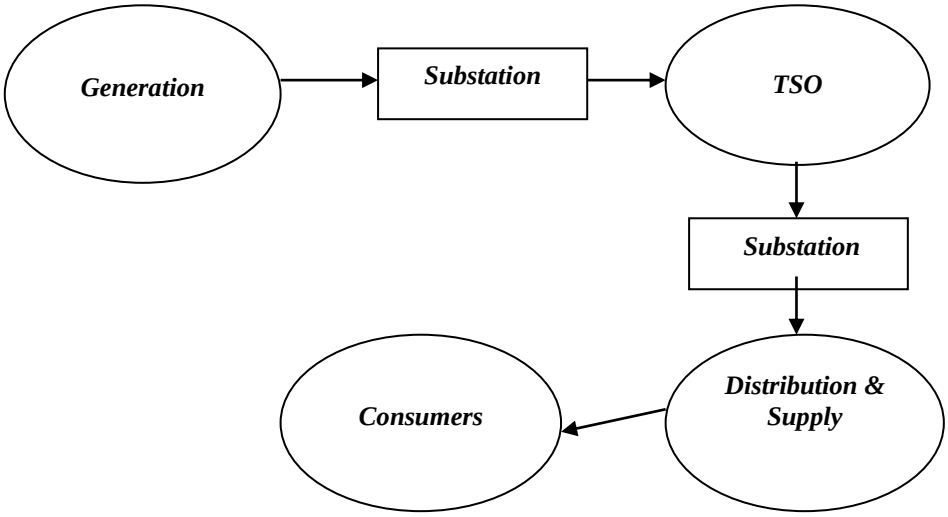



*Fig. 3. A schema of Power Systems*

In this article have been used the data from Load profiles of the electric grid in Kosovo. The Load profile is a comma-separated (CSV) file, which contains various data of electricity, such is max power, reactive power, on/off switching or disconnecting and other data. These data are real based on measurements from the smart metering in every 15 minutes. Smart meters are installed on the boundary between TSO and Distribution parts based on the Automatic Meter Reading (AMR) system. So, AMR is used as a source of Big Data in electricity, because collects, stores and transmits data of Electric Power, Reactive Power, load consumption etc., to the central system. Load profiles contain a population of data of 54 Substations for about 450.000 consumers. In our treatment, we have taken a sample of data which presents the Power of electricity and the date of measurements.

Big Data can be treated by various tools. So, the Hadoop, or MangoDB is used on database platform. The Amazon S3 is used for storage basis, but there are tools such as R language, Oracle framework, Revolution R Enterprise etc., which are based on data processing. The R language is a very comfortable tool for data analyzing, visualization and modeling. But, it has limitations in processing of data, especially when they are too large. The R tool uses the main memory for data

processing, so it is limited by this memory size. The Revolution R Enterprise (RRE) is an extension of the R language. RRE tool has increased the speed and the power of data processing, because it is designed to work in multi-threaded computations. It uses all cores and processors to handle the data. Thus, it has better performance during Big Data treatment. Our data are treated using RRE tool.

First of all the data in comma separated files have been imported in RRE, coding the appropriate script for this operation. The data are imported and must be converted into a data frame format, because this format is specific to RRE tool and facilitates the manipulations with data. Electricity data and dates are in the format of the character, so are used POSIXct () and POSIXlt () functions to transform the character format in the data frame.

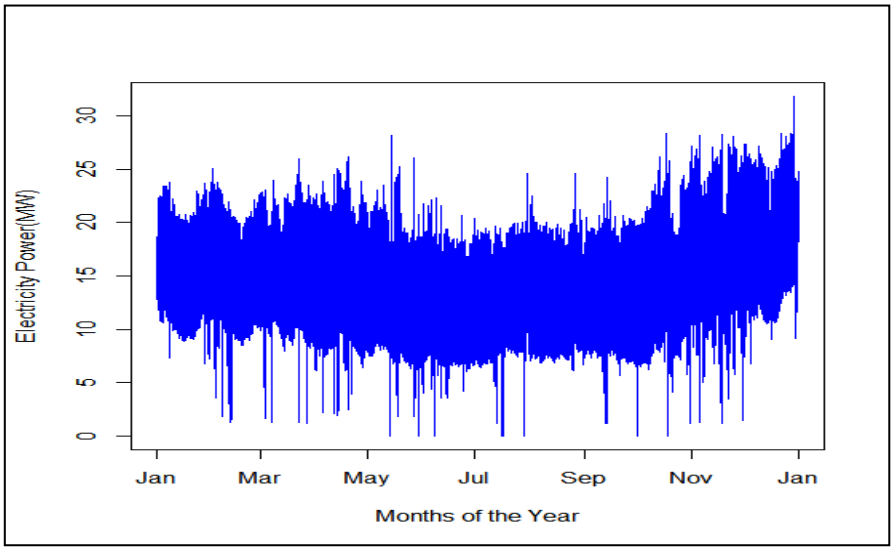



*Fig. 4. Data dissemination during a year*

From Figure 4 can be seen how much is the electrical load over time. The load is greater during the winter months. The maximum was reached in January. However, it can be seen that at the time from May until October the expenditure of energy is smaller. The minimum value of electric power measured in substations is 4.5 MW and the maximum is 29 MW.

Writing scripts in the editor of RRE, it is possible to obtain a visual view for the histogram and for the density frequency. The Figure 5 above shows that the distribution of the electric power is not normal, because most of the data are stretched in the band between 18 and 22. But, there is another stretching of the data in the band between 10 and 13. It looks like a bimodal distribution. This distribution is based on range, bin= 25, which is default by RRE tool.

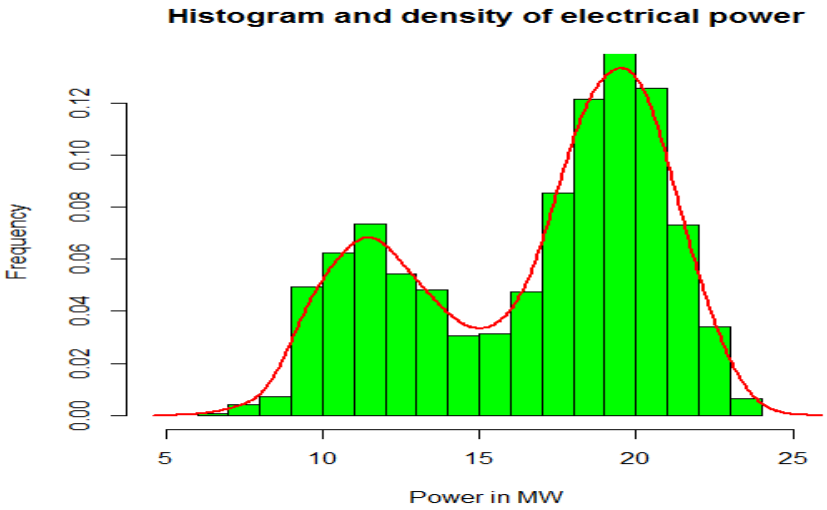



*Fig. 5. Histogram and frequency density of data in Load profile*

To make more normal this distribution, it is required to do smoothing of the range. That means that the value of the bin must be changed. In [9] it is presented the mixed distribution, but in our article have been calculated the parameters of the normal distribution, which are shown in Table 1.

*Table 1. Calculated parameters*

| ***Quantile()*** | ***0%*** | ***25 %*** | ***50 %*** | | ***75 %*** | ***100 %*** |
|---|---|---|---|---|---|---|
| ***Summary()*** | ***Min.*** | ***1st Qu.*** | ***Median*** | ***Mean*** | ***3rd Qu.*** | ***Max.*** |
| *Values* | *6.7* | *12.9* | *18.04* | *16.7* | *19.9* | *23.7* |

The mean is 16.7 which is smaller than median=18.1. This fact shows how the distribution is skewed to the left. The standard deviation is calculated based on the formula of the Gaussian distribution and the value is 3.89. The central tendency of data distribution is now completed. The InterQuartile Range (IQR) presented in Table 1, has the value IQR = 3rd Quartile – 1st Quartile =19.9-12.9=7. The data can be considered as Outliers if they extend outside the whiskers, in the case when the maximum and minimum values are not greater than 1.5 X IQR.

Writing the script in the RRE, it could be visualized the expenditure of electricity like in Figure 6, where is presented the monthly average load of three Substations during a year.

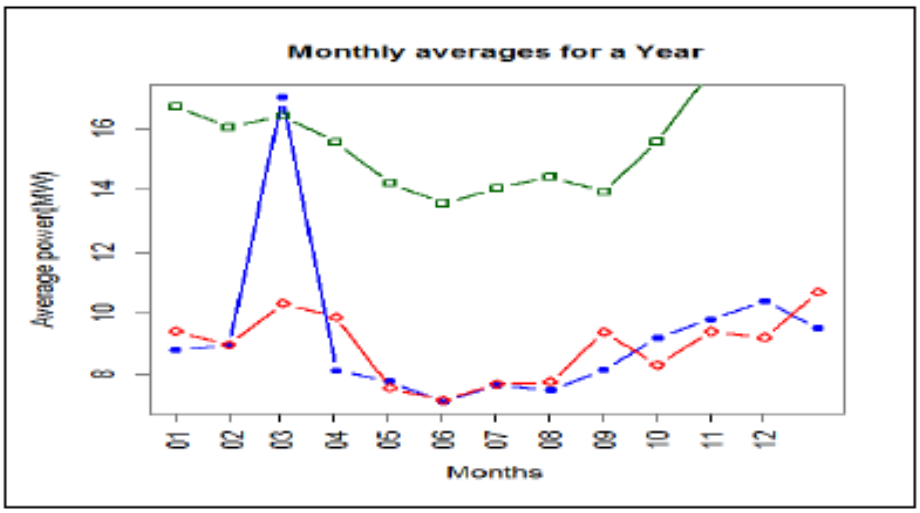



*Fig. 6. Monthly averages of loading*

First Substation has a max load about 16.5MW and min load about 13.8 MW. The second and the third substation have almost an equivalent load. However, the second substation has a higher load in three months February, Mart and April. If we analyze the data, it can be found that for these three months in this substation have been connected three industrial customers who cause this higher load. The number of customers for the first substation is about 12000, but the second and the third substation have an equal number of customers about 5000 each of them. This is the reason why the loading is higher in the first substation.

## 5. REGRESSION MODEL AND EVALUATION OF THE MODEL

Writing and executing scripts on RRE, we have achieved to visualize the distribution of the data and calculate interception, and the slope of the regression model [8]. Data distribution and the best-fit line are presented in Figure 7. Furthermore, the figure shows the linear line of the regression model based on the equation 1.

$$\begin{aligned} y &= f(x) = \alpha + \beta^{*}.x + \varepsilon \\ P &= load = 3.491e+02 - 2.598e-07 * t \end{aligned} \quad (1)$$

The values of Intercept and Slope are calculated using RRE based on the observational data of Load profiles in order to minimize the errors: $\varepsilon_i = Y_i - (a_0 + a_i X_i)$ from 2.

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [Y_i - (a_0 + a_1.X_i)]^2 \quad (2)$$

where, $Y_i$ is the electric power calculated in function of time, in a moment *i*.

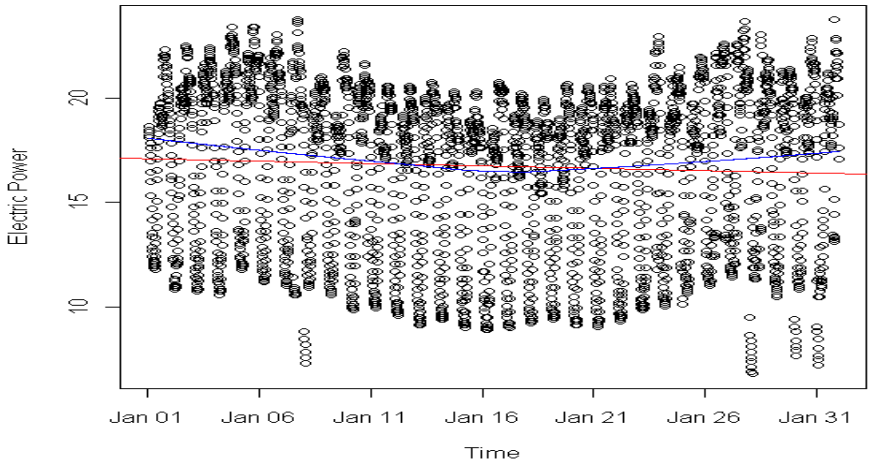



*Fig. 7. Best fit line in regression model*

We have used the method of Least Square Error-LEE to minimize the errors. In addition, it is important to minimize the sum of LEE from 3, which is known as residuals.

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left[ Y_i - \hat{Y}_i \right]^2 \tag{3}$$

Where, $\hat{Y}_i$ is the predicted value of electric power in the function of time, calculated in a moment i.

The Figure 7 shows that the best-fit line is a decreasing function. Moreover, it can be seen that the consumed energy is almost linear in January.

For evaluation of the regression model, we have used parameters *of p-value, degree of freedom, adjusted R-squared and confidence interval*. From the scripts in the RRE is found out that the *p-value* is 0.006, which is smaller than 0.05. This indicates that variables in the regression model have a good relationship, although the *adjusted R-squared* is 0.002. This is not the best value to represent the correlation between variables. However, a better value is found out for the *degree of freedom* = 0.035. So, this value meets the requirements of the model.

The degree of freedom from 4 explains explicitly the square root of the ratio between the residual standard error and a degree of freedom.

$$\sqrt{\frac{residual_standard_error}{degree_of_freedom}} = \sqrt{\frac{3.983}{2974}} \tag{4}$$

The first graph in Figure 8 above shows that residual's distributions seem to have a normal trend of distribution points, with a little increase for small values of

fitting values and with a little increase for greater values of fitting values. So, in this case, it breaks or falls down the homogeneity assumption.

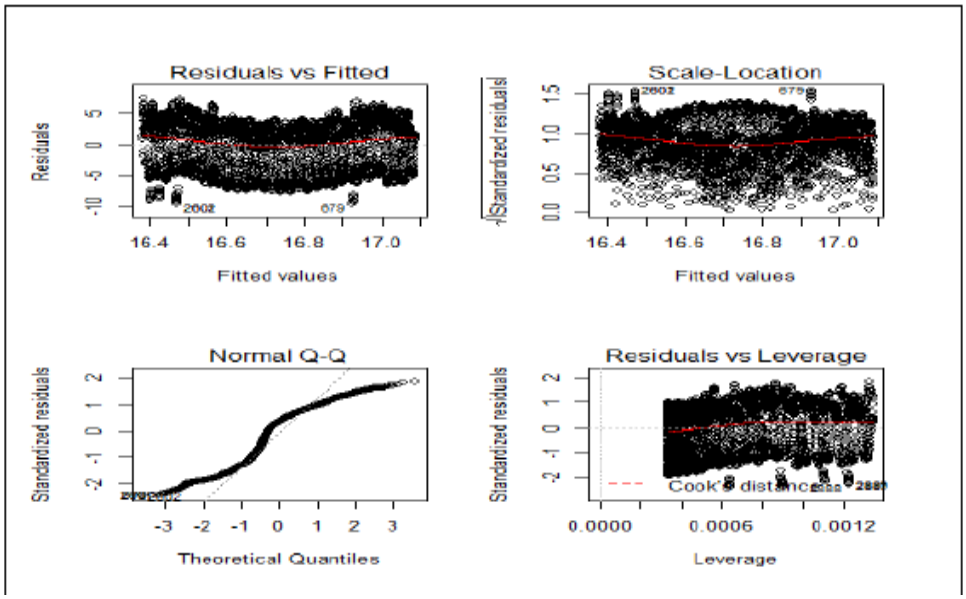



*Fig. 8. Evaluation of regression model*

The QQ (quantile-quantile) graph shows a nonlinear dispersion and shows that the residual errors are not normally distributed. The last graph shows the residuals against leverage calculated by Cook's distance. For as much as the values of the residuals are smaller than one, it can be deleted because their impact is very low in the model [7].

Writing the script in RRE, we have calculated the confidence interval for the prediction of the electrical load based on regression model 1. The confidence interval shows the possibilities range in which falls the mean when the sample will be repeated. In this case, the confidence interval gives the range of prediction, and the range value is 95%.

## 6. CONCLUSION

At the beginning of this article, it was discussed about Big Data and their importance in the industry. It was discussed for the characteristics of Big Data, how they are stored in Data warehouse and how it looks the relationship between DW, OLAP and Data Mining in the three-tier architecture.

For treatment of Big Data, we have used the RRE tool, which is very powerful to analyze, visualize and evaluate data models. Writing scripts and running them in

RRE, we have found the relationship between electric power and times. This relationship is expressed and modelled through the regression line, in order to meet the criteria of minimization of the sum of Least Square Error. The adjusted $R^2$ gives a value which is low, but the p-value, the degree of freedom and the confidence interval have the values that argue this relationship.In the article, also is made a descriptive statistic, using the RRE tool allowing us to see that data distribution is skewed to the left.

The aim of the paper was to find a model for the expenditure of electricity in order to apply it to the Big Data in the electric companies.

For the future, we recommend adding other variables to the model such as temperature or humidity, in order to calculate parameters and to compare those models. Based on various data of Load Profiles measured in electricity, we recommend to provide a repository system like Data warehouse, to manage better with Big Data.

## REFERENCES

[1] White paper. Big Data - Challenges and opportunities for the energy industry. *ALIGNE Energy trading, risk & operations*, SunGard, USA, 2013.

[2] Silipo, R., Winters, P. Time Series Prediction of Smart Energy Data. Big Data, *Smart Energy, and Predictive Analytics*, Knime, 2013.

[3] Jun, S., Liu, M., Lee, S., Hicks, J., Ankcorn, J., King, M., Xu, S., Arvind. BlueDBM: An Appliance for Big Data Analytics. *ACM- ISCA '15, June 13 - 17*, 2015, Portland, OR, USA, 2015.

[4] Ryohei, F., Satoshi, M. The Most Advanced Data Mining of the Big Data Era. *NEC Technical Journal- Big Data*, (Vol.7 No.2), 2012.

[5] Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T.M., Seliya, N., Wald, R., Muharemagic, E. Deep learning applications and challenges in big data analytics. *Journal of Big Data-Springer Open Journal*, 2:1, 2015.

[6] Kang, D.J., Park, S. A Conceptual Approach to Data Visualization for User Interface Design of Smart Grid Operation Tools. *International Journal of Energy, Information and Communications*,Vol. 1, Issue 1, November, 2010.

[7] Keka, I., Çiço, B. Big data in electricity-visualization aspect. *Proceedings of the 16th International Conference on Computer Systems and Technologies, CompSysTech*, Dublin, Ireland, June 25 - 26, 2015. ACM 2015.

[8] Keka, I., Hamiti, M. Load Profile analyses using R language. *IEEE - Information Technology Interfaces (ITI), Proc. of the ITI 2013 35th International Conference* on 24-27 June 2013, Cavtat, pp.240-250.

[9] Keka, I., Çiço, B. Statistical treatment for trend detection and analyzing of electrical load using programming language R. *IEEE - Proceedings of Embedded Computing (MECO), 2015 4th Mediterranean Conference on 14-18 June 2015*, pp.216 - 219, Budva.

[10] Satyanarayana Reddy,G., Srinivasu, R., Poorna Chander Rao, M., & Rikkula, S.R. Data Warehousing, Data Mining, OLAP and OLTP Technologies are essential elements to support decision making process in Industries. *(IJCSE) International Journal on Computer Science and Engineering*. 2010.

### *Information about the authors:*

**Ilir Keka** – Affiliation: PhD Candidat, Faculty of Contemporary Sciences and Technologies, South East European University (SEEU), Tetovo, Macedonia, Phone: +377 44 602 721, E-mail: ilir_keka@hotmail.com

**Betim Çiço** – Affiliation: South East European University, Faculty of Contemporary Sciences and Technologies, Tetovo, Macedonia. Professor in Computer Engineering. Areas of scientific research: Digital Logic Design, Computer Architecture, Cloud Computing, etc. Phone +389 44 356 176, E-mail: b.cico@seeu.edu.mk

**Manuscript received on 15 January 2016**
**Revised version received on 21 March 2016**