IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)
e-ISSN: 2278-2834,p- ISSN: 2278-8735.Volume 9, Issue 1, Ver. III (Jan. 2014), PP 88-96
www.iosrjournals.org

Application of Methods for Data Analysis and Predictive Modeling Scenarios in Mobile Communications Systems

Orlando Gea[1], Julio Cesar R. Dal Bello[2] and Eduardo Rodrigues Vale[3]
[1,2,3](Telecommunications Department, School of Engineering / Federal Fluminense University, Brazil)

Abstract: *The Strategic Management consists of the planning and the risky analysis of the activities involved on the decision process. Once the possible settings are a result of the variability of the parameter trends involved in the phenomenon, the risk is unavoidable. These concepts make the strategic management more efficient, because they support the theoretical basis of the theme. This work is based on the "Science of Data Analysis" and its comprehension allows a better basis in order to make up a decision through the creation of knowledge, which comes from the information obtained from the available network data. This work is devoted to data analysis methods and results of its application in the technical and operational data of a GSM mobile communication network used as a case study . The use of these techniques is possible in other mobile networks such as Wi-Fi (Wireless Fidelity), WiMAX (Worldwide Interoperability for Microwave Access), UMTS (Universal Mobile Telecommunication System), LTE (Long Term Evolution) and NGN (Next Generation Networks), among others. The results obtained allow the research of possible future scenarios for the development of effective strategic management. The authors are not aware of similar work published in the open technical literature.*

Keywords: *Mobile Communications Systems, Methods of Decision Support, Data Analysis, Predictive Models*

I. Introduction

Strategic management is the planning and risk analysis activities involved in decision process. Once the possible scenarios resulting from the variability of the trends of the parameters involved in the phenomenon, the existence of risk is inevitable. At the time that the study of the elements of strategic management - planning and risk - is modeled by the analysis of network parameters there is availability of indicators that provide information and knowledge. Strategic management in the mobile segment occurs when these indicators derived from the analysis of the data allow the construction of methods to support decision based on predictive models, since they help in the visualization of possible scenarios for the evaluation of operational actions carried out in the sector. In this work the results obtained by the application of methods for data analysis in the technical and operational information of a mobile GSM communication network are presented and discussed.

This paper is structured as follows: the item II presents considerations about strategic management; in the item III are shown the results obtained by application of methods for data analysis in a technical and operational information of a typical mobile GSM communication network; the item III shows the results of time series analysis performed and the item IV presents the conclusions of this work.

II. Strategic Management

2.1 Application

The treatment of the data obtained from the technical and operational information systems of a mobile communication network enables the acquisition of differential knowledge and important indicators about the network constituents. These indicators allow the formulation of a strategic planning to get the best performance of the desired business through the factors of production (technology and management) [1]. A strategic planning presupposes the need for a continuous decision-making process for the company to have conditions and means to act on variables and factors to exert some influence on them. This can be achieved by using a basic system.

2.2 Basic System and the Decision Process

The composition of the basic system mentioned in the previous item which serves to the purpose of this work consists of the union of the following elements [2]:

- ***Input***: statistical data.
- ***System***: processes for modeling.
- ***Output***: scenario development and performance indicators.
- ***Feedback:*** adjustments to update the model with the variability of the behavior of the external and internal environments.

A continuous decision-making process for the company to have conditions and means to act on variables and factors used by authors in this work is shown in Fig. 1.

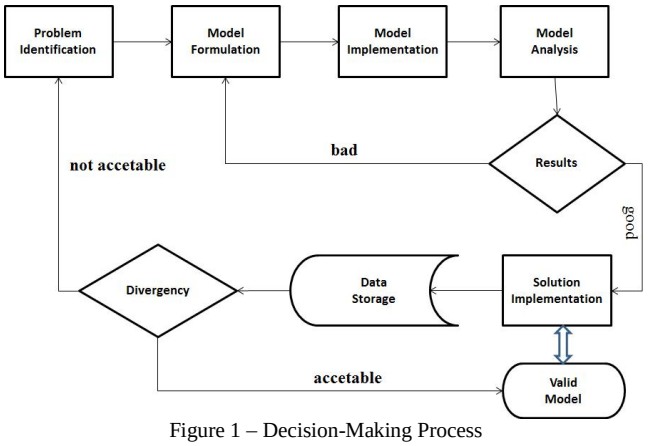


Figure 1 – Decision-Making Process

The process illustrated in Figure 1 consists of the following stages:

- ***Model Formulation*** selects the modeling technique that fits the ***Problem Identified***;
- ***Model Implementation*** consists of the application of the parameters used by the model;
- ***Model Analysis*** generates and evaluates alternatives that can lead to the solution of the problem;
- ***Results*** tests the viability and quality of each potential solution and the ***Validity of the Model***;
- ***Data Storage*** contains the history of the results obtained; and
- ***Divergency*** tests the model results compared to known results and the validity of the model with time.

The application of the above decision-making process avoids human error in the judgment due to the effects of anchoring and structure associated with decision problems. The effects of anchoring arise when a seemingly trivial factor serves as a starting point (or anchor) for estimates on a problem of decision analysis. It is important to note that the good decisions do not always result in good results due to the risks involved.

2.3 Risk Management

In the area of exact sciences, as engineering, the risk can be defined as the product of the probability of an undesirable event occurs and the estimated loss, if it occurs. The risk depends on the amplitude of the perceived impact of phenomenon or activity [3].

The measurement of the risk is possible by the *Scenario Analysis, Decision Trees and Simulation*, that is the most appropriate for this work. The use of *Simulation* enables the examination of the risks continuously and not discretely, as occurs in the first two.

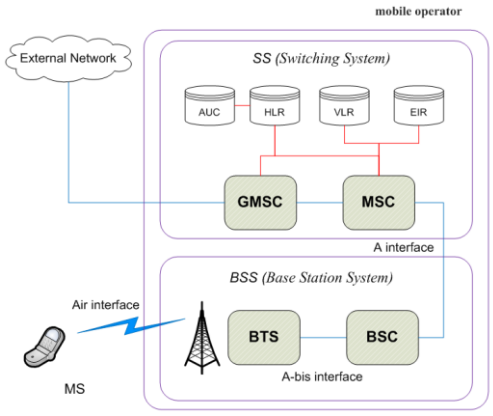
The knowledge of the structure of the information available and turning them into knowledge through Data Analysis helps the maximum exploitation of resources and the minimization of the possible risks involved. In the field of telecommunications, this understanding is essential for professionals in the area of decision because they need to direct limited resources to meet a growing demand for sophisticated systems.

2.4 Data Analysis

There are several techniques for Data Analysis. Some of them are subjectives and others objectives, as the used in this work. The applications of these techniques must be made with care and under theoretical basis to avoid errors and misapplication. The next item begins with a brief discussion of some concepts of the GSM structure, necessary for understanding the text, and among the possible operational indicators are presented some results obtained from the data analysis of three logical GSM channels. It will be evident the challenge of designing a strategic management for the GSM network.

III. Case Study: GSM Channel

In this item, the results obtained by the application of methods for data analysis in the technical and operational information of a typical mobile GSM communication network shown in Fig. 2 are presented and discussed. The logical structure of the network consists of BTS connected to the BSC system, controlled by a MSC system. The numerical data used in this study were obtained from a secondary source of a mobile operator placed in Rio de Janeiro State, as shown in Fig. 3. The primary source (BTS), sends the data collected to a data collector. This collector performs an initial compilation of the data in order to change the variables into technical ones that represent the performance and quality indicators used by the operator. The categorized data are from a primary source, the operator itself, as it defines the attributes according to its own criteria.



Components:

- Mobile Services Switching Center (MSC)
- Home Location Register (HLR)
- Visitor Location Register (VLR)
- Authentication Center (AUC)
- Equipment Identity Register (EIR)
- Base Station Controller (BSC)
- Base Transceiver Station (BTS)
- Gateway MSC

Figure 2 – GSM mobile communication network

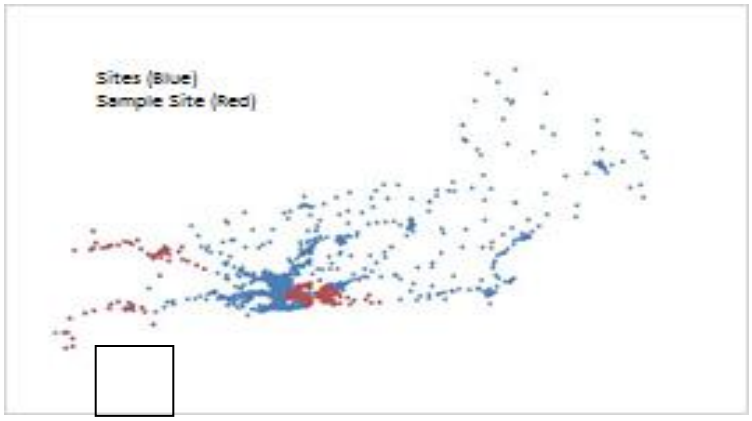


Figure 3 – Distribution of the mobile network sites sampled - in red - Rio de Janeiro State
(The square shows the position of Cluster 1 controlled by the BSC 1)

Among the possible operational indicators [5,6], in this item are presented some results obtained from the analysis of three logical GSM channels of the sites sampled shown in red in Fig. 3:

- Control Channel *SDCCH (Stand alone Dedicated Control Channel)*
- Traffic Channel *TCH* (*TCH Full Rate*)
- Traffic Channel *TCH/2* (*TCH Half Rate*)

The region shown in Fig. 3 is comprised of sites distributed in four areas of control (MSC) that contain six subareas of control (BSC) and a total of 21 clusters.

For this work, the choice of elements of the GSM network was defined by the operator of the system, who also provided the variables (attributes of the elements) as follow:

- Five tables containing daily data of five BSC sites during November 2009. Each daily record consists of date and time of the variables of the channels measured.
- One table containing the daily data from three sites of another BSC during May to November 2009.
- Eight tables containing information (logical and physical) of the network.

3.1 Control Graphic

The graphics shown in in Figs 4 to 7 for the cluster 1 and the descriptive analysis shown in Table 1 represent the operational profile of the channel traffic and are the "Control Graphics" for the operator which will use them in order to monitor and optimize the network. The graphs of traffic volume (SDCCH, TCH and TCH/2) for all six BSCs were carefully analyzed by the authors [1].

A Control Graphic is a manner to monitor variations in the characteristics of a process. Besides offering data visualization, its main focus is an attempt to separate the *"special causes of variation"* (identifiable) from the *"common causes of variations"* (random). The distinction between the two causes of variation is crucial, since special causes of variation are considered to be those that are not part of a process and are subject to correction

without modifying the system, while the common causes of variation can only be reduced by modifying the system.

For a better diagnosis of the variations and its causes, it is recommended to establish three reference lines on the Control Graphics.

- Central Line: average of the interval
- Upper Limit Control = process mean + 3 standard deviations
- Lower Limit Control = process mean - 3 standard deviations

The SDCCH, TCH and TCH/2 variables are determined and configured for each sector of the site by the Company. Consequently, the total availability and actual availability of these channels may differ from one sector to another between sites as they do not necessarily have the same configuration. Because of this, in this work a variable called beta (β) of the channel was defined, which is the percentage of the ratio between the effective channel availability in the sector and the overall availability of this site. The β variable quantifies the actual rate of system availability at the time of measurement. Thus, it eliminates misinterpretations in assessing absolute values whose baselines may differ. The beta (β) of the channel shows, at the time of measurement, the condition of "effectively available" since for various reasons - interference, noise and other - not always the full availability is effective for reporting.

The Fig. 4 represents the profile for the SDCCH, TCH and TCH/2 channels in the Cluster 1 of the BSC 1 with the β parameters of the channels.

Figure 4 – Traffic data of SDCCH, TCH and TCH/2 Channels in the Cluster 1 of the BSC 1 (November 2009)

The descriptive analysis of the SDCCH, TCH and TCH/2 channels in the Cluster 1 of the BSC 1 (Fig.1) are shown in Table 1 and illustrated by the histograms of Figures 5, 6 and 7 respectively.

Table 1 - Descriptive Analysis of the SDCCH, TCH and TCH/2 Channels in the Cluster 1 of the BSC 1

VARIABLE	SDCCH	TCH	TCH/2
Elements in the sample	6990	6990	6990
Arithmetic Average	1.72	3.92	6.75
Maximum Value	28.95	27.95	59.96
Minimum Value	0.00	0.00	0.00
Median	1.24	3.16	4.39
Mode	0.00	0.00	0.00
Average of the Interval	14.47	13.98	29.98
Average Joints	1.37	3.43	5.42
First Quartile	0.67	1.65	1.16
Third Quartile	2.06	5.20	9.69
Amplitude	26.95	27.95	59.96

VARIABLE	SDCCH	TCH	TCH/2
Interquartile Amplitude	1.39	3.55	8.52
Relative Variance	1.12	0.78	1.31
Absolute Variance	3.33	12.02	59.70
Standard Deviation	1.82	3.47	7.72
Variation Coefficient	106.04%	88.47%	114.47%
First Pearson Coefficient	0.943	1130	0.874
Second Pearson Coefficient	0.788	0.661	0.917
First Fisher Coefficient	3540	2412	2.296
Second Fisher Coefficient	21396	9132	7.915
Percentile Coefficient	0.190	0.277	0.240
Possible Outliers	3.79%	1.65%	2.47%
Probable Outliers	1.66%	0.53%	0.93%

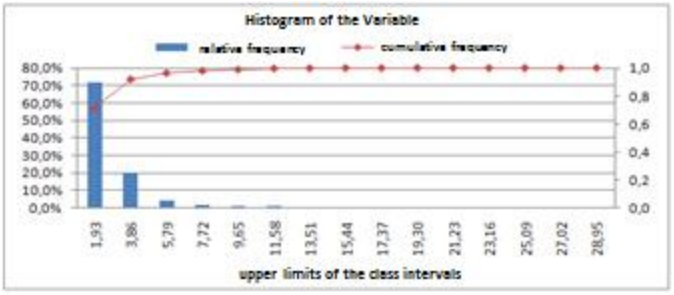


Figure 5 – Histogram of the SDCCH channel

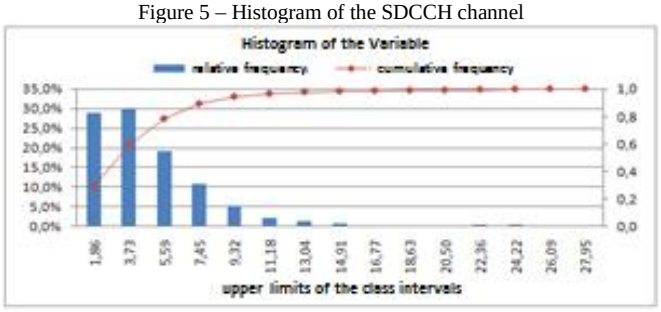


Figure 6 – Histogram of the channel TCH

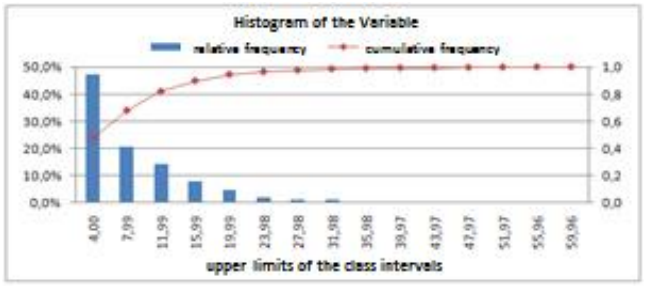


Figure 7 – Histogram of the TCH/2 channel

The analysis of Fig. 4 to 7 and Table 1 within the time period (November 2009) shows that the variable TCH/2 channel presents a greater dispersion around the average than the TCH and SDCCH channels due to its higher standard deviation. The TCH/2 channel presents a more heterogeneous distribution, with the highest *Variation Coefficient* (106.04%) and TCH presents the more homogeneous distribution with the lowest *Variation Coefficient* (88.47%). The three channels have few O*utliers* (1.66%, 0.53% and 0.93%, respectively). The first Pearson coefficient measures the distance, in number of standard deviations, from the average to the mode, while the second Pearson coefficient measures the distance, in number of standard deviations, from the average to the median. The presence of zero as *Mode* value for all channels can be understood as measurement errors.

The same analysis was carefully and exhaustively performed by the authors [1] for all six BSCs focused in this work.

3.2 Data Modeling

Multivariate analysis has an important role in strategic management, because the amount of data that can be extracted from the system can be high and this can lead to difficulty in determining the inter-relationships between variables and the definition of the most appropriate model [7,8,9].

The objectives of the multivariate analysis refer to:

- data reduction to structural simplification;
- selection or grouping;
- investigation of the existence of dependence between variables;
- forecasting and
- construction and testing of hypotheses.

3.3 Predictive Analysis

In predictive analysis the following models are used [10,11]:

- Prescriptive models: mathematical formulation known and defined, the values of the independent variables are known or are under control of the decision maker.

- Predictive models: mathematical formulation unknown and defined, the values of the independent variables are known or are under the control of the decision maker.
- Descriptive models: mathematical formulation known and defined, the values of the independent variables are unknown or uncertain.

The predictive model represents a class of decision problems in order to predict or estimate what value the dependent variable will assume as a function of the independent variables. However, the mathematical model to be used may be unknown and therefore must be estimated so that the decision maker can predict the dependent variable. The predictive models used in this work are Discriminant Analysis, Regression Analysis and Time Series Analysis.

A. Discriminant Analysis

The objective of the Discriminant Analysis is to create a rule to predict which group one observation belongs based on the values that the independent variables assume. Analysis for *Coverage* and *Quality* sites of clusters 4 and 18 are shown in Fig. 8, where the pairs of SDCCH and TCH traffic are examined (x = sdcch_traffic; y = tch_hr_traffic + tch_fr_traffic) for 1,860 points (cluster 4) and 1,677 points (cluster 18). It is observed that the groups of points of *Coverage* and *Quality* in cluster 4 and 8 are scattered and there is no clear distinction between the points of the groups. Thus, it is not possible a description of groups using the monitored traffics.

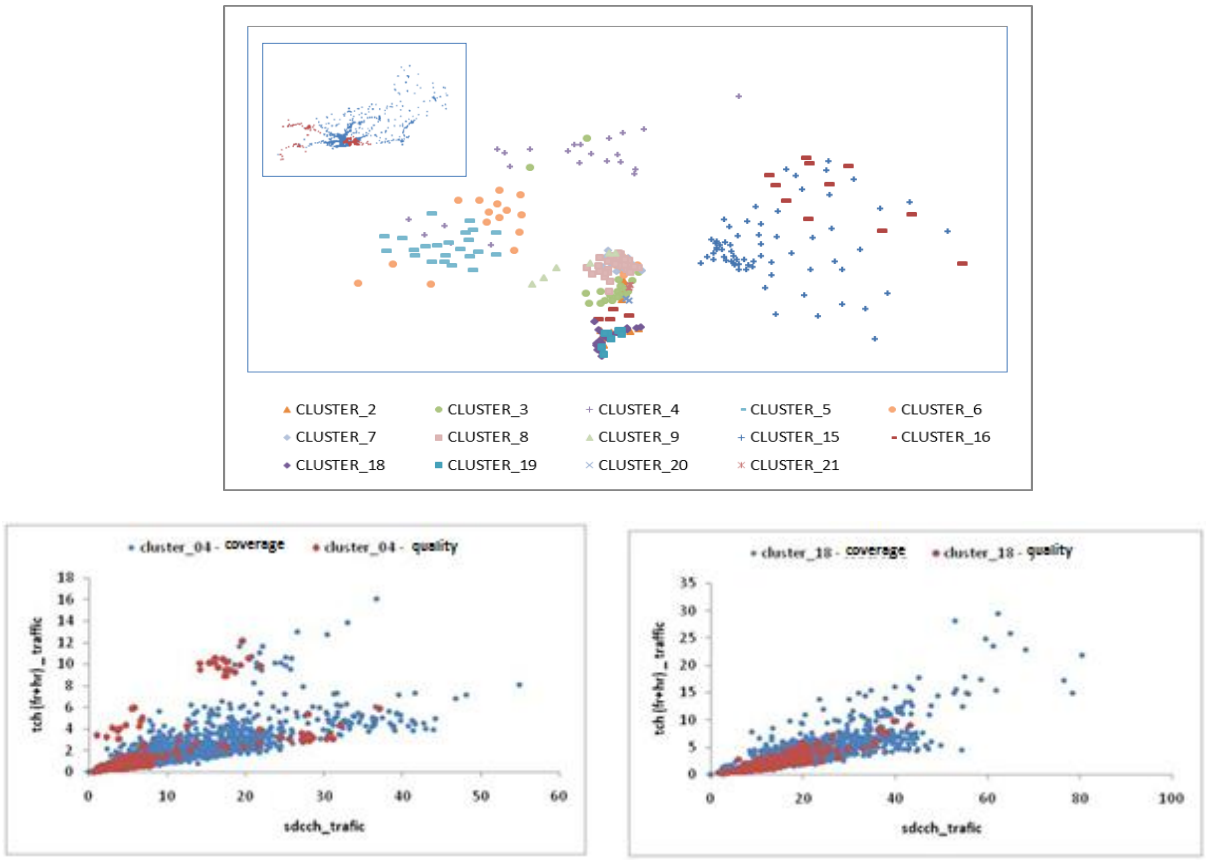


Figure 8 – Analysis for Coverage and Quality sites of clusters 4 and 18

B. Regression Analysis

Regression analysis is used mainly for the purpose of statistical development of a standard between a dependent and independent variables. The model for the analysis of the relationship between a continuous dependent variable Y and one or more independent variables X, aims to identify the function that describes the behavior of these variables

The linear regression used in this work is determined by the *Least Squares Method (min SQE)*, which will be determined by the regression coefficients (b_0 and b_1), used as estimates of their population parameters (β_0 and β_1):

$$\min(SQE) = \min\left\{\sum_{i=1}^{n}(Y_i - \hat{y}_i)^2\right\} = \min\left\{\sum_{i=1}^{n}\left[Y_i - (b_0 + b_1 x_i)\right]^2\right\}$$

$$(Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i) \longrightarrow (\hat{y}_i = b_0 + b_1 x_i) \quad (1)$$

According to this technique, the determination of the coefficients is possible when they minimize the difference between the actual value and the value predicted by the regression - estimation error (residues). Minimizing the sum of squared errors results in Equation (2):

$$b_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^{\,2} - n\bar{X}^{\,2}} \qquad b_0 = \bar{Y} - b_1 \bar{X} \quad (2)$$

Where:

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} \qquad \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

The Residue is the difference between the observed value and the predicted value. The adequacy of the fitted model by the absence of apparent pattern it is possible to be verified in a Residue Graph. The Student residues are adjusted for the average X (independent variables).

The independence of the residues is typically violated when the data is collected over sequential time periods, since a residue at any time may tend to be identical to residues at adjacent points in time - standard called autocorrelation.

The Fig 9 and 10 shows the results obtained with the linear regression model applied in a site placed in Rio de Janeiro City.

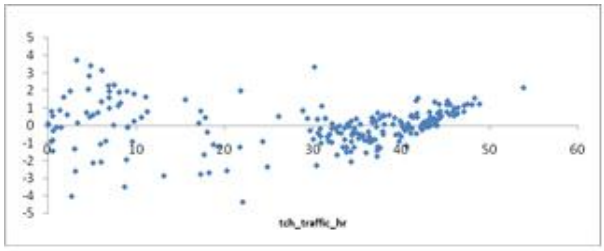


Figure 9 – Residue Graph (Student)

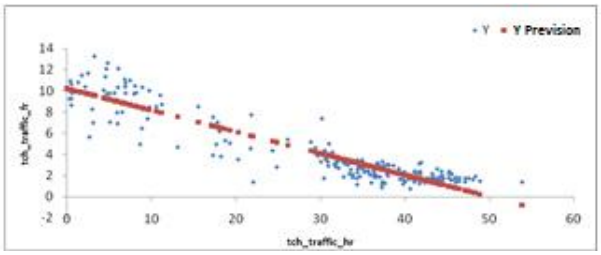


Figure 10 – Linear regression

For the graph of Fig. 9 the residues (Student) falls within a horizontal band centered at zero with a concentration only in the end. The Fig. 10 shows a straight line fit (linear regression) on the dispersion between TCH e TCH/2 channels. We note that the dispersion has two distinct regions with a concentration of points on the line that could distort the results by the influence of outliers in these groups. This fact is clearly seen in the graph of Fig. 9 where the assumption of normality is violated. The correction can be performed using multiple regression and nonlinear models of heteroscedasticity.

The same analysis was performed by the authors [1] for several other sites focused in this work.

C. Time Series Analysis

Time series is a set of observations collected in a quantitative variable over time as shown in Fig 12 for a site placed in Rio de Janeiro city. The past behavior of a variable allows the construction of a model to predict its future behavior. This prediction helps in planning future operational needs as it provides ways to know how these changes cause effects on operations [10, 11].

Exponential fit is a technique for stationary data that allows weights to be assigned to past data. Weighting occurs in descending order from newest to oldest observation. This technique is shown in Fig. 13.

Autoregressive modeling for adjustment and trend forecasting is a technique where the values of a data series, at certain points in time, are closely related with the values that precede them and those that follow. A first-order autocorrelation refers to the magnitude of the association between consecutive values in a time series, while the second order refers to the magnitude of the relation between values in two separate periods, as shown in Fig. 14.

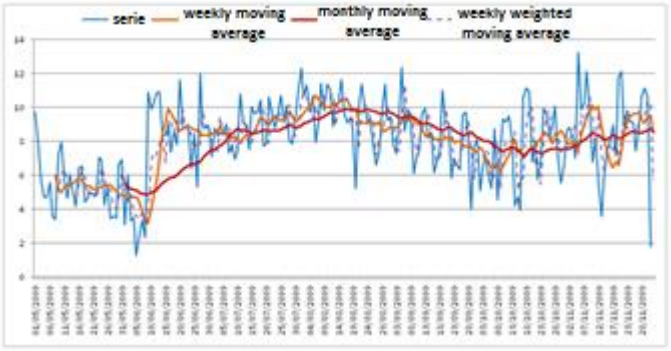


Figure 12 - Time series with moving average

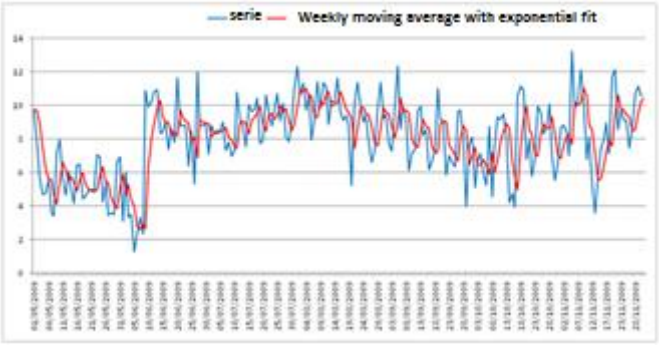


Figure 13 - Time series with exponential fit

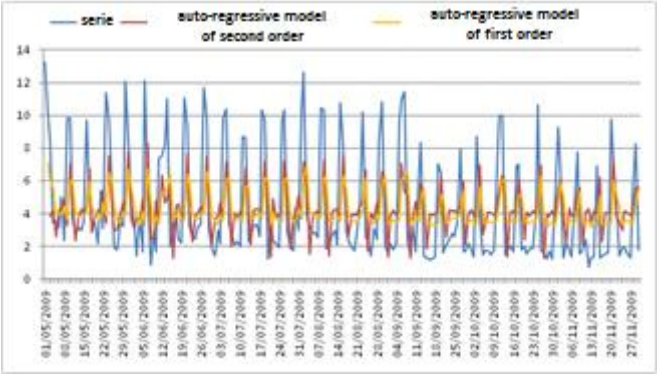


Figure 14 - Time series regression model with 1st order and 2nd order

IV. Conclusion

The objective of this work is the application of techniques of data analysis in the setting of cellular networks to assist in the decision-making process in relation to the dynamic update and expansion of the installed structure. Mobile cellular networks undergo continuous changes requiring that the companies must be prepared to predict future scenarios, transforming obstacles into opportunities sources. The technologically competitive companies have constant demand for investments in the infrastructure.

The science of data analysis was presented in this work within a dynamic and difficult context for modeling: a mobile telecommunications network. The knowledge obtained and discussed here can be used in the optimization and planning of the network structure modification, minimizing the impacts. In this context, it becomes a decisive factor for organizations the ability to realize the strategic implications of the choices to be made (technological, economic, managerial or any other).

This work is devoted to data analysis methods and results of its application in the technical and operational data of a GSM mobile communication network used as a case study. The results allow the elaboration of possible future scenarios for the development of effective strategic management. The authors are not aware of similar work published in the open technical literature.

For purposes of future works the authors suggest the use of these techniques in other mobile systems such as Wi-Fi (Wireless Fidelity), WiMAX (Worldwide Interoperability for Microwave Access), UMTS (Universal Mobile Telecommunication System), LTE (Long Term Evolution) and NGN (Next Generation Networks).

References

[1] O. Gea, *Application of Methods for Data Analysis on Mobile Communication Systems* (Brazil: MSc Thesis - Federal Fluminense University, 2010).

[2] A. Damodaran, *Gestão Estratégica do Risco: uma referência para a tomada de decisão de riscos empresariais* (Brazil: 1st Edition, Bookman, 2009).

[3] D. Oliveira, *Planejamento Estratégico - Conceitos, Metodologia e Práticas* (Brazil: 6th Edition, Atlas, 1992).

[4] L. P. Fávero, *Análise de Dados: modelagem multivariada para tomada de decisões* (Brazil: 1st Edition, Elsevier, 2009).

[5] L. Jukka and M. Matti, *Radio Interface System Planning for GSM/GPRS/UMTS* (New York: 1st Edition, Klumer Academic Publishers, 2002).

[6] M. D. Yacoub, *Foundations of Mobile Radio Engineering* (Florida : CRC Press, 1993).

[7] C. Ragsdale, *Modelagem e Análise de Decisão* (Brazil: 2nd Edition, Cengage Learning, 2009).

[8] M. C. Goldbarg and H. P. L. Luna , *Otimização Combinatória e Programação Linear* (Brazil: 1st Edition, Campos, 2000).

[9] M. P. E. Lins and G. M. Calôba, *Programação Linear - com aplicações em teoria dos jogos e avaliação de desempenho (data envelopment analysis)* (Brazil: 1st Edition, Interciência Ltda., 2006).

[10] D. N. Gujarati, *Econometria Básica* (Brazil: 3th Edition, Makron Books Ltda, 2000).

[11] R. C. Hill, G. G. Judge and W. E. Griffits, *Econometria* (Brazil: 2nd Edition Saraiva, 2003).