

A methodology for statistical modeling of water losses and seepage in hydrotechnical objects

Milovan Milivojević*, Srdjan Obradović*, Slobodan Radovanović**,***, Boban Stojanović****, Nikola Milivojević***

* Technical and Business College, Uzice, Serbia

** University of Belgrade, Faculty of Civil Engineering

*** Jaroslav Černi Institute, Belgrade, Serbia

**** University of Kragujevac, Faculty of Science, Kragujevac, Serbia

milovan.milivojevic@vpts.edu.rs, srdjan.obradovic@outlook.com, sradovanovic@grf.bg.ac.rs, bobi@kg.ac.rs, nikola.milivojevic@gmail.com

Abstract—For a long time, novel analytical models have been applied in modeling, prediction and monitoring of water losses and seepage in hydrotechnical objects (HO). Statistical models based on multiple linear regression (MLR) have been shown to be more or less successful for modeling processes in many domains, but have not been a common choice for modeling the above mentioned processes. The rapid improvement of sensor and data acquisition technologies enables the development of statistical models in basic and advanced MLR forms, such as hierarchical regression, stepwise regression (SWR), robust regression, and ridge regression. This paper presents a framework for the application of advanced statistical tools in modeling and evaluation of water losses and seepage in hydrotechnical objects, with the focus on stepwise regressions. The developed framework provides a platform for software generation of adequate regression models of seepage and water losses, considering both the number, and the type of regressors.

I. INTRODUCTION

The creation of mathematical models for water losses and seepage processes in hydraulic objects (HO) is of great importance concerning HO functionality and safety. In the last few decades analytical models, which describe the above mentioned processes, have been developed. The use of statistical models has not been significant although they are characterized by the simplicity of formulation, the speed of execution and the availability of any type of correlation between independent and response variables. To the best of our knowledge, statistical models in the multiple linear regression (MLR) form have not been the common approach for modeling water losses and seepage processes in HO.

Based on the review of current literature we found that there are few studies in this domain of research. In the field of modeling water losses and seepage processes in HO, a number of models based on experimental methods with resulting analytical solutions have been shown to be more or less successful [1, 2]. In [1], an experimental method is used for modeling water losses in a hydraulic tunnel - an analytical model was established and the relationship between water losses and the main influences upon which they depend, such as: internal water pressure in the tunnel, groundwater pressure in the area above the tunnel and tunnel lining temperature. In [2], analytical

solutions based on conformal mapping of the complex variable methods are derived for two-dimensional, steady seepage into an underwater circular tunnel.

Models that are based on numerical methods, such as the finite element method, have been developed as well. In [3], an interlaced algorithm based on the finite element method was suggested to solve the coupled processes of non-steady seepage flow and non-linear deformation for a concrete-faced rockfill dam. In [4], stress-strain analysis, underground water flow influence analysis and filtration analysis were performed for the hydrotechnical tunnel in the phase of excavation.

Recently, numerical and statistical methods have been enriched with various heuristics from the artificial intelligence (AI) domain, creating hybrid models that combine their advantages. In [5], phreatic line detection, which is a major challenge in seepage problems, is accomplished with the use of Natural Element Method (NEM) and Genetic Algorithm (GA).

Although AI models have high performance, mathematical models in MLR forms are still very useful. The problem of determining the appropriate number and type of regressors in MLR models is still present. This open question preoccupies many researchers. In light of the above, we posed the following research question: Is it possible, at a sufficiently high quality level to automate the procedure for selection of the number and type of regressors, which describe the water losses and seepage processes in HO such as dams, hydraulic tunnels, etc.?

In general, the modern AI approach provides certain hybrid solutions for this problem. In [6], in the domain of structural behavior of concrete dams, authors have developed a method for model optimization in terms of model complexity and accuracy (regularization) [7], based on hybrid GA and MLR approach.

In this paper, for the purpose of solving the problem of regularization in MLR model for seepage and water losses processes in HO, we focused on the enhanced statistical technique of stepwise regression, mainly due to the high execution speed it offers.

II. THEORETICAL BACKGROUND

The following sections provide a brief outline of the employed theoretical base for the statistical modeling of water losses and seepage in hydrotechnical objects.

A. Principal Component Analysis

Besides data quantity, the number of variables and potential regressors also affects the complexity of the regression model and the required processing time for its generation. With this in mind, it is advisable to consider methods of reducing the potential complexity of the model using the technique of factor analysis, in the phase of data preprocessing. For this purpose, we utilized Principal Component Analysis (PCA).

Principal Component Analysis is a statistical approach that is utilized to analyze inter-relationships among a large number of variables and to describe these variables in terms of their common underlying dimensions (factors). The goal is to condense the information contained in a number of original variables into a smaller set of dimensions (components) with a minimal loss of information [8].

PCA is concerned either with the covariances or correlations between a set of observed variables x_1, x_2, \dots, x_q that can be explained in terms of a smaller number of unobservable latent variables or common factors, f_1, f_2, \dots, f_k , where $k < q$. In mathematical terms, the factor analysis model is expressed in (1):

$$\begin{aligned} x_1 &= \lambda_{11} \cdot f_1 + \lambda_{12} \cdot f_2 + \dots + \lambda_{1k} \cdot f_k + u_1, \\ &\dots \\ x_q &= \lambda_{q1} \cdot f_1 + \lambda_{q2} \cdot f_2 + \dots + \lambda_{qk} \cdot f_k + u_q, \end{aligned} \tag{1}$$

where λ_{ij} are the $q \times k$ factor loadings, and u_i are the residual terms, also known as specific variates [9]. The correlation matrix \mathbf{R} (2) gives the inter-correlations among the set of variables, which is the basis for dimension reduction:

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \dots & \dots & \dots & \dots \\ r_{p1} & r_{p2} & \dots & 1 \end{bmatrix} \tag{2}$$

The elements of the correlation matrix, $r_{ij} = r_{ji}$, represent Pearson's linear correlation coefficients, and the sum of the elements on the main diagonal is equal to the sum of eigenvalues and thus to the number of predictors. The eigenvalues of the correlation matrix, λ_i , also represent the variance of the principal components, and they are the roots of the \mathbf{R} matrix characteristic polynomial given by (3), where \mathbf{I} is the identity matrix.

$$\det(\mathbf{R} - \lambda \cdot \mathbf{I}) = k(\lambda) \tag{3}$$

The correlation matrix \mathbf{R} , can be transformed according to (4):

$$\mathbf{P}^{-1} \cdot \mathbf{R} \cdot \mathbf{P} = \mathbf{D} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \lambda_p \end{bmatrix} \tag{4}$$

where \mathbf{P} is the matrix whose columns are eigenvectors of matrix \mathbf{R} , and \mathbf{D} is the diagonal form matrix of \mathbf{R} . Matrix \mathbf{D} keeps the variability of the original matrix \mathbf{R} , now expressed through eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_p$. In order to simplify the system, only some of the principal components, which correspond to the chosen number of first k largest eigenvalues, $\lambda_1, \lambda_2, \dots, \lambda_k$, are kept without significant loss of variance of the original dataset.

PCA in detail is given by [9, 10].

B. Multiple Linear Regression

Linear regression models that comprise of more than one predictor variable are multiple linear regression models - MLR. The general form of an MLR, can be written as follows in (5) [6]:

$$y_i = \beta_0 + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + \dots + \beta_j \cdot x_{ji} + \dots + \varepsilon_i \tag{5}$$

where y_i is a response variable, x_{ji} are the predictor variables ($j=1,2,\dots,k$), k the number of significant predictors, the index i shows the sample number, and ε_i are the independent and normally distributed random variables that have a mean of zero and a variance σ^2 . In the majority of the applications of linear regression models, the functional forms of the predictors (basis functions) are not clear in advance and are dependent on the nature of the modeled phenomenon [6]. Coefficients, β_0, \dots, β_k are the unknown parameters of the model, which are estimated for a given set of data using the least squares method, which minimizes the sum of squared errors (SSE) in (6):

$$SSE = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - b_0 - \sum_{j=1}^k b_j \cdot x_{ji})^2 \tag{6}$$

by taking the derivatives of SSE with respect to the b 's and setting them equal to zero. In (5), the b 's are the estimated values of the model parameters.

C. Stepwise Regression

Stepwise regression is essentially based on semipartial correlation, which is expressed through the semipartial correlation coefficient (sr) and the square of this coefficient (sr^2). Semipartial correlation expresses changes in R^2 for one variable in one regression model transformation (it shows how much does each single variable uniquely contribute to the coefficient of determination R^2). Or, in other words, the square of the semipartial correlation coefficient, for a specific single variable, indicates by how much will the R^2 value be reduced if this single variable is removed from the regression equation.

Let χ be the set of all independent variables \mathbf{X} , and ψ_k be the set of all independent variables \mathbf{X} except for x_k . Then the squared semipartial correlation coefficient is expressed as follows in (7):

$$sr_k^2 = R_\chi^2 - R_{\psi_k}^2 \quad (7)$$

Therefore, in order to obtain the unique contribution of predictor x_k to the coefficient of determination of the regression model, R^2 , it is first necessary to regress the dependent variable y , based on all independent variables (χ), and then regress the dependent variable y on the basis of all variables except x_k (ψ_k). The difference in obtained R^2 values, represents squared semipartial correlation coefficients. The meaning of semipartial correlation coefficient can be expressed in different ways. One way is through the Ballantine chart in Fig. 1:

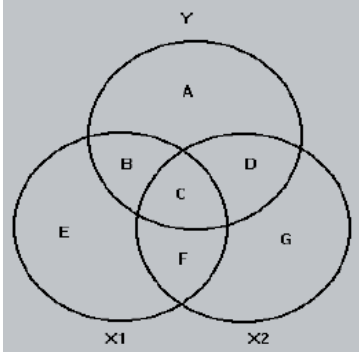


Figure 1. Ballantine chart – squared semi partial correlation coefficient (semipartial correlation)

In this chart the variance of each variable is represented by a circle of unit area (each variable is standardized to have its mean equal to 0 and its standard deviation equal to 1). Each overlapping area of two circles represents the square of their mutual correlation (e.g., $C + F = r_{12}^2$). The total area for variable Y , which is formed by circles X_1 and X_2 , represents the portion of total variance for variable Y which is caused by variables X_1 and X_2 (r_{Y12}^2). Areas B and D represent unique contributions of variables X_1 and X_2 to the variance of variable Y . Unique areas (B, D) correspond to the squared semipartial correlation coefficients ($B = sr_1^2, D = sr_2^2$) and represents the portion of variable Y variance, that is the amount for which the squared value of multiple correlation increases when one of the variables, X_1 or X_2 , is added to the pool of other independent variables.

Form of expressing semipartial correlation coefficients may vary, and one of the most common forms is (8):

$$sr_k = \frac{t_k \cdot \sqrt{1 - R_\chi^2}}{\sqrt{\text{residualDF}}} \quad (8)$$

where t_k is Student's t-statistic value for the k -th regressor in the MLR model, $\text{residualDF} = N - K - 1$, is the number of degrees of freedom for the sum of residuals,

N is number of measurements, and K is the number of regressors.

Stepwise regression is explained in detail in [11, 12].

III. MODELING METHODOLOGY

Development of statistical mathematical models of real objects and systems is a very demanding research and engineering activity that involves the use of subtle mathematical apparatus and making of a large number of decisions based on both the theoretical knowledge as well as empirical an evidence. Although it is impossible to develop turnkey solutions, it is possible to define critical steps, and an indicative algorithm to generate adequate regression models.

A. Proposed algorithm

For the realization of statistical modeling of water losses and seepage in hydrotechnical objects algorithm the given in Fig. 2 is proposed. Key features of the proposed algorithm are described in the following sections.

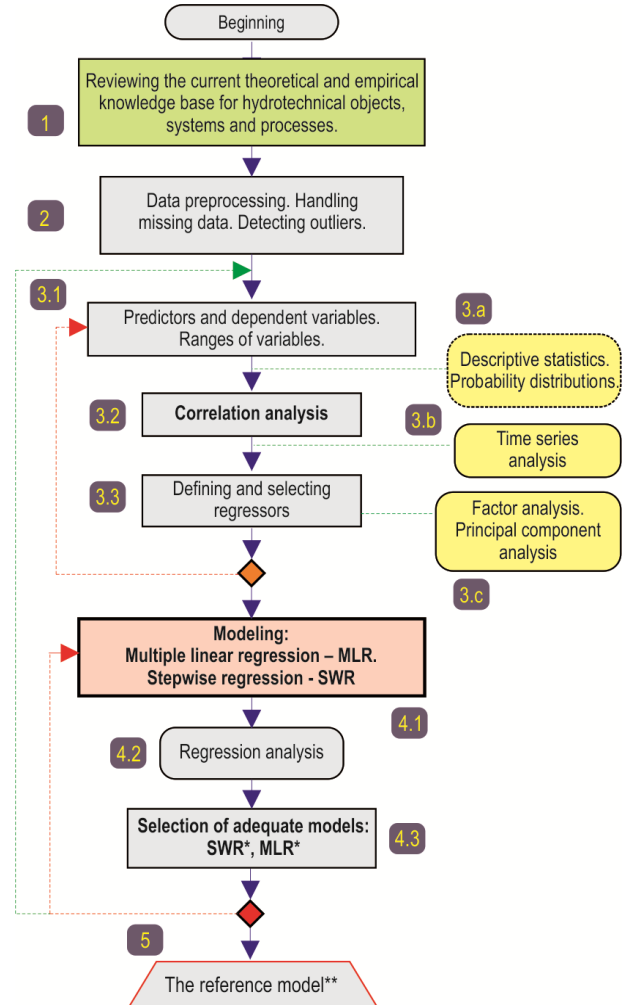


Figure 2. Algorithm for statistical modeling of water losses and seepage in hydrotechnical objects

B. Acquisition of domain knowledge

The initial phase in model creation is certainly acquisition of knowledge from the domain of hydraulics, fluid mechanics, stress and deformation, and also the

domains of applied statistics and computer science. This refers to the theoretical knowledge and laws, and also the experience of researchers and the results achieved in previous periods. Without well-grounded theoretical platform software tools can easily miss the target.

C. Handling missing data and outliers

Regression models are very sensitive to the occurrence of outliers, and the occurrence of missing data, so the process of anomaly detection in the data, or among the predictors, is of crucial importance.

For dealing with missing data, different strategies can be applied: linear interpolation, spline interpolation, regression interpolation, replacing missing data with mean values, moving averages, etc.

The problem of outlier detection and dealing with outliers is far more complex. For this purpose we recommend the use of the following packages developed in the R programming language environment [13]:

- *zoo* – provides various methods for replacing data, including cubic spline interpolation, linear missing interpolation, etc.
- *tsoutliers* – provides time series analysis and outlier detection methods based on an iterative outlier detection and adjustment procedure that obtains joint estimates of model parameters and outlier effects [14]. Five types of outliers are considered: innovational outliers, additive outliers, level shifts, temporary changes and seasonal level shifts.
- *forecast* – provides methods and tools for analyzing univariate time series.
- *AnomalyDetection* – provides time series outlier detection on the basis of the Seasonal Hybrid extreme studentized deviate (ESD) algorithm, based on [15].

An example of outlier detection for time series of water losses in a hydrotechnical object, is given in Fig. 3:

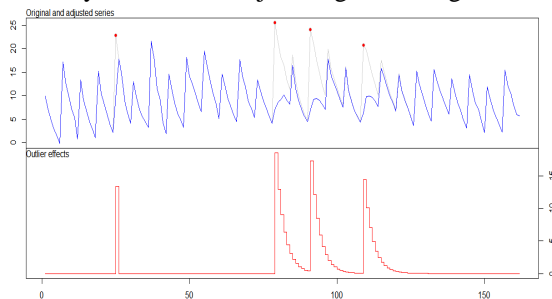


Figure 3. Potential outliers detected with the use of R package *tsoutliers*

Using the *tsoutliers* package *tso* function, four outliers were detected (red points). The plot also displays outlier effects, the original and suggested adjusted time series (red, grey, and blue, respectively).

D. Preparation for modeling and modeling

Generating high quality models involves proper selection of predictors and the dependent variable and the ranges of their values (Fig. 2 - step 3.1). In other words, the success of the methodology depends upon defining the multidimensional space (hyperspace), in which the model should be created. Correlation analysis (Fig. 2 - step 3.2)

is of great importance in this matter, because it provides a strong foundation for the proper selection of input and output variables, and their regressors. Regression models are very sensitive to multicollinearity and singularity, so the steps 3.2 and 3.3 are necessary to avoid the pitfalls of model overfitting (models with high accuracy, but small ability of generalization).

Furthermore, this complex issue can be supported by elements of descriptive statistics and probability distribution laws of physical values upon which water losses and seepage depend (Fig. 2 - step 3.a). Since the largest number of variables are dependent on a function of time, it is useful to use the tools for time series analysis (Fig. 2 - step 3.b). In this way, trends, cyclical and seasonal changes, can be observed, which should facilitate the modeling process.

It is also useful to support the correlation analysis with the paradigms of dimension reduction. In this context, the technique of factor analysis, Principal component analysis, is very effective (Fig. 2 - step 3.c). Its results provide an overview of supporting, influential factors that underlie the processes.

Often the methodological steps in the third stage (Fig. 2 - steps: 3.1, 3.2, 3.a, 3.b, 3.c) are implemented in several iterative steps.

E. Iterative generation and selection of models

As previously mentioned, a key issue in creating a model in the form of multiple regression is the selection of appropriate regressors, considering both the number and type of these regressors. To solve this problem, according to the proposed methodology, the techniques engaged were Stepwise regressions. These techniques in Fig. 2 - step 4.1, based on partial correlation coefficients, perform automated entering and removing of regressors, thus creating optimal regression models. At this stage, we defined the following critical steps:

- Generating a regression model using stepwise regression algorithm in multiple variants that are associated with different criteria of exclusion or inclusion of regressors from or in the regression model during the iterative procedure (stepwise, forward, backward, best subsets).
- Analysis of the resulting, reduced number of regressors.
- Analysis of regressor importance. To the extent that a predictor is important in the model, leaving it out of the model should produce a substantial increase in the residual sum of squares; to the extent that a predictor is not important in the model, leaving it out of the model should produce a minor increase in the residual sum of squares. A series of models is computed excluding each predictor in each successive model, record the sum of squares associated with the residuals for each model, adds $1/p$ to each residual where p is the total number of predictors, and then determines the ratio of each subtotal to the grand total (these ratios will sum to 1.00) [16].
- Selection of an acceptable model from the obtained variants. Choice is based on criteria of: accuracy (adjusted R^2 , root mean squared error, etc.), complexity, and the type of candidate regressors (Akaike information criterion, F-statistics, etc.)

- Testing the model by the forward stepwise algorithm to verify and to avoid multicollinearity.
- The final selection of the model within a session – choosing a model that meets the criteria of regularization and non multicollinearity - Variance Inflation Factor (VIF) and Tolerance [17]. Tolerance is an indicator of how much of the variability of the specified independent variable is not explained by the other independent variables in the model, and is calculated for each variable using the formula (9):

$$\textit{tolerance} = 1 - R^2 \quad (9)$$

where R^2 is the coefficient of determination. If this value is very small (less than .10) it indicates that the multiple correlation with other variables is high, suggesting the possibility of multicollinearity. VIF is the inverse of tolerance. VIF values above 10 indicate multicollinearity.

The quality of obtained models is measured by the adopted criteria such as: adjusted coefficient of determination, root mean squared error for the test dataset and the training dataset, Pearson product-moment correlation coefficient of modeled and measured values, and other criteria.

Detailed performance indicators of generated models such as confidence intervals for predicted mean values, significance and confidence intervals of regressors, and other indicators of interest, are produced by the tools of regression analysis (Fig. 2 – step 4.2).

Based on the above-mentioned criteria and the results of regression analysis, as well as model complexity analysis, in step Fig. 2 – step 4.3, the appropriate selection of MLR models for water losses and seepage is being made.

IV. CONCLUSION

Modern technology, sensors and measuring equipment, enable the collection of vast amounts of data on HO and processes that are the subject of modeling. The increase in the volume of information expands the knowledge about the observed object and allows the creation of advanced MLR models such as: hierarchical regression, stepwise multiple regression, robust regression, and partial least squares regression.

The problem of determining the appropriate number and type of regressors in a MLR model, in the domain of modeling water losses and seepage in HO, was addressed in this paper. Solutions offered so far, represent an attempt to balance between the fitting ability of regression models, and the ability of generalization (Occam's razor).

In order to contribute to solving the defined problem, the authors have created a methodology focused on stepwise regression, which is also supported with PCA (Principal Component Analysis), times series outliers detection, as well as correlation and regression analysis.

The methodology presented in this paper, allows the creation the reference MLR model for summary water losses and seepage in HO, with high performance and reliability.

ACKNOWLEDGMENT

The part of this research is supported by Ministry of Education and Science of Republic of Serbia. Project TR37013 (Development of a system for safety management of high dams in the Republic of Serbia).

REFERENCES

- [1] Andjelkovic V., Lazarevic Z., Nedovic V., Stojanovic Z., "Application of the pressure grouting in the hydraulic tunnels", Tunn. Undergr. Space Technol., vol. 37, 2013, pp. 165–179.
- [2] Huang, F.M., Wang, M.S., Tan, Z.S. and Wang, X.Y. (2010), "Analytical solutions for steady seepage into an underwater circular tunnel", Tunn. Undergr. Space Technol., vol. 25, pp. 391–396.
- [3] Chen, Y., Hu, R., Lu, W., Li, D., and Zhou, C. (2011). "Modeling coupled processes of non-steady seepage flow and non-linear deformation for a concrete-faced rockfill dam." Comput. Struct., vol. 89, 2011, pp. 1333–1351.
- [4] Radovanović S., Rakić D., Divac D., Živković M., *Stress-Strain Analysis and Global Stability of Tunnel Excavation*, 2nd International Conference for PhD students in Civil Engineering and Architecture „CE-PhD 2014“, 10-13 December 2014, Cluj-Napoca, Romania, Editor: Cosmin G. Chiorean, Publisher: Technical University of Cluj-Napoca, ISSN 2392-9715, pp. 248-255, 2014.
- [5] Shahrokhbadi S.H., Toufigh M.M., "The solution of unconfined seepage problem using Natural Element Method (NEM) coupled with Genetic Algorithm (GA)", Applied Mathematical Modelling, vol. 37(5), 2013, pp. 2775-2786.
- [6] Stojanovic B., Milivojevic M., Ivanovic M., Milivojevic N., Divac D., "Adaptive system for dam behavior modeling based on linear regression and genetic algorithms", Adv. Eng. Softw., vol. 65, 2013, pp. 182–90.
- [7] Russel S., Norvig P., "Artificial Intelligence: A modern Approach, 3rd edition", New York, Pearson Education, 2010.
- [8] Milivojevic M., Obradovic S. Kurcubic J. Djokovic K., "Application of radial basis function neural networks to prediction of raw water quality parameters", 9. Int. Conf. SED 2016, Uzice, Serbia, 30 Sep.-01 Oct., 2016. (2.81 – 2.94), ISBN 978-86-83573-82-02, COBIS.SR-ID 227527948
- [9] Der G., Everitt B. S., "A Handbook of Statistical Analyses using SAS", Third Edition, Chapman and Hall/CRC, 2008, pp. 300-302.
- [10] B. G. Tabachnick and F. S. Linda, *Using Multivariate Statistics* (5th Edition), Pearson, 2006, pp. 607–675.
- [11] Milivojevic M., Stopic S., Stojanovic B., Drndarevic D., Bernd F., "Forward stepwise regression in determining dimensions of forming and sizing tools for self-lubricated bearings", METTAL Internationale Fachzeitschrift fur metallurgie, April, 2013, vol. 67, pp. 147-153.
- [12] Cohen J., Cohen P., West S.G., Aiken L.S.. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, vol. 3rd Edition, Taylor & Francis, 2002.
- [13] R Development Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2012. ISBN 3-900051-07-0. <<http://www.R-project.org/>>.
- [14] Chen C., Liu L.M., "Joint estimation of model parameters and outlier effects in time series", Journal of the American Statistical Association, vol. 88, 1993, pp. 284–297.
- [15] Bernard R., "Percentage Points for a Generalized ESD Many-Outlier Procedure", Technometrics, vol. 25, No. 2, 1983, pp. 165-172.
- [16] Meyers L. S., Gamst G. C., Guarino A. J., "Performing data analysis using IBM SPSS", 1st edition, Hoboken, NJ: John Wiley, 2013, pp 209.
- [17] Pallant J., "A Step by Step Guide to Data Analysis Using SPSS for Windows (Version 15)", 3rd edition, Berkshire: Open University Press, McGraw-Hill Education, 2007, pp. 158.