*Journal of the Nigerian Association of Mathematical Physics*
*Volume 46* **(May,** *2018 Issue***), pp227 –232**


# MODELING RELATIONSHIP BETWEEN SOIL PROPERTIES AND YIELD COMPONENT OF PLANTAIN USING MULTIPLE REGRESSION AND STRUCTURAL EQUATION MODELING.

***T. A. Ojurongbe[1*], K. A. Bashiru[1]and S.O.S Akinyemi[2]***

**[1]Department of Mathematical Sciences College of Science, Osun State University, P.M.B 4494, Osogbo, Nigeria.**
**[2] Fruits and Biotechnology Research Unit, National Horticultural Research Institute, P.M.B 5432, Idi-Ishin Jericho, Ibadan.**

## *Abstract*

***Regression analysis is a useful multivariate technique for investigating and modelling the relationship between variables.Structural Equation Modelling (SEM) is a combination of factor analysis and multiple regression. Since soil properties influence the behavior of soils, the knowledge related to these properties is important in using them for different purposes. The aim of this work is to obtain an expression using multiple linear regression to evaluate relationship between the dependent variables(yield and Nitrogen), and the soil physical and chemical properties (pH, Orgcarbon, P, Ca, Mg, Mn, K, CEC, Sand, Silt, clay). SPSS 21.0 was used to carry out the analyses on multiple regression. Secondly, to develop a structural equation model of yield component on plantain (YCP) in some areas in Osun state and Oyo State, using soil chemical and physical properties, and path analysis was performed using LISREL 9.1. The main purposeof this study is to develop a conceptual model in order to determine the sources of variation within the dataset and to explore equations for the sampled soil. The result revealed thatsoil physical and chemical properties works were important in explaining YCP.***
***Considering the relative importance of the estimation of YCP and from the view of regression equation, magnesium and Cation exchange capacity (CEC) made the biggest contribution through the proposed models for the soil yield. According to the structural equation modeling (SEM) outcome, the final model proved that YCP was controlled by soil chemical properties more than the physical properties.***

## 1.0 INTRODUCTION

Regression analysis is a statistical process for estimating the relationship among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between dependent variables and one or more independent variables. The two basics types of regression are linear regression and multiple regression. Linear regression is an approach for modeling the relationship between a scalar dependent variable Y and one or more explanatory variables denoted by X. The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple regression. Linear regression is performed either to predict the response variable based on the predictor variables, or to study the relationship between the response variable and predictor variables which could be represented by the linear model: $Y = \beta_0 + \beta_1 + \varepsilon$, where $\beta_0$ $and$ $\beta_1$ are constants called, the model regression coefficients or parameters and $\varepsilon$ is a random disturbance error[1].

Corresponding Author**:** Ojurongbe T.A., Email: taiwo.ojurongbe@uniosun.edu.ng, Tel: +2348132617229

The earliest form of regression was the method of least squares, which was published by Legendre in 1805 and by Gauss in 1809. Legendre and Gauss both applied the method to the problem of determining, from astronomical observations, the orbits of bodies about the Sun (mostly comets, but also later the then newly discovered minor planets)[2]. Gauss published a further development of the theory of least squares in 1821, including a version of the GaussMarkus theorem.

On the other hand, Structural equation modeling (SEM) or path analysis is the statistical technique used to examine causal relationship between two or more variables. It is based upon a linear equation system and was developed by Sewall Wright in the 1930s for use in phylogenetic studies.[3].Path analysis was adopted by the social sciences in the 1960s and has been used in increasing frequency in the ecological literatures since the 1970s.In ecological studies, path analysis is used mainly in the attempt to understand comparative strengths of direct and indirect relationships among a set of variables[4]. In this way, path analysis is unique from other linear equation models: In path analysis mediated pathways (those acting through a mediating variable, which is "Y" in the pathway X-Y-Z) can be examined [5]. Pathways in path models represent hypotheses of researchers, and can never be statistically tested for directionality[6].

In this study, multiple regression will be carried out on several independent or predictor variables (i.e. the physical and chemical properties of some soil where plantain was planted in Oyo and Osun states) with yield and Nitrogen content of the soil (N) as dependent variables. All the variables would be included simultaneously into a single model in order to test the potential interaction between the independent variables using structural equation modeling. We decided to use plantain seeing that there's a higher demand for it and its consumption is on the increase worldwide [7].

## 2.0 MATERIALS AND METHOD

Data on plantain yield as well as chemical and physical properties of the soil on which the plantain was planted were collected from National Horticultural Research Institute (NIHORT), Idi-Ishin,Ibadan.

In this study, we want to investigate the relationship between soil physical and chemical properties and plantain yield in some locations in Osun and Oyo states, Nigeria.

Soil is the mixture of minerals, organic matters, gases, liquids and the myriad of organisms that together support plant life. Soil has different types of properties, but in this study, physical and chemical properties will be considered. Physical properties are those characteristics which can be seen with the eye or felt between the thumbs and fingers, while chemical properties cannot be determined just by viewing or touching the substance, the internal structure must be affected for its chemical properties to be investigated.

Multiple linear regressions (MLR) and structural equation modeling (SEM) were used to analyse the secondary data collected from NIHORT. The data were subjected to regression analysis whereby a regression model was generated for the dependent and independent variables. The contributions of each variable were given in the model equations.

Statistical analyses were carried out using LISREL 9.1[8]application and SPSS software version 21.0. MLR was performed on the data and the results were used to construct equations that explained the relationship between the different properties and the yield component of plantain and also the Nitrogen content of the soils. SEM was also summarized using structural equation modeling and multiple regression.

In general, the multiple regression case can be written as

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \cdots + \beta_k x_{tk} + u_k \quad \ldots (1)$$

Where the $\beta's$ are k unknown parameters, the u is the error or residual terms, t refers to the observation number, and $x_{ti}$refers to the $i^{th}$ independent variable for observation t.

The individual equations for k parameters and n observations are:

$$\begin{aligned} y_{1=} &\quad \beta_1 + \beta_2 x_{12} + \beta_3 x_{13} + \cdots + \beta_k x_{ik} + u_1 \\ y_2 &= \beta_1 + \beta_2 x_{22} + \beta_3 x_{23} + \ldots + \beta_k x_{2k} + u_2 \\ y_3 &= \beta_1 + \beta_2 x_{32} + \beta_3 x_{33} + \cdots + \beta_k x_{3k} + u_3 \quad \ldots (2) \\ \vdots &\qquad\qquad \vdots \\ y_n &= \beta_1 + \beta_2 x_{n2} + \beta_3 x_{n3} + \cdots + \beta_k x_{nk} + u_n \end{aligned}$$

The difference between the simple linear and multiple linear case is the complicating issue of additional independent variables, $x$'s. Using the procedure to derive the simple linear case, the derivation of the OLS estimator results in k equations. This is where matrix algebra enter the mix. Three matrix algebra operations are necessary, multiplication, transpose and inverse.

To write equation (2) in matrix form, four matrices must be defined, one for the dependent variables, one for the independent variables, one for the unknown parameters, and finally one for the error terms. These four matrices are:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ . \\ . \\ y_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ . \\ . \\ . \\ \beta_n \end{pmatrix}, \quad and \quad U = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \\ \\ \\ u_n \end{pmatrix} \qquad \ldots(3)$$

$$X = \begin{pmatrix} 1 & x_{12} & x_{13} & . & . & . & x_{1k} \\ 1 & x_{22} & x_{23} & . & . & . & x_{2k} \\ 1 & x_{32} & x_{33} & . & . & . & x_{3k} \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ 1 & x_{n2} & x_{n3} & . & . & . & x_{nk} \end{pmatrix} \qquad \ldots(4)$$

SEM is an extension of the general linear model (GLM) that enables one to test a set of regression equations simultaneously [9]. SEM software can test traditional models, but it also permits examination of more complex relationship and models, such as confirmatory factor analysis and time series analyses.

The basic approach to performing a SEM analysis is as follows:

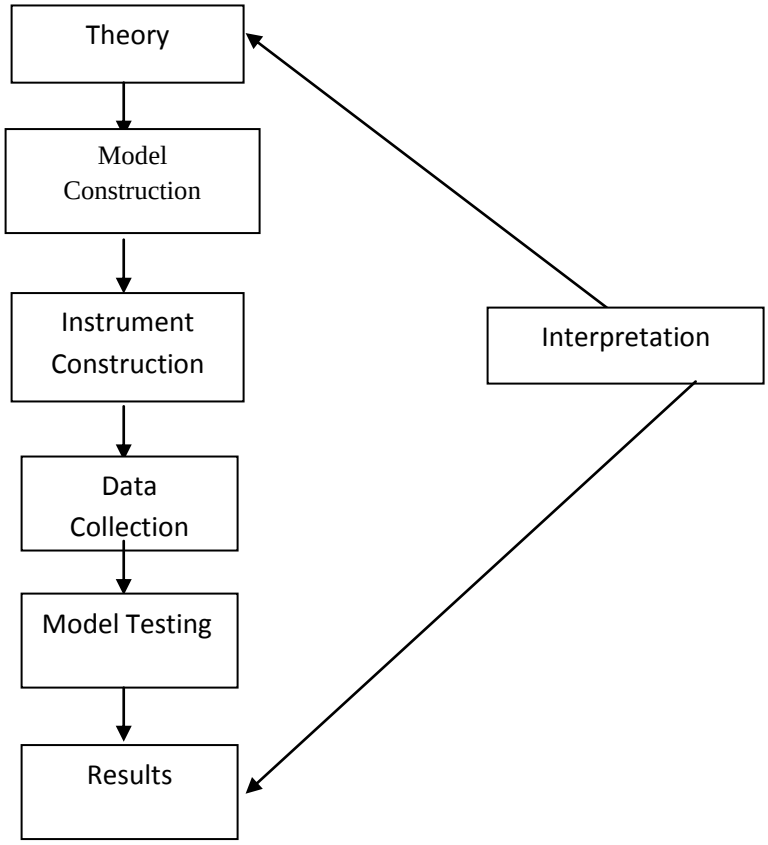


Figure 1: SEM Analysis

The contribution to the analysis is usually a covariance matrix of measured variables such as survey item scores, though sometimes matrices of correlations or matrices of covariance and means are used. In practice, the data analyst usually supplies SEM programs with raw data, and the programs convert these data into covariance and means for its own use[10].

The model consists of a set of relationships among the measured variables. These relationships are then expressed as restrictions on the total set of possible relationships[11].

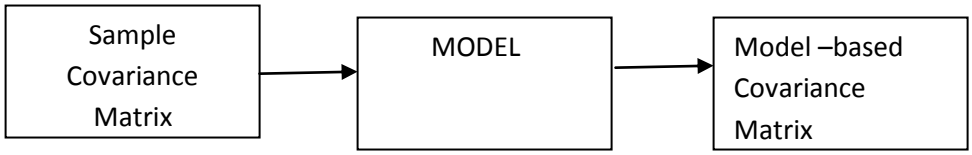


Figure 2: Set of Possible Relationships

## 3.0 RESULT AND DISCUSSION

MLR analyses was carried out on plantain data collected from NIHORT. The theoretical regression model includes a set of nine independent explanatory variables for chemical properties and three physical properties. MLR supplied equation connecting dependent variable (YCP) to the independent variables with the following form of equation:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip} + \varepsilon_i$$

$\beta_0$ isthe intercept (constant) and $\beta_1$, $\beta_2 \ldots . \beta_p$, are the regression coefficients. $X_i$ Descriptors are used to describe soil chemical and physical properties.

Regression analysis was conducted to investigate the relationship between the YCP and soil properties using SPSS v21. The dependent variable is known as YCP and the independent variables are measured to be PH, ORGCARBON, N, P, CA, MG, MN, K, CEC, SAND, SILT, CLAY. Considering yield as the dependent variable while performing the analysis, the result as well as the descriptors and the regression coefficients are presented in Table 1 and with nitrogen as the dependent variable, the result is presented in Table 2.

According to the result from the analysis, R square statistics is 36% for the total variation for the estimation of YCP was explained by the multiple linear regression models. In terms of the relative value of the estimation of a dependent variable, it fall out that the MAGNESIUM (MG) and CATION EXCHANGE (CEC) made the biggest contribution across the model. And the theory test of t values also revealed that the same factors contributed to the estimation of YCP.

When the dependent variable is yield, and the soil physical and chemical properties are PH, ORGCARBON, N, P, CA, MG, MN, K,CEC, SAND, SILT, CLAY, The selected equations for the soil physical and chemical properties with the coefficients are:

$$\text{YCP}= 17.152 - (0.965 * PH) + (0 \cdot 151 * ORGCARBON) - (1.266 * N) + (0 \cdot 62 * P) + -(0.433 * CA) + (1.438 * MG) + (0 \cdot 384 * MN) - (0 \cdot 313 * K) + (0 \cdot 537 * CEC) - (0 \cdot 0404 * SAND) - (O \cdot O130 * SILT) - (0 \cdot 064 * CLAY) \quad \ldots(5)$$

Also, When the dependent variable is Nitrogen, and the soil properties are PH, ORGCARBON, P, CA, MG, MN, K,CEC, SAND, SILT, CLAY and YIELD the selected equations with the coefficient are:

$$\text{NITROGEN}=2 \cdot 157 - (0 \cdot 35 * ph) + (0 \cdot 106 * orgcarbon) - (0 \cdot 008 * p) + (0 \cdot 002 * ca) - (0.177 * mg) + (0 \cdot 239 * mn) - (0 \cdot 084 * k) + (0 \cdot 011 * cec) - (0 \cdot 018 * sand) - (0.025 * silt) + (0.001 * clay) - (0.024 * yield) \quad \ldots(6)$$

Figure 3 presents the diagram representing the YCP and 11% of the YCP level achieved variance was explained by the diagram, the accepted model explained some amount of the variance in YCP ($R^2 = 0.360$), that is when the dependent variable is yield and ($R^2 = 0.974$) when the dependent variable is Nitrogen. The regression analysis performed to evaluate the plantain yield model resulted in equation (1) that soil properties prediction was significant for YCP ($p < 0.0001$, $R^2 = 0.360$ (yield as the dependent variable) and $R^2 = 0.974$ (Nitrogen as the dependent variable).

In terms of the relative value of the estimation of a dependent variable, it falls out that the MAGNESIUM (MG) (Path Coefficient = 0.38) and CATION EXCHANGE (CEC) (Path coefficient = 2.18) made the biggest contribution across the model.

Hypothesis testing by structural equation modeling includes all the variables in a conceptual model in order to test the possible communication between them. The SEM was used to achieve this goal and the unstandardized and standardized coefficients were used to evaluate the strength of path coefficient expected.

Figure 3 shows that the theoretical model adequately constructed the observed covariance of the data with all of the aforesaid indicators of good model fit exceeding the minimum specifications. The $X^2$ (chi-square) statistics was non-significant, indicating a good model fit to sample variance ($X^2 = 259.59$, d.f= 53, RMSEA = 0.349).

The chi-square ($X^2$) test, gave a value of 259.59, which had a related p-value < 0.0001, the ratio chi-square/degree of freedom yield 4.898 ($X^2$/df= 4.898), with a root mean square squared error of approximation (RMSEA) of 0.35. These index values showed that there is a strong correspondence between predicted and observed covariance. For RMSEA, a p-value greater than 0.05 indicated no significant deviation between observed and expected covariance. The path analysis result revealed Organic Carbon, CEC and Ca as the major contributors within the chemical properties and Sand as the major contributor within the physical properties.

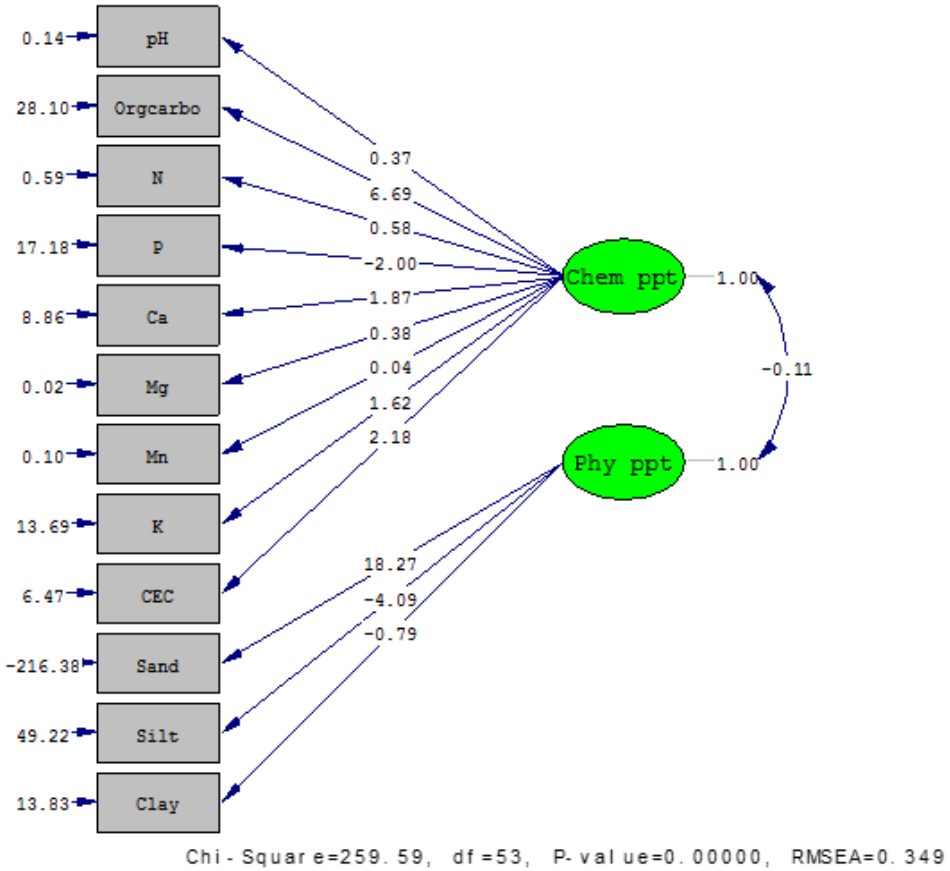


Figure3: Theoretical model

**TABLE 1: Regression analysis result with yield as the dependent variable**

| MODEL | Unstandardized coefficients | | Standardized coefficients | t | Sig. | 95.0 % confidence interval for B | |
|---|---|---|---|---|---|---|---|
| | B | Std error | Beta | | | Lower bound | Upper bound |
| Constant | 17.152 | 7.785 | | 2.203 | 0.040 | 0.857 | 33.446 |
| pH | -0.965 | 0.831 | -0.360 | -1.161 | 0.260 | -2.705 | 0.775 |
| orgcarbon | 0.151 | 0.185 | 0.916 | 0.814 | 0.426 | -0.237 | 0.539 |
| N | -1.266 | 1.650 | -0.863 | -0.768 | 0.452 | -4.720 | 2.187 |
| P | 0.062 | 0.073 | 0.203 | 0.854 | 0.404 | -0.090 | 0.214 |
| Ca | -0.433 | 0.356 | -1.085 | -1.217 | 0.238 | -1.179 | 0.312 |
| Mg | 1.438 | 1.426 | 0.413 | 1.008 | 0.326 | -1.548 | 4.423 |
| Mn | 0.384 | 1.327 | 0.088 | 0.289 | 0.776 | -2.394 | 3.161 |
| K | -0.313 | 0.212 | -0.899 | -1.478 | 0.156 | -0.756 | 0.130 |
| CEC | 0.537 | 0.399 | 1.279 | 1.345 | 0.195 | -0.299 | 1.372 |
| Sand | -0.040 | 0.066 | -0.309 | -0.603 | 0.554 | -0.179 | 0.099 |
| Silt | -0.130 | 0.085 | -0.751 | -1.535 | 0.141 | -0.307 | 0.047 |
| Clay | -0.064 | 0.083 | -0.174 | -0.776 | 0.447 | -0.238 | 0.109 |

**Table 2: Regression analysis result with Nitrogen as the dependent variable**

| Model | Unstandardized coefficients | | Standardized coefficients | t | Sig | 95% confidence interval for B | |
|---|---|---|---|---|---|---|---|
| | B | Std error | beta | | | Lower Bound | Upper bound |
| (constant) | 2.157 | 1.087 | | 1.985 | 0.062 | -0.118 | 4.433 |
| pH | -0.035 | 0.118 | -0.19 | -0.298 | 0.769 | -0.281 | 0.211 |
| orgcarbon | 0.106 | 0.009 | 0.947 | 12.484 | 0.000 | 0.089 | 0.124 |
| p | -0.008 | 0.010 | -0.038 | -0.800 | 0.433 | -0.29 | 0.013 |
| Ca | 0.002 | 0.051 | 0.006 | 0.031 | 0.975 | -0.104 | 0.108 |
| Mg | -0.177 | 0.196 | -0.075 | -0.900 | 0.379 | -0.588 | 0.234 |
| Mn | 0.239 | 0.174 | 0.080 | 1.379 | 0.184 | -0.124 | 0.603 |
| K | -0.084 | 0.024 | -0.354 | -3.540 | 0.002 | -0.134 | -0.034 |
| CEC | 0.011 | 0.057 | 0.038 | 0.189 | 0.852 | -0.109 | 0.130 |
| Sand | -0.018 | 0.008 | -0.202 | -2.172 | 0.043 | 0.035 | -0.001 |
| Silt | -0.025 | 0.011 | -0.215 | -2.347 | 0.030 | -0.048 | -0.003 |
| Clay | 0.001 | 0.012 | 0.005 | 0.115 | 0.910 | -0.023 | 0.025 |
| Yield | -0.024 | 0.031 | -0.035 | -0.768 | 0.452 | -0.089 | 0.041 |

## 4.0 CONCLUSION

The result from the regression analysis revealed that MLR can be effectively used for soil properties and to predict plantain yield. MLR analysis also showed that the soil chemical properties contributed the most important parameters in determining plantain yield, organic carbon,Phosphorus, Magnesiumand CEC represented the main variables responsible for plantain yield. This method can also be used to predict yield for other crops.

Furthermore, SEM providedan explanation of the simultaneous interactions among the variables included in the conceptual model. SEM revealed that the soil chemical factors were better indicators of plantain yield than the physical properties, yield component on plantain is dominated by organic carbon, Magnesium and CEC results of regression equations.

The integration of MLR and SEM analysis to evaluate the yield component of plantain showed that the fruit yield was related to soil properties and the interaction with YCP. Using Nitrogen as the dependent variable revealed a significant effect of organic carbon, potassium, as well as sand on silt on the Nitrogen content in the soil. The methodology used in this study shows the effect of incorporating soil properties information into a statistical model that involves yield and takes all possible interactions among the variables into consideration.

## REFERENCES


[1] Uyanık GK, Güler N. A Study on Multiple Linear Regression Analysis. *Procedia - SocBehavSci* 2013; 106: 234–240.

[2] Yan X, Gang Su X. *Linear Regression Analysis: Theory and Computing*. 5 Toh Tuck lin, Singapore 596224: World Scientific, 2009.

[3] Karimi L, Meyer D. Structural Equation Modeling in Psychology: The History, Development and Current Challenges. *Int J Psychol Stud*; 6, 2014.

[4] Igwenagu CM. Path Modeling of Global warming with CO2 Emission as a Surrogate. *IOSR J Math* 2014; 10: 83–88.

[5] Rengiah P. *Effectiveness of entrepreneurship education in developing entrepreneurial intentions among Malaysian university students*. DBA Thesis, Southern Cross University, 2013.

[6] Khine MS. *Knowing, Knowledge and Beliefs: Epistemological Studies across Diverse Cultures*. Springer Science & Business Media, 2007.

[7] McCullough EB, Pingali PL, Stamoulis KG. *The Transformation of Agri-food Systems: Globalization, Supply Chains and Smallholder Farmers*. Food & Agriculture Org., 2008.

[8] Jöreskog K, Sörbom D. *LISREL 9.10 for Windows [Computer software]. Skokie, IL: Scientific Software International, Inc.* Scientific Software International, Inc, 2012.

[9] CheRusuli M, Tasmin R, Takala J, et al. The Report of Structural Equation Modeling Analysis Results: A Study at Malaysian University Libraries. *Aust J Basic ApplSci* 2013; 7: 688–693.

[10] Nachtgall C, Kroehne U, Funke F, et al. (Why) Should We Use SEM? Pros and Cons of Structural Equation Modeling. *Methods Psychol Res Online* 2003; 8: 1–22.

[11] MOUTINHO LETA. *Quantitative Modelling in Marketing and Management (second Edition)*. World Scientific, 2015.