Journal of Experimental Research

December 2016, Vol 4 No 2
www.er-journal.com
Email: editor-in-cheif@er-journal.com

Received: September, 2016
Accepted for Publication: October, 2016

EXPLORATORY DATA ANALYSIS AND MULTIVARIATE STRATEGIES FOR REVEALING MULTIVARIATE STRUCTURES IN CLIMATE DATA

Igwenagu CM [1,2]

[1]Department of Industrial Mathematics/Applied Statistics, Enugu State University of Science and Technology, Nigeria
[2]Merc Data Consulting; http://www.mercdataconsulting .org

Author for Correspondence: *chineloigwenagu@yahoo.com*,

ABSTRACT

This paper is on data analysis strategy in a complex, multidimensional, and dynamic domain. The focus is on the use of data mining techniques to explore the importance of multivariate structures; using climate variables which influences climate change. Techniques involved in data mining exercise vary according to the data structures. The multivariate analysis strategy considered here involved choosing an appropriate tool to analyze a process. Factor analysis is introduced into data mining technique in order to reveal the influencing impacts of factors involved as well as solving for multicolinearity effect among the variables. The temporal nature and multidimensionality of the target variables is revealed in the model using multidimensional regression estimates. The strategy of integrating the method of several statistical techniques, using climate variables in Nigeria was employed. R^2 of 0.518 was obtained from the ordinary least square regression analysis carried out and the test was not significant at 5% level of significance. However, factor analysis regression strategy gave a good fit with R^2 of 0.811 and the test was significant at 5% level of significance. Based on this study, model building should go beyond the usual confirmatory data analysis (CDA), rather it should be complemented with exploratory data analysis (EDA) in order to achieve a desired result

Keywords: Climate variables, Factor analysis, Multivariate structure, Strategy, Technique,

INTRODUCTION

For some time, the scope of statistics has gone beyond inferences and estimation. It has been broadened to include exploratory data analysis and visualization to examine behaviours and trends in data beyond what ordinarily should be expected. (John, 1997). These techniques are typically applied before formal modeling commences and can help inform the development of more complex statistical models. (Roger et al. 2016) Exploratory techniques are also important for eliminating or sharpening potential hypotheses about the world that can be addressed by the data (Roger et al, 2016)

Exploratory data analysis (EDA) is a well established statistical tradition that provides conceptual and computational tools for discovering patterns to foster hypothesis development and refinement. Its goal is indictment in nature (Behrens and Smith, 1996). It is the process of making a sought of "rough cut" for a data analysis, by identifying relationships between variables that are particularly interesting or unexpected, checking to see if there is any evidence for or against a stated hypothesis, checking for problems with the collected data, such as missing data or measurement error, or identifying certain areas where more data need to be collected. EDA is necessary as it allows the investigator to make critical decisions about what needs to be followed up on and what probably is not worth pursuing because the data do not provide the evidence and may not provide the evidence, even with follow up. (Roger, 2015)These tools and attitudes complement the use of significant and hypothesis tests used in confirmatory data analysis CDA. EDA does not replace CDA instead it complements it.

As such, CDA is hardly done without EDA, which usually helps in interpreting the result of CDA and most often reveals misleading patterns in the data (John, 1997). Both EDA and CDA can be viewed as methods for comparing observed data to what would be obtained under an implicit (when patterns in a two-way plot are compared to an assumed model of e, exploratory independence) or explicit (when data are compared to sets of simulated data) statistical model (Bode et al, 1986). Most often, exploratory data analysis is considered when model building is of less interest. While, in Bayesian inference, exploratory data analysis is usually considered only in the early stages of model formulation and is of less importance once a model has actually been fit. However, (Andrew, 2004) argues that exploratory and graphical methods can be effective especially when used in conjunction with models, and that model-based inference can also be effective especially when checked graphically. If this argument holds it supports the idea that EDA and CDA complements each other.

The methodology employed in EDA has three major benefits which includes; explicit identification of a comparison model which allows one to simulate replicated data to be used as a reference distribution for an exploratory plot, symmetries in the underlying model can be used to construct exploratory graphs that are easier to interpret, sometimes (as with a residual plot) without the need for explicit comparison to a reference distribution and inclusion of imputed missing and latent data which can allow more understandable completed-data exploratory plots.

(Andrew, 2004). However, recent improvements in computation have spurred developments both in exploratory data analysis and in complex modeling which this paper intends to evaluate. As earlier mentioned that EDA is often presented as model-free; but the study by Tukey, (1972) which focused on "graphs intended to let us see what may be happening over and above what we have already discussed," which suggests that these graphs can be built upon existing models. He contrasted exploratory analysis with calculations of p -values, or *confirmatory data analysis*.

These two sets of methods are both forms of model checking. While exploratory data analysis is the search for unanticipated areas of model misfit, confirmatory data analysis quantifies the extent to which these discrepancies could be expected to occur by chance. This method tends to be based on fairly simple models such as additive fits and the Poisson distribution for counts. The study by (Andrew, 2004) applied the same principles to more complex models that can be fit using methods of Bayesian inference and nonparametric statistics. He observed that in complex models, test variables can be constructed using structure in the model or data; such that the average of the data and the residuals at the group level can be plotted against group-level predictors, and vectors of exchangeable variables at any level of the model can be displayed as histograms. He stated that more complicated cross-level structures, such as occurs in latent class models could also be plotted. He further viewed that the structure in the model could define default structures in the test variables, which generalizes the ideas of Tukey (1972, 1977) on two-way plots. Therefore, the objective of this study is to use data mining techniques in line with the EDA strategies, to explore the importance of multivariate structures; using climate variables which influences climate change.

MATERIAL AND METHODS

The Basic strategy in exploratory data analysis is to first examine the variables of interest one after the other, and establish the nature of the relationships among the different variables. (Roger,2015) The second strategy is to plot the graphs, then add numerical summaries of specific aspects of the data. In this paper, the exploratory multivariate strategy considered is the methods of ordinary least square (OLS) Regression analysis and Factor analysis. These two methods of analyses reduce the variable dimensions to the significant ones (Simon, 2015). The data used are climate variables in Nigeria recorded for a period of seventeen years. The OLS multiple regression analysis will be used to find the relationship among the different variables which is the first strategy in an exploratory data analysis. The model is presented in matrix form as

$$\underset{nx1}{Y} = \underset{nx2}{X}\underset{2x1}{\beta} + \underset{nx1}{\square} \quad ...... \quad .(1)$$

Factor analysis is a generic term for a family of statistical techniques concerned with the reduction of a set of observable variables in terms of a small number of latent factors. It has been developed primarily for analyzing relationships among a number of measurable entities. The underlying assumption of factor analysis is that there exist a number of unobservable latent variables or "factors" that account for the correlations among observed variables, such that if the latent variables are partialled out or held constant, the partial correlations among observed variables all become Zero. In other words, the latent factors determine the values of the observed variables. Each observed variable (y) can be expressed as a weighted composite of a set of latent variables (f's) such that,

$$yi = ai_1f_1 + ai_2f_2 + ...+ ai_kf_k \;+ ei \;.......\; (2)$$

Where y_i is the i_{th} observed variable on the factors and e_i is the residual of y_i on the factors.

Given the assumption that the residuals are uncorrelated across the observed variables, the correlations among the observed variables are accounted for by the factors. The observable random vector x, with P components has mean μ and covariance matrix Σ. The factor model postulates that x is linearly dependent upon a few unobservable random variables

$$x_1 - \mu_1 = L_{11}F_1 + L_{12}F_2 + ...L_{1m}F_m + E_1$$

$$x_2 - \mu_2 = L_{12}F_2 + L_{22}F_2 + ...L_{2m}F_m + E_2$$

$$\vdots \qquad \vdots \qquad \vdots \qquad\qquad (3)$$

$$x_p - \mu_p = L_{p1}F_1 + L_{p2}F_2 + ...L_{pm}F_m + E_p$$

This can be represented in matrix form as;

$$\underset{(px1)}{X - \mu} \;=\; \underset{(pxm)}{L}\;\underset{(mx1)}{F} \;+\; \underset{(px1)}{E} \qquad ...... \quad (4)$$

Where, μ = mean of variable 1, E_i = i_{th} specific factor, F_i = i_{th} specific factor, L_{ii} is called the loading of the i_{th} variable on the j_{th} factor, so the matrix L is the matrix of factor loadings. The i_{th} specific factor E_i is associated only with the ${}_ith$ response X_i. The P deviations $x_1 - \mu_2$, $\xi_2 - \mu_2$, $X_p - \mu_{p,}$ are expressed in terms of p + m random variables F_1, F_2 . . . F_m, E_1, E_2. . ., E_p which are unobservable.

That is what distinguishes factor model from multivariate regression model in (1) above, in which the independent variables whose position is occupied by F in factor model above (2) can be observed. (Igwenagu, 2012)

The exploratory data analysis strategy was illustrated using Climate Data collected from Metrological service department Headquarter located at Oshodi in Lagos State, Nigeria. The data were subjected to the first stage of EDA strategy as mentioned above and the variables of interest were extracted as presented in table 1 below. The data was also subjected to the second strategy and the graphs were as shown on figures 1,2 &3 below. Finally numerical summaries of specific aspects of the data were analysed using the chosen multivariate strategies. In contrast to traditional hypothesis testing; where the testing of hypotheses always requires an a priori assumption (or hypothesis) about the data, exploratory data analysis is not based on any a priori assumptions. Any method can be used to identify systematic relations between the variables (Lohninger , 2012) Therefore in this study, the best strategy was considered base on the one with a good fit.

RESULTS

Table 1.Climate Variables in Nigeria from 1998 to 2014

Yr/Var.	Tmax(oc)	Tmin (oc)	RF(mm)	Rh09(%)	Rh15(%)	Rad(ml)	Eva.(ml)	W/C (m/s)	CO2e(tons)
1998	387.8611	260.1	1403.04	775.8	619.89	285.17	65.82	58.05	0.5
1999	384.6278	252.4	1358.72	748.7	589.06	289.61	71.47	58.88	0.1
2000	391.7778	255.8	1347.34	761.1	578.67	295.28	69.38	59.73	0.7
2001	390.2889	254.7	1373.11	760.8	601.83	291.67	69.59	59.26	0.6
2002	394.0944	254.7	1438.76	765.2	588.50	295.18	72.66	59.27	0.5
2003	393.8278	252.9	1508.38	782.5	611.17	296.39	67.91	56.678	0.3
2004	391.7222	254.3	1441.78	782.6	612.17	288.58	66.29	56.75	0.4
2005	397.2722	260.3	1373.41	774.1	613.33	287.76	74.33	58.372	0.4
2006	391.3556	257.6	1607.77	797.4	627.33	279.88	69.98	54.34	0.4
2007	391.6889	255.3	1317.47	764.5	603.72	288.56	74.63	55.53	0.4
2008	392.0778	257.4	1390.17	764.5	598.56	294.08	70.34	56.69	0.4
2009	392.0111	260.9	1394.92	762.3	594.78	293.32	68.55	56.26	0.4
2010	394.5056	262.8	1412.79	776.4	602.22	288.79	67.12	60.92	0.4
2011	394.7167	262.4	1371.07	770.5	599.67	282.54	71.80	58.54	0.4
2012	395.8222	264.9	1362.52	773.3	595.22	288.28	67.73	58.52	0.8
2013	392.4056	258.1	1429.14	772.2	594.11	288.99	68.81	60.46	0.8
2014	393.5667	260. 2	1457.27	764.3	595.11	297.943	69.41	62.39	0.8

Source: Calculated from data from Metrological Service Department Headquarter at Oshodi Lagos State, Nigeria

The following variables: Wind current, Evaporation Pitch, minimum temperature maximum temperature, Solar Radiation, Rainfall, Relative Humidity at late hours of the day, Relative Humidity in the early hours of the day and Carbon dioxide (CO_2) emission as shown on the table above were extracted and calculated from the climate data collected from Metrological service department Headquarter at Oshodi Lagos State, Nigeria.

The second strategy of EDA carried out yielded the graph on fig.1 below.

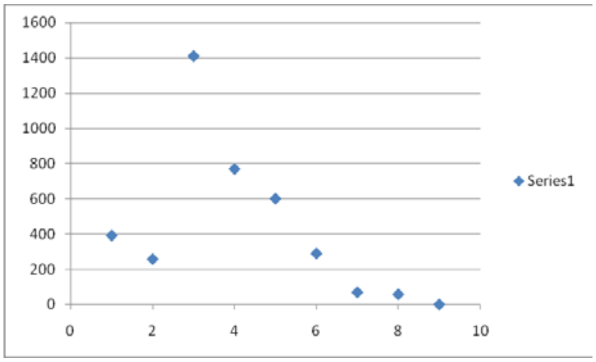

Figure 1.Scatter plot of the Climate variables for the period under study.

From fig.1 above there seems to be the presence of outliers in the data set. For this reason the average of the two variables with the outliers was taken and the scatter plot was plotted as shown on fig 2 below

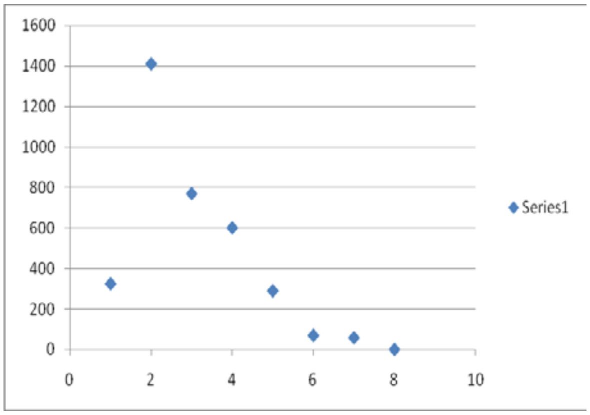

Figure 2: Scatter plot of the Climate variables with average temperature .

From fig 2 above, there is still the presence of one outlier. The variable average temperature(Tave) was eliminated entirely and the new scatter plot is as shown below:

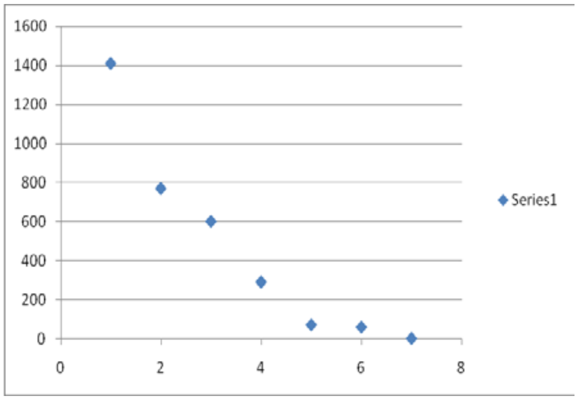

Figure 3: Scatter plot of the Climate variables without temperature variable.

Figure 3 above appears to be free from outliers and the variables used seems to be linear with downward slope.

Based on the EDA strategy, the nature of this data were established using the scatter plots. The final stage of the EDA strategy in order to establishe the complex model using regression model as earlier mentioned, gave the reults are as shown below:

Table 2 OLS Regression Analysis of Climate variables for the period under study.

					Change Statistics				
Model	**R**	**R Square Change**	**Adjusted R Square**	**Std. Error of the Estimate**	**F Change**	**df1**	**df2**	**Sig. F Change**	**Durbin- Watson**
1	.719(a)	.518	.035	.20364	1.073	8	8	.461	1.593

Table 3 ANOVA

Model		**Sum of Squares**	**df**	**Mean Square**	**F**	**Sig.**
1	Regression	.356	8	.045	1.073	.461(a)
	Residual	.332	8	.041		
	Total	.688	16			

a. Rel. Humidity, Early Rel. Humidity
b. Dependent Variable: CO_2emission

The F-value of 1.073 and the corresponding P-value of 0.461 in table 3 above indicates that the model established by this method is not significant and cannot be used to model climate change.

Table 4 Coefficients (a)

Model	Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
1	B	Std. Error	Beta	t	sig	Tolerance	VIF
(Constant)	ˉ17.602	13.083		ˉ1.345	.215		
Tmax	ˉ.036	.060	ˉ.506	ˉ.591	.571	.082	12.170
Tmin	.030	.030	.535	.987	.353	.205	4.873
Rainfall	ˉ.001	.002	ˉ.415	ˉ.708	.499	.176	5.677
Early Rel. Humidity	.027	.026	1.426	1.037	.330	.032	31.350
Late Rel. Humidity	-.010	.009	-.568	-1.067	.317	.213	4.697
Radiation	.025	.029	.586	.843	.424	.125	8.010
Evaporation	.023	.048	.287	.482	.642	.171	5.853
Wind current	.042	.035	.422	1.217	.258	.502	1.991

a. Dependent Variable: Co_2emission

Although the result of multiple regression analysis from table 2 above indicates that variables used accounted for 51.8% of the variation from the response variables, Durbin-Watson value of 1.593 shows some effects of auto-correlation. The complex model can be written as

$$\mathbf{Y}_{CO2} = -17.602 - 0.036T_{\max} + 0.030\ T_{\min} -0.001_{\text{Rainfall}} + 0.027\ _{\text{Early Rel.Humidity}} -0.010_{\text{Late Rel.Humidity}} + 0.025\ _{\text{Radiation}} +0.023_{\text{Evaporation}} + 0.042\ _{\text{Wind Current}}. \qquad \dots\dots\dots\dots\dots\dots \qquad (5)$$

From the model summary, the standard error is not that high with value of 0.20364. However, the collinearity statistics shows effect of multicolinearity in the variables used in the complex model with variance inflation Factor (VIF) > 5. Hence the variables affected where eliminated and a new analysis is as shown in table 5 below

Table 5 OLS Regression Analysis of Climate variables after Elimination

Model Summary

					Change Statistics				
Model	R	R Square Change	Adjusted R Square	Std. Error of the Estimate	F Change	df1	df2	Sig. F Change	Durbin-Watson
1	.598(a)	.358	.209	.18436	2.412	3	13	.114	1.687

a. Predictors: (Constant), Wind current, Tmin, Late Rel.Humidity
b. Dependent Variable: Co2emission

Table 6 ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.246	3	.082	2.412	.114(a)
	Residual	.442	13	.034		
	Total	.688	16			

a. Predictors: (Constant), Wind current, Tmin, Late Rel. Humidity
b. Dependent Variable: Co$_2$emission

Table 7 Coefficients(a)

Model	Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
1	B	Std. Error	Beta	t	sig	Tolerance	VIF
(Constant)	-5.572	4.394		-1.268	.227		
Tmin	.019	.013	.337	1.431	.176	.891	1.123
Late Rel. Humidity	-.002	.005	-.098	-.359	.725	.664	1.507
Wind current	.037	.028	.371	1.323	.209	.627	1.595

a. Dependent Variable: CO$_2$emission

From the model summary on table 5 above, the regression analysis with R^2 value of 0.358; still does not give a good fit.
Hence factor analysis was introduced to contend with this effect thus:

Factor Analysis

Table 8. Sampling Adequacy Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.421
Bartlett's Test of Sphericity	Approx. Chi-Square	6.660
	Df	3
	Sig.	.084

KMO and Bartlett's Test

Table 9 Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.578	52.587	52.587	1.578	52.587	52.587
2	1.042	34.745	87.332	1.042	34.745	87.332
3	.380	12.668	100.000			

Extraction Method: Principal Component Analysis.

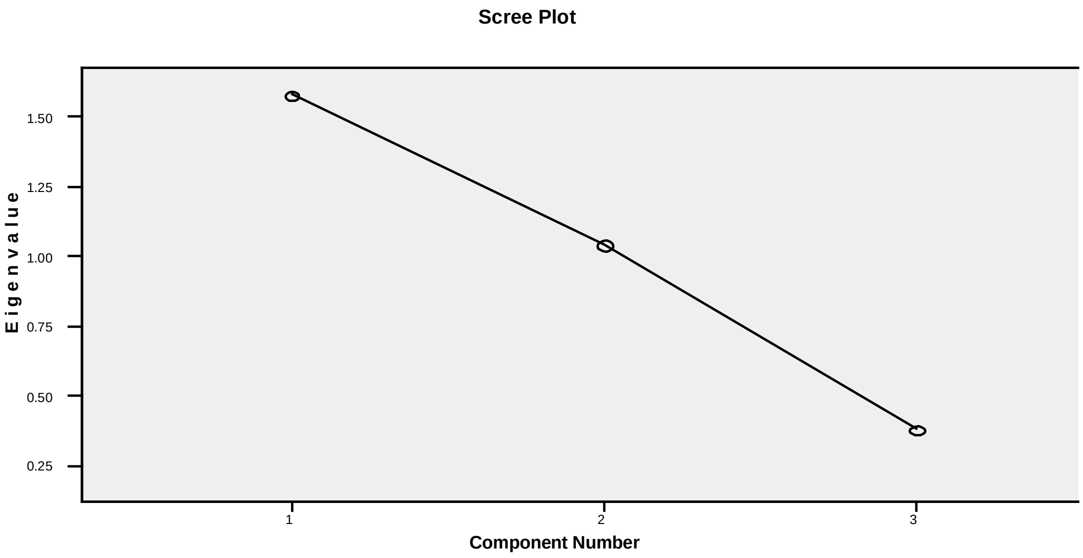

Figure 4 the Scree Plot

The scree plot selected two components as shown on the graph in fig4 above.

Table 10 Component Matrix (a)

	Component	
	1	2
Tmin	.295	.934
Late Rel. Humidity	.8-23	.408
Wind current	.902	.067

Extraction Method: Principal Component Analysis.
a. 2 components extracted.

Table 11 Factor scores Regression Analysis

					Change Statistics				
Model	R	R Square Change	Adjusted R Square	Std. Error of the Estimate	F Change	df1	df2	Sig. F Change	Durbin-Watson
1	.900(a)	.811	.725	.10881	9.419	5	11	.001	1.229

a. Predictors: selected factor scores
b. Dependent Variable: CO$_2$emission

The analysis result in table 11 above shows coefficient of multiple determination (R^2) value of 0.811, indicating that the variables used accounted for 81.1 %, variation in the value of the response variables used. The standard error of the estimate with value of 0.10881 is not high and could be considered okay. This model gave a good fit.

Table 12 ANOVA (b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.558	5	.112	9.419	.001(a)
	Residual	.130	11	.012		
	Total	.688	16			

a. Predictors: Selected factor scores (FS1 and FS2)
b. Dependent Variable: CO$_2$emission

The P- value of 0.001 in Table 12 above indicates that the test is significant.

Table 13 Coefficients (a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta	B	Std. Error
1	(Constant)	.491	.026		18.612	.000
	FS 1	.269	.057	1.297	4.737	.001
	FS 2	-.132	.085	-.637	-1.562	.147

a. Dependent Variable: CO$_2$emission

The complex model can be written as

$$Co_2 \text{ emission} = b_0 + 0.26FS_1 - 0.132FS_2$$

DISCUSSION

From figures 1and 2 above, the scatter plots showed linear relationship among the variables, although there seemed to be the presence of outliers. This was eliminated by eliminating some of the variable that were assumed to have multiple colinearity; a more desireable plot was obtained in figure 3.

From table 2 above, the model generated from multiple regression analysis carried out did not give a good fit with R^2 value of 0.518 and the test was not significant with *P-value* of 0.46.

Durbin-Watson value of 1.593 showed some effects of auto-correlation; also effect of multicolinearity was noticed among the variables used. These anomalies noticed in the data set used could not yield any meaningful result upon which a good model could be built. Situations like this could be dictated early enough with EDA strategies. As seen in the graph on fig1 above. The advantage of EDA over CDA is that the strategies involved makes it possible to examine the nature and behaviour of the data set and efforts are made to solve for possible defects as seen in this study. This follows the idea portrayed by John (1997).

Furthermore unlike EDA, most analysis done based on CDA alone fails to address the anomalies that results in the early stage of the analysis as agued by McGuire (1989). The presence of outlier already suggested that there could be effect of multicolinearity among the variables used this suggested the nature of the data from the onset. Using the EDA strategies as discussed, a successful elimination was used to select suitable data set; this is in line with the study by Henderson and Velleman (1981). The data set were analyzed using the factor scores regression, The result gave a good fit with R^2 value of 0.811 and the test was significant with *P-value* of 0.001. This shows that the use of data mining techniques is very useful in exploring the importance of multivariate structures as it concerns model building.

Factor analysis as used in this study as one of the multivariate analysis strategy; introduced into data mining technique, in order to reveal the influencing impacts of factors involved, as well as solving for multicolinearity effect among the variables, gave a good fit against the OLS regression. Although a set of climate data was used to illustrate this strategy, the strategy can also be applied in any multivariate data set. The major interest should be to follow the steps outlined in achieving this strategy.

CONCLUSION

Data mining technique is an important aspect in statistical modeling. From the EDA strategies employed in this study, factor analysis regression gave a good fit with R^2 value of 0.811 against the OLS regression and the test was significant at 5% level of significance. Unlike the CDA, without prior assumptions and hypothesis; the EDA strategy was used to select a data set that gave a good fit for modeling. Based on this study, the strategy of integrating the method of several statistical techniques, in data analysis should be encouraged. Model building should go beyond the usual confirmatory data analysis (CDA), rather it should be complemented with exploratory data analysis (EDA) in order to achieve a desired result.

REFERENCES

Andrew G. *(2004)*. Exploratory Data Analysis for Complex Models. *Journal of Computational and Graphical Statistics. 13(4): 755–779*

Behrens JT, Smith ML. (1996). Data and data analysis. In D. Berliner & B. Calfee (Eds.). The handbook of educational psychology: New York: Macmillan. (pp. 945-989).

Bode H, Mosteller F, Tukey JW, Winsor C. (1986). The education of scientific generalist. *In L. V. Jones (Ed.), the collected works of John W. Tukey, (4): 1949-1964.*

Henderson HV, Velleman PF. (1981). Building multiple regression models interactively. *Biometrics*, (37): 391--411.

Igwenagu CM. (2012). An Exploratory modeling of global warming. Ph.D Dissertation Published

Behrens JT. (1997). Principles and procedures of Exploratory Data Analysis. *Psychological methods*. 2(2): 13-160.

Lohninger H. (2012). Exploratory Data analysis. *Fundamentals of Statistics*. http://www.statistics4u.info/fundstat_eng/copyright.html

McGruire WJ. (1989). A perspectivist approach to the strategic planning of programmatic scientific result. *In B Gholson W. R Shadish R.A, Neimeyer and A.C Houts (Eds)Psychology of science: Contribution to meta science Cambridge University press*.

Peng RD. (2015). Exploratory Data Analysis with R. *http://leanpub.com/exdata. Google Assessed 31/10/2015*

Peng RD, Leek J, Caffo B. (2016). Exploratory Data Analysis *Coursera https://www.coursera.org/learn/exploratory-data-analysis*

French S (2015). Exploratory Data Analysis and Multivariate Strategies. *Google Assessed 31/10/2015.*

Tukey JW. (1972). Some Graphic and Semi-graphic Displays. In *Statistical Papers in Honor of George W. Google Assessed 31/10/ 2015*

Tukey JW. (1977). *Exploratory Data Analysis, New York: Addison-Wesley ISBN 978-0201076165. Google Assessed 31/10/ 2015*

Tukey JW, Wilk MB. (1986). Data analysis and statistics: An expository overview. *In L. V. Jones (Ed.), the collected works of John W. Tukey.* (4): 549- 578).