Enhanced Knowledge in Sciences and Technology Vol. 3 No. 1 (2023) 216-225
© Universiti Tun Hussein Onn Malaysia Publisher's Office

EKST

Homepage: http://publisher.uthm.edu.my/periodicals/index.php/ekst
e-ISSN : 2773-6385

A Multivariate Analysis on The Response of Crop To Fertilizer and Soil Type

Muhammad Awra Najwan Norhisham[1], Mohd Saifullah Rusiman[2*], Norziha Che Him[2]

[1] D'Tempat Matrix Country Club Sdn. Bhd.
Pt 12653, Jalan Pusat Dagangan Sendayan 1, Bandar Sri Sendayan,
71950 Seremban Negeri Sembilan, MALAYSIA

[2] Department of Mathematics and Statistics,
Faculty of Applied Sciences and Technology,
Universiti Tun Hussein Onn Malaysia (Pagoh Campus),
84600 Pagoh, Muar, Johor, MALAYSIA.

*Corresponding Author Designation

DOI: https://doi.org/10.30880/ekst.2023.03.01.025
Received 15 January 2023; Accepted 11 April 2023; Available online 3 August 2023

Abstract: This research work is primarily aimed at determining the significance effect of factors and to test the determinants of farmers participating behaviour in agricultural. A model of Multinomial Logistic Regression (MNL) is employed and other method that have been chosen which are Principal Component Analysis (PCA), and Factor Analysis (FA). Using multinomial regression, the dependent variable in question is a nominal where more there are more than two categories, analyse of series of data which observations are explained. Analysis of eigenvalues using PCA and FA represent the total amount of variance that can be explained by a given methods. Analysis of variance that replicates soil types is individually main effect and statistically significance at 5% significance level since P-value 0.002 which is below 0.05. The soil type achieved the highest variance of 50.1% compared to fertilizer type with 49.9% for being a factor and response for a crop. As concluded, Factor Analysis was taken to be the main factor as the percent of eigenvalues was higher 66% compare to Principal Component Analysis 50.76%.

Keywords: Multinomial Logistic Regression, Principal Component Analysis, Factor Analysis, Soil Type

1 Introduction

Crop production is one of the fundamental branches of agriculture. Crop production can be done on a commercial or subsistence foundation. Subsistence farming is when a farmer raises food to sell; commercial farming is when farmer raises food in huge quantities for market use. Before the discovery of oil, agriculture had been a significant factor in the economy of Nigeria [1]. Since being relegated to

*Corresponding author: saifulah@uthm.edu.my the margins, it has been performed in Nigeria at low level, with the bulk of active participants essentially being subsistence farmers. The farmers detected soil and fertilizer types have a big impact on crop yield which explains why different parcels of land planted with the same crop at the same time and with the same management package grow at different rates [2].

Crop production has been largely inconsistent resulting in the lack of knowledge of the combination of soil type. According to [3] in addition to being a medium for plant growth, soil also acts as a source of fertilizer, a place for plants to root, and a place for biological activity. A reduction in crop productivity is caused by a number of variables, including farmer ignorance of harvest glut, unpredictability of weather and seasonal rainfall patterns, and other [4]. The first objective in this study is to identify the significant variable toward soil types using Multinomial Logistic Regression. Based on the values of the independent variables, the multinomial logistic regression level forecasts the likelihood that the dependent variable will fall into different categories. According to the principle of highest probability of membership, the predictive category for the dependent variable is ultimately selected.

The past accomplishments in agriculture demonstrate the strength and capability of man in meeting the agricultural demand despite population rise. However, there has been an increase in the usage of fertilizer in nations that offer input subsidies, such as Malawi, Mali, Nigeria, and Tanzania [5], and this trend is likely to continue in the years to come. The second objective is to determine the main factors by using Principal Component Analysis and Factor Analysis. PCA is a multivariate statistical data analysis technique that divides a set of raw data into a number of primary components that preserve the maximum variance in the original data.

Although application rates have for many years been based on general recommendations, fertilizer is frequently not suited to individual crop, soil, or agro-ecological circumstances [6]. An iterative principal component analysis (PCA) data reduction process similar to [7] as a data-driven approach to determine important covariate layers. The comparison of the performance of Principal Component Analysis and Factor Analysis used by using eigenvalues will be the last objective. Most factor analyses reported in the literature use only strength of a series of factors in explaining the variance of water quality data to identify factors that explain the dominant variables in the datasets [8]. The scope of this study focused on how crops respond to different types of fertilizer in each of the different types of soils by using chosen methods, Principal Component Analysis and Factor Analysis.

2 Materials and methods

This paper aim to look at the response of crop towards fertilizer and soil type. Figure 1 shows the research framework for the implementation of the study. This paper proposes a comparative analysis of statistical and multivariate methods for prediction of response crops toward fertilizer and soil type. Figure 1 shows the research framework for the implementation of the analysis that involved in predict the models between Principle Component Analysis and Factor Analysis. Multinomial logistic regression showing soil characteristics responsible for the allocation of fields to specific yield-nutrients response clusters. The framework illustrates the steps done in the study towards achieving all the objectives. In analysing the data collected for the purpose of carrying out this research, the statistical tool known as the multivariate analysis was used. The use of sample percentage was also employed. Tables were used in presenting data for the purpose of the simplicity and clarity.

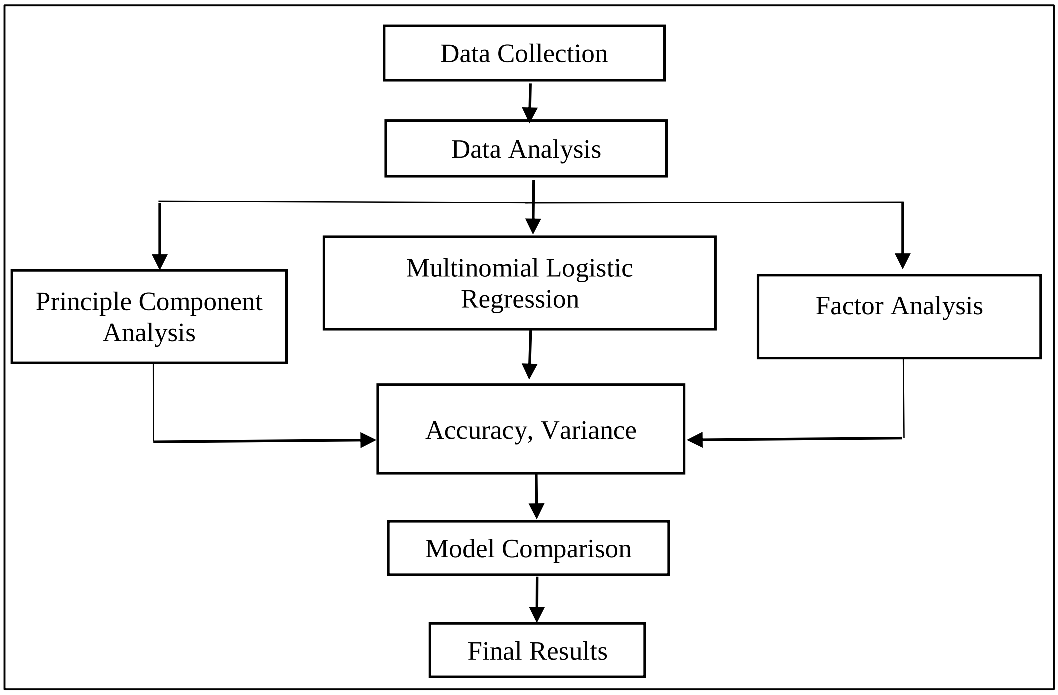

Figure 1: Research framework

2.1 2.1 Data description

The dataset was obtained from kaggle website (www.kaggle.com/code/fertilizer-type-prediction). The causality approach to this study is most preferred because the study will be investigating the response of crops to fertilizers and soil types.

Table 1 : List of Variables

Variables	Description	Type of variable
Y	Soil type	Qualitative
X1	Fertilizer type	Qualitative
X2	Crop type	Qualitative
X3	Temperature	Quantitative
X4	Humidity	Quantitative
X5	Moisture	Quantitative
X6	Potassium	Quantitative
X7	Nitrogen	Quantitative
X8	Phosphorous	Quantitative

2.2 Methodology

2.2.1 Multinomial Logistic Regression

Multinomial Logistic Regression is one of the most common classification algorithms used for analysing binary and categorical target. The connection between the categorical dependent variable and continuous independent variable is measured by changing the dependent variables into probability scores. According to [9], the Multinomial Logistic Regression method has a number of important benefits over other regression models and had an interesting interpretation in terms of logistic regressions. The logistic regression can be extending to models with multiple explanatory variables. Let *k* denotes number of predictors for a binary response *Y* by $X_1, X_2 \ldots X_k$, the model for log odds is

$$\text{Logit } [P(Y = 1)] = \alpha + \beta_1 X_1, + \beta_2 X_2 + \ldots \beta_k X_k \qquad Eq.\ 1$$

And the alternative formula, directly specifying *π(x)*, is

$$\pi(\mathrm{x}) = \frac{exp(\alpha+\beta 1jX1i+\beta 2jX2i+\ldots+\beta\rho jX\rho i)}{1+exp(\alpha 0i+\beta 1jX1i+\beta 2jX2i+\ldots+\beta\rho jX\rho i} \qquad Eq.\ 2$$

The parameter β, refers to the effect of xi, on the log odds that Y =1, controlling other xj. For instance, exp(βi) is the multiplicative effect on the odds of a one-unit increase in xi, at fixed levels of other xj. The effects of the predictors vary according to the response paired with the baseline, i.e., the regression coefficient βj, are specific to the corresponding logistic model. The multinomial probabilities may be obtained as :

$$log(Xj(Xi)) = \frac{exp(\alpha 0i+\beta 1jX1i+\beta 2jX2i+\ldots+\beta\rho jX\rho i)}{1+\sum_{j=1}^{h-1} exp(\alpha 0i+\beta 1jX1i+\beta 2jX2i+\ldots+\beta\rho jX\rho i)} \qquad Eq.\ 3$$

There are also other approaches for building regression models for multinomial responses. One such approach is to consider a multivariate generalized linear model (GLM), assuming the multinomial distribution for the response. Specifically, let yi =(yi1. ... yim), where yij = 1 if the response of individual *i* is in category *j* and yij = 0 otherwise (so $\sum yij = 1$), *i* - 1,2,...*n: j* = 1,2.. ..,*m.*

2.2.2 Principal Component Analysis

Principal component analysis illustrates the most significant parameters, which describe the whole dataset providing data reduction with minimum loss of original information. For a theoretical development of the principal component analysis, it is necessary to use some results on the canonical reduction of matrices, which are summarized in this section for use in the later sections [10]. The model of principal component analysis (PCA) is expressed as Eigenvalues and vectors of a matrix. Let $\sum$ be a non-negative (*i.e*., positive definition or positive semi-definition) matrix of order *p*X*p*. Corresponding to each root λI, there exists a column rector *P*, such that

$$\sum Pi = \lambda P\mathrm{I} \qquad Eq.\ 4$$

which is called an eigenvector. An eigenvector is a nonzero vector that changes at most by a scalar factor when that linear transformation is applied to it. The corresponding eigenvalue is the factor by which the eigenvector is scaled [11]. Yi =e_j^iX are the principal components obtained from the covariance matrix $\sum$,then

$$\rho Y_i,X_k=\frac{e\sqrt{\lambda}}{\sqrt{\sigma}}, \qquad l,k = 1,2,\dots, p \qquad Eq.\ 5$$

are the correlation coefficients between Yi and variable Xk, Here $e_j^i = [ei1\ ei2 \dots eip]$ is the eigenvector of $\sum$, corresponding to the eigenvalue λI. Also, X= [X_1 X_2 ... X_p].

2.2.3 Factor Analysis

Meanwhile, factor analysis (FA) attempts to extract a lower dimensional linear structure from the data set. The inclusion of chemical parameters only for factor analysis is the standard approach for most studies of stream chemistry that use factor analysis [12]. It further reduces the contribution of less significant variables obtained from PCA and the new group of variables known as varifactors (VFs) which is extracted through rotating the axis defined by PCA. The two approaches, PCA and FA, are primarily described in identical equations, with the exception of PC, which is expressed as a linear combination of measured variables. A VF, on the other hand, might incorporate unobservable, hypothetical, latent variables because the measured variable is expressed as a mixture of factors and the equation comprises the residual term [13]. The purpose of factor analysis is to attain parsimony by utilizing the fewest possible explanatory ideas to explain as much shared variance as possible in a correlation matrix. An overview of factor analysis approaches was provided in this article, as well as a conceptual explanation of factor in assessing the suitability of factor analysis. The model of factor analysis equation can be represented in matrix form as:

$$X=\mu+LF+e \qquad Eq.\ 6$$

where **X** is the p x 1 vector of measurements, $\boldsymbol{\mu}$ is the p x 1 vector of means, $\boldsymbol{L}$ is a $p \times m$ matrix of loadings, $\boldsymbol{F}$ is a $m \times 1$ vector of common factors, and $\boldsymbol{e}$ is a $p \times 1$ vector of residuals. Here, p represents the number of measurements on a subject or item and m represents the number of common factors. For the greater part of the twentieth century, it was widely utilized as a data analytic technique [14]. Data reduction, instrument building, classification and description of data, data transformation, hypothesis testing, exploring relationships in new domains of interest, and mapping construct space have all been utilised extensively by social scientists [15]. The latent vectors η and ξ are related with the manifest random vectors by the measurement model which composes the following Factor Analysis models by let y=$\left(X_1^T, X_2^T\right)^T\omega = (\eta^T,\xi^T)^T$ *Eq.5* can be expressed as

$$vvvY=\begin{pmatrix} X1 \\ X2 \end{pmatrix}=\begin{pmatrix} \Lambda1 & 0 \\ 0 & \Lambda2 \end{pmatrix}\begin{pmatrix} \eta \\ \xi \end{pmatrix}+\begin{pmatrix} e1 \\ e2 \end{pmatrix}=\Lambda\omega+e \qquad Eq.\ 7$$

These underlying factors can be extracted using principal component analysis (PCA) or factor analysis (FA), which is a mathematical process that converts a large number of (possibly) correlated variables into a (smaller) number of uncorrelated (PCA) or correlated (FA) variables known as principal components or factors. The shared variance of a variable is separated from its unique variance and error variance during the PCA/FA extraction to show the underlying factor/PC structure. In the solution, there is only the shared variance. Finally, people tend to use PCA to reduce the data into a smaller number of components, while they use FA to understand what constructs underlie the data.

3 Results and Discussions

This section shows the result of the modelling and discusses the comparison of the three algorithms; Multinomial Logistic Regression, Principle Component Analysis and Factor Analysis.

3.1 3.1 Multinomial Logistic Regression

Based on Table 2, it is shown the selected variable that significant towards soil type.

Table 2: Standard Normal Distribution for Soil Type

Soil Type	B	Std. Error	Wald	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
Crop Type	-.248	.127	3.815	.051	.780	.608	1.001
Moisture	.102	.044	5.398	.020	1.107	1.016	1.207
Moisture	.070	.044	2.521	.112	1.072	.984	1.168
Humidity	.598	.343	3.032	.082	1.818	.928	3.563

Generally, 95% confidence interval or 5% level of the significance level is chosen for the study. Thus, the *p*-value should be less than 0.05. In the above table, there are four significant variables chosen which are significant for a soil type. The model shows as

$$Y_1=0.51-0.248+0.127+3.815 \qquad \textit{Eq.}\ 8$$

$$Y_2=0.02+0.102+0.044+5.398 \qquad \textit{Eq.}\ 9$$

$$Y_3=0.112+0.07+0.044+2.521 \qquad \textit{Eq.}\ 10$$

$$Y_4=0.082+0.598+0.343+3.032 \qquad \textit{Eq.}\ 11$$

Based on Table 3, it is shown that the overall percentage for the observation on each component of soil type.

Table 3: Classification table output for the multinomial logistic regression

	Predicted					
Observed	1	2	3	4	5	Percent Correct
1	6	1	0	10	3	30.0%
2	3	11	6	1	0	52.4%
3	0	7	0	4	8	0.0%
4	4	5	1	4	5	21.1%
5	5	0	2	5	8	40.0%
Overall Percentage	18.2%	24.2%	9.1%	24.2%	24.2%	29.3%

It shows the crop type, moisture and humidity give more influenced towards crop type. Classification table reporting the overall percentage which 29.3% of correctly classified cases, and showing not very good classificatory power of the Logistic Regression model.

3.2 Principal Component Analysis

Based on Table 4, it is shown that the total variance in every component that influenced the crop production by extract the Principal Component Analysis method while Figure 2 shows the scree plot of eigenvalues on each nine components.

Table 4: Total Variance Explained in nine components PCA

	Initial Eigenvalues		
Component	Total	% of Variance	Cumulative %
1	273.595	50.761	50.761
2	131.035	24.311	75.072
3	67.872	12.592	87.664
4	40.884	7.585	95.250
5	15.865	2.943	98.193
6	6.495	1.205	99.398
7	1.792	.333	99.731
8	.992	.184	99.915
9	.459	.085	100.000

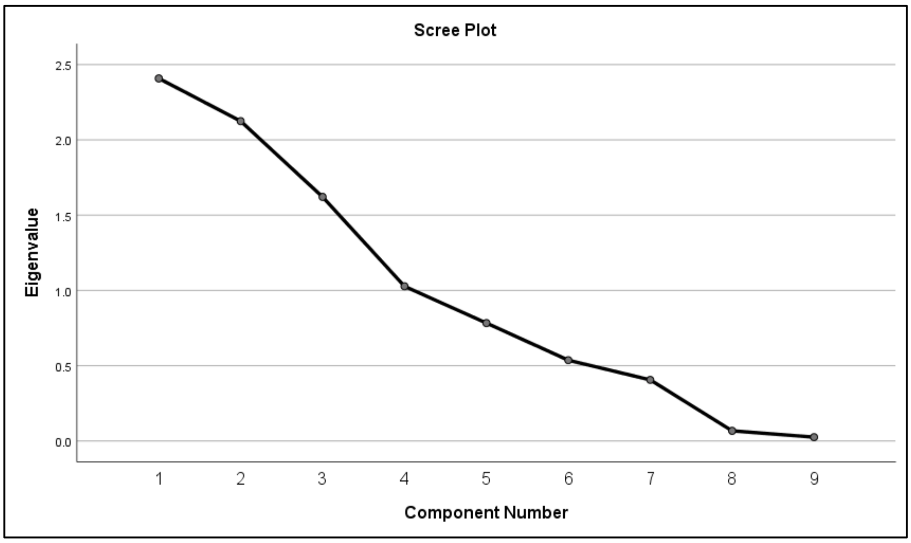

Figure 2: Scree Plot of Eigenvalue

From the Fig.2, it shows that the first four principal components have eigenvalues greater than 1. The highest fraction of explained variance among these variables is 50.76%, and the lowest one is 0.85%. The more spread the data, the larger the variance is in relation to be the factor component[16]. The scree plot shows that the eigenvalues start to form a straight line after the four principal component. If 26.75 % is an adequate amount of variation explained in the data, then the first four principal components were used. The objective is to provide a relatively straightforward technique for automatically and fairly objectively locating the gap in the scree plot. Other than the one described by [17], which is based on resampling approaches, we are not currently aware of any other such automatic techniques.

3.3 Factor Analysis

Based on Table 5, it is shown that the factor matrix generates by Maximum Likelihood method to find the significant value to build a model for eigen analysis.

Table 5: Factor matrix

	1	2	3	4
Soil Type	-.085	.007	-.045	.346
Fertilizer Type	.317	-.362	.696	.106
Crop Type	-.126	-.022	-.159	.842
Temparature	.311	.945	.093	.001
Humidity	.318	.913	.123	-.008
Moisture	.071	.060	.139	-.620
Nitrogen	-.883	.291	-.366	-.001
Potassium	.299	-.173	.506	-.004
Phosphorous	.922	-.046	-.384	.000

Model generate:

Y_1= $[(-0.085) + (0.317) + (-0.126) + (0.311) + (0.318) + (0.071) + (-0.883) + (0.299) + (0.922)]$=0.66

Y_2= $[(0.007) + (-0.362) + (-0.022) + (0.945) + (0.913) + (0.060) + (0.291) + (-0.173) + (-0.046)]$=0.24

Y_3= $[(-0.045) + (0.696) + (-0.159) + (0.093) + (0.123) + (0.139) + (-0.366) + (0.506) + (-0.384)]$=0.51

Y_4= $[(0.346) + (0.106) + (0.842) + (0.001) + (-0.008) + (-0.620) + (-0.001) + (0.004)]$=0.40

Based on the eigenvalue calculation on Maximum Likelihood extraction method above, the calculation show that all the model are lower than zero. So the independent variable for Y_1 have a greater value compare to another. The higher is the eigenvalue, the higher will be the variance along the covariance matrix's eigenvector direction. From that, Maximum Likelihood as a method for Factor Analysis show the best method with the higher percent of eigenvalue 66% compare to the Principal Component Analysis 50.76%.

3.4 Comparisons

Principal Component Analysis and Factor Analysis are both techniques to reduce the data from the higher dimensions to lower dimensional space without losing the information content of the data variance. Both of these have similarities yet are certainly not synonyms of each other. The comparison between both methods of multivariate between Principal Components Analysis and Factor Analysis made by looking at the eigenvalues of the model.

Λ_i=50.761 (Principal Component Analysis)

$$Y_i = [(-0.085) + (0.317) + (-0.126) + (0.311) + (0.318) + (0.071) + (-0.883) + (0.299) + (0.922) = 0.66 \text{ (Factor Analysis)}$$

So as a result, the higher the eigenvalue, the higher will be the variance that can influenced towards crop type. Correlation matrix reports that the eigenvalues values are greater than 1, which is a means to address the choice of the number of factors. The method of PCA is used for comparison purposes involving the statistic package SPSS. The results clearly report the usefulness of multivariate statistical analysis (Principal Component Analysis).

3.5 Conclusions

Based on the study of correlations between large numbers of quantitative variables, the factor analysis (FA) method aims at finding structural anomalies of a communality composed of p-variables and a large number of data (large sample size). It reduces the number of original (observed) variables by calculating a smaller number of new variables, which are called factors. In PCA the original variables are transformed into the smaller set of linear combination, with all of the variance in the variables being used. In FA, however, factors are estimated using mathematical model, where only the shared variance is analysed. However, principal components analysis is often preferred as a method for data reduction, while principal factors analysis is often preferred when the goal of the analysis is to detect structure. Three techniques were applied for the determination model which are Multinomial Logistic Regression (MNL), Principal Component Analysis (PCA), and Factor Analysis (FA) to determine which response is more effective for the farmers. The experiments showed that the crop type, moisture and humidity give more influenced towards crop type for a significant variable toward soil types using Multinomial Logistic Regression. The soil type achieved the highest variance of 50.1% compare to fertilizer type with 49.9% for being a factor and response for a crop. The greater the variance, the greater the spread in the data. For purposes of comparison, the statistical software SPSS is utilised in conjunction with the PCA approach. The outcomes unequivocally demonstrate the value of multivariate statistical analysis (Principal Component Analysis).

Acknowledgement

The research work was supported by Ministry of Higher Education (MOHE) through Fundamental Research Grant Scheme (FRGS/1/2020/STG06/UTHM/02/4).

References

[1] A.Suryawanshi., M.Savasani., & J.Shah., (2015). “A Multinomial Logistic Regression Study of Married Men and Women's Stress Levels.” Journal of Scientific and Technical Publications, 5, 1–8(2015).

[2] H.Abdi., and L.J.WilliamS.,“Principle Component Analysis.” Reviews Computational Statistics. 2: 433-459 (2010).

[3] P. Sanchez, G. Denning, G. Nziguheba “The African green revolution moves”, Journal of agriculture production.,pp. 37-44 (2009).

[4] B.G.Tabachnick, L.S.Fidell, & S.J.Osterlind. “Using multivariate statistics.”Journal of Allyn and Bacon, (2001).

[5] J.M.Hilbe, “Logistic regression models.” CRC press, (2009).

[6] R.L.Gorsuch,” Factor analysis” W.B. Saunders Company, Phila-delphia (1974).

[7] A.Alsharif, B.Pradhan “analysis of Tripoli Metropolitan city (Libya) multivariate logistic regression model” J Indian Soc Remote Sens 42(1):149–163F,(2014).

[8] T.Hengl., N.Toomanian., H.I.Reuter., M.J.Malakouti., 2007b. “Methods to interpolate soil categorical variables from profile observations:” lessons from Iran. Geoderma 140 (4), 417–427, (2011).

[9] S.Wold, K.Esbensen, and P.Geladi., “Principal component analysis” Chemometrics and Intelligent Laboratory Systems 2: 37 -52,(1987).

[10] S.A.Mohammadi., and B.M.Prasanna B.M. “Analysis of genetic diversity in crop plants salient statistical tools and considerations” Crop Sci., 43: 1235–1248,(2003).

[11] K.R.Akshatha., and K.S.Shreedhara, “Implementation of machine learning algorithms for crop recommendation using precision agriculture”. Journal of Machine Learning (2018).

[12] B.Parinet, A,Lhote, B.Legube “Principal component analysis: an appropriate tool for water quality evaluation” Ecol Model 178:295–311,(2004).

[13] C.D.Evans., T.D. Davies, P.J. Wigington, Jr., M. Tranter, and W.A.Kretser. “Use of factor analysis to investigate processes con-trolling the chemical composition of four streams in the Adirondack Mountains,” New York. J. Hydrol. (Amsterdam) 185:297–316, (1996).

[14] G.D.Garson.,“Factoranalysis” Retrieve from chass.ncsu.edu(2010).

[15] T.J.Hastie., R.J.Tibshirani., J.H Friedman.,“The Elements of Statistical Learning” Data-Mining, Inference and Prediction,(2001).

[16] Z.Springer, L.Berlinspatial “data analysis with R, Use R” Springer, New York; London xiv (374 p),(2012).

[17] W.F.Velicer, & D.N.Jackson, “Component Analysis Versus Common Factor AnalysisSome Further Observation” Multivariate Behavioral Research, 25(1), 97-114,(1996).