APPLICATION OF PRINCIPAL COMPONENT ANALYSIS TO THE DETERMINATION OF IMPACT OF DOWN TIME ON PRODUCTION OF WHEAT PRODUCTS

ISSN: 2141 – 3290
www.wojast.com

UMOREN, M. U[1]. AND OFEM, E. I[2].
Department of Mathematics and Statistics
University of Uyo, Uyo, Nigeria
E-mail: umorenmfoniso@yahoo.com
Department of Mathematics/Statistics and Computer Science , University of Calabar
Calabar, Nigeria

ABSTRACT: For any company to stay in business, it must be making satisfactory profit and cash flow, and for it to make profit, a quality system for both the company and customers must be provided so that there could be a remarkable market for its products. The problem considered in this work was to determine the impact of Down Time on an optimal production capacity of wheat products so as to meet the daily demands of customers in order to minimize outstanding demand by reducing Down Time to its barest minimum. Niger Mills Company Plc, Calabar was used as the case study, and the data for the years 2004-2006 in which the company was unable to meet customers' demands were employed in the analysis. Principal Component Analysis (PCA), a Multivariate Statistical Analysis Method was used as the method of investigation. Results show that there was no significant impact of Down Time on the production of the company's product as may be claimed by the management of the company.

INTRODUCTION

Niger Mill Company (PLC) with a production capacity of 250 tonnes per mill was established by the South Eastern State (now, Cross River State) Government, Nigeria in the early 1970's. The sitting of the company, in its present location, usually referred to as "Green field site" may have been due to its numerous advantages, one of which is its nearness to the 'local dock side grain terminal' capable of handling large bulk carrier vessels. The company's major raw material, wheat, is imported from United States of America (Raychaudhuri et al 2000).

Down Time is the period of time the company is not producing either because of poor machines' maintenance or technical milking problem, with a minimum of 48 hours in a month. This means that the company will be engaged in its production process for at most 28 days of a month. The question is, does Down Time of about two days a month significantly impact the production of wheat product? Also, given the DT, can the company's wheat production capacity meet outstanding customers' demands? Can we find the minimum DT that will overcome the outstanding supply of wheat product? The answers to these questions formed our major consideration in this work.

We note here that, by wheat product we mean Flour (FL), Semovita (SE), and Wheat Offals (OF). Hence, we shall in general be speaking about the impact of Down Time on the production of Flour, Semovita and wheat Offals. Also, by outstanding supply we mean products which have been paid for and are yet to be produced and delivered to customers.

The data employed in this study were collected from Niger Mills Company PLC, Calabar for the period 2004 – 2006 and are displayed in terms of Down Time and Wheat Products (Flour), Semovita, and Wheat Offals and were observed for twelve months. Thus,

$x_{ijk}, i = 1,2,3,4, \quad j = 1,2,...,12, \quad k = 1,2,3,$ is the ith observation for the jth month and in the kth year. For example, x_{111} is the observed value of Down Time in January, 2004, x_{212} is the observed quantity of flour for January, 2005, x_{323} is the observed quantity of Semovita in February, 2006, and x_{433} is the observed value of Wheat Offals for March 2006, etc.

In order to make for easy management of the matrices that will arise in the analysis, we shall merge the fourth variable, wheat offal by subtracting it from each of the standardized observations of the other variables. Thus, this study will from now on consider standardized scores for Down-Time, Flour, and Semovita, i.e. the number of variables will now be three (i=1,2,3).

Suppose we let X represent the vector of variables; i.e. $X = (x_1, x_2, x_3)$ so that $E(X) = (\mu_1, \mu_2, \mu_3)'$, and

$$V = \text{cov}(X, X) = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22}^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33}^2 \end{pmatrix}$$

with the corresponding correlation matrix

$$C = \begin{pmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{pmatrix}$$

Our interest is to compute the eigen values and eigenvectors of the correlation matrix. The scalars, $\lambda_1, \lambda_2, \lambda_3$ satisfying the polynomial equation $|C - \lambda I| = 0$ are called eigenvalues, and $\underline{x}, \underline{x} \neq 0$, such that $Cx = \lambda x$ are the eigenvectors (or characteristic vectors) of the matrix associated with the eigenvalue λ.

FUNDAMENTALS OF PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis (PCA) largely developed by Hotelling (1933) is a technique for reducing a large set of correlated variables to smaller number of uncorrelated hypothetical components, the purpose being to interpret the total observed variance with as few variables as possible, which is usually, accomplished by a linear transformation of the original variables (Yeung, and Ruzzo, 2001; Richard and Dean, 1992).

The first principal component is the normalized linear combination of the original variates and has maximum variance, with its coefficients squared to one. The second principal component is another normalized linear combination of the original variates orthogonal to the first PC with coefficients also squared to one (Jyh-Cheng and Samuel, 1999). The other components are similarly defined, each component being orthogonal to the others and providing maximum variance of the remaining residual data.

In order to obtain the first PC, a linear combination $c_1'x$ is formed such that $\text{var}(Y_1) = \text{var}(c'x) = E(c'xx'c) = c'vc$ is maximum, subject to the constraint $c'c = 1$. That is, using Lagrange's multipliers,

Maximize $$\Phi(c, \lambda) = c'vc - \lambda(c'c - 1) \qquad (1)$$

where λ are the Lagrange's multipliers

$$\frac{\partial \Phi(c'\lambda)}{\partial c_1} = 2vc - 2\lambda c = 0$$

i.e $\quad vc - \lambda c = 0$

or $\quad (v - \lambda I)c = 0$

or $\quad |v - \lambda I| = 0$

since $c \neq 0$, which is a polynomial of degree p. Therefore c_1 is the first eigenvector of $v - \lambda_1$, with λ_1 being the first eigen-value. Now, since c_1 is normalized, then

$$c_1' v c_1 = \lambda_1 c_1' c_1 = \lambda_1 = \mathrm{var}(c_1' x).$$

The second PC is also obtained by forming a linear combination $c_2' x$ such that $\mathrm{var}(c_2' x) = c_2' v c_2$ is maximum, subject to $c_2' c_2 = 1$. $\mathrm{cov}(c_2' x, x' c_2) = 0$.
That is, using the Lagrange's multipliers,

$$E(c_2' v c_2) = \left(c_2' \lambda_2 \theta - 1, \ \ c_2' c_2 - 1 \right) - 2\theta (c_2' v c_2 - 1)$$

is maximum, so that

$$\frac{\partial \phi(c_2, \lambda_2, \theta)}{\partial c_2} = 2vc_2 - 2\lambda_2 c_2 - 2\theta v c_2 = 0$$

or $\quad vc_2 - \lambda_2 c_2 = \theta v c_2$

or $\quad (v - \lambda I)c_2 = 0$

or $\quad c_2' v c_2 = \lambda_2 c_2' c_2 = \lambda_2 v c_2 c_2' = 0$

$$\Rightarrow \quad c_2' v c_2 = \lambda_2 c_2' c_2 = \lambda_2$$

Hence $(v - \lambda_2 I)c_2 = 0$, since $c_2 \neq 0$. Therefore, c_2 is the second eigen vector corresponding to the second eigen-value λ_2. The remaining eigenvalues are similarly obtained.

ILLUSTRATION

The following steps are involved in the computations of the eigenvalues/eigenvectors:

S_1: Raise the correlation matrix R to power 16; i.e $R^{16} = R^*$, and then multiply by an initial column vector (1,1,1)′.

S_2: The entries in each column of the matrix R* are added and their sums written down at the bottom of each column. Each of the sums is divided by the largest and the result written down as column vector besides R*

i.e

$$R_1^* = \begin{pmatrix} * & * & * \\ * & * & * \\ * & * & * \end{pmatrix} \overset{a^*}{\begin{pmatrix} * \\ * \\ * \end{pmatrix}}$$

S_3: Multiply each row of R* by the corresponding entries in column a*; this gives rise to a new matrix R_2^*

S_4: Sum the entries in each column of the R_1^* matrix and write down the sum at the bottom of each column, then divide by the largest to obtain the vector a^{*2}.

S_5: Continue the process by repeating S_3 using the latest matrix until the process converges. That is, until a^{*n} is not different from a^{*n-1}. The normalized value of a^{*n} equals c_1.

Then $\lambda_1 = \left(c_1 R^* c_1\right)^{-1/2}$

Replacing R^* with $R_1^* = R^* - \lambda_1 c_1 c_1'$ we have $\lambda_2 = \left(c_1 R_1^* c_1\right)^{-1/2}$

$$R_2^* = R_1^* - \lambda_1 c_1 c_1'$$

or $$R_2^* = R^* - \lambda_1 c_1 c_1' - \lambda_2 c_2 c_2'$$

and $$a'_1 = \left(c_2' c_2\right)^{-1} c_2'$$

DATA ANALYSIS

Given the twelve observations for the three standardized variables, the variance covariance matrix for the pooled observations for the years 2004-2006 is

$$C_p = \begin{matrix} z_{1p} \\ z_{2p} \\ z_{3p} \end{matrix} \begin{pmatrix} 1.4067 & & \\ 1.0040 & 1.5777 & \\ 0.8216 & 1.1845 & 2.2129 \end{pmatrix}$$

with the corresponding correlation matrix

$$R_p = \begin{matrix} z_{1p} \\ z_{2p} \\ z_{3p} \end{matrix} \begin{pmatrix} 1 & & \\ 0.6870 & 1 & \\ 0.4660 & 0.6460 & 1 \end{pmatrix}$$

The eigenvalues of R_p are given as:

Variable	Eigenvalues	Difference	Proportion	Cumulative
z_{1p}	2.2042	1.6682	0.7350	0.7350
z_{2p}	0.5360	0.2763	0.1790	0.9130
z_{3p}	0.2597		0.0870	1.0000

with corresponding eigenvectors (principal components)

Variable	PC_1	PC_2	PC_3
z_{1p}	-0.5640	0.6630	0.4920
z_{2p}	-0.6170	0.0580	-0.7850
z_{3p}	-0.5490	-0.7470	0.3760

The coefficients for the principal components of the standardized variables is summarized in Table 1

Table 1: Coefficient of the principal components (standardized variables)

Variable	e_1	e_2	e_3
DT(z_{1p})	-0.5640	0.6630	0.4920
FL(z_{2p})	-0.6170	0.0580	-0.7850
SE(z_{3p})	-0.5490	-0.7470	0.3760
Variance	2.4042	0.5360	0.2597
Cumulative percentage of total variance	73.5000	91.3000	100.00

Thus, using the coefficients in Table 1, we have the principal components as

$$Y_{1p} = e_i z_{ij} = e_{11} z_1 + e_{12} z_2 + e_{13} z_p$$

Therefore

$$Y_{1P} = -0.5640z_1 - 0.6170z_r - 0.5490z,$$
$$Y_{2P} = e_{21}z_1 + e_{22}z_2 + e_{23}z_3$$
$$= 0.6630z + 0.0580 - 0.7470z_3$$
$$Y_{3p} = e_{31}z_1 + e_{22}z_2 + e_{33}z_3$$
$$= 0.4920z_1 - 0.7850z_2 + 0.3760z_3$$

The proportion of total variance accounted for by the first PC is

$$\frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3} = \frac{2.2042}{3} = 0.7350$$

and the proportion of total variance accounted for by the second PC is

$$\frac{\lambda_2}{\lambda_1 + \lambda_2 + \lambda_3} = \frac{0.5360}{3} = 0.1760$$

Therefore the proportion of total variance accounted for the by the first two PCs = (0.7350 + 0.1790)=0.914 or 91.4%. Hence the first two PC's Y_{1P} and Y_{2p} replace the components since they account for more than 80% of the total variance.

We observe that in the second PC, production of flour and the Down Time (DT) are high while the production of Semovita is low. This indicates that probably, as the time of production of Semovita went on, there was a breakdown of production started after some minutes. Following from the foregone results of the analysis, we conclude that Down Time does not have significant impact on the production of wheat products.

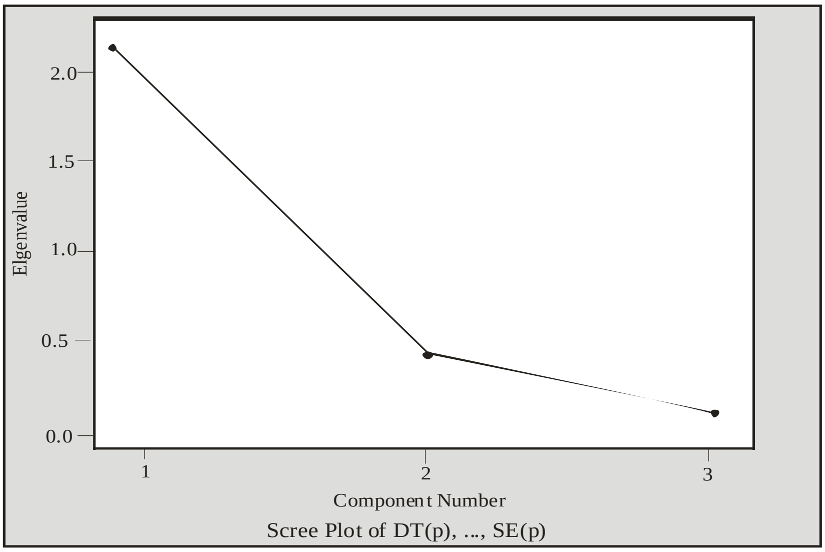

Figure 1: Scree plot of eigenvalues against DT, FL and SE

In Figure 1, the eigenvalues are plotted from the largest to the smallest, thus there is a clean break between the first two eigenvalues, and these two eigenvalues account for 91.4% of the total variability.

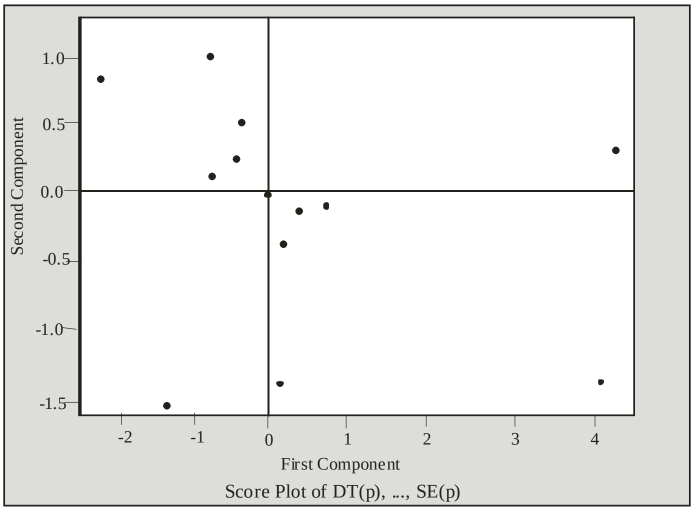

Figure 2: Score plot of first component against second component of DT, FL and SE

In Figure 2, dots indicate the observation of the variables for 12 months. The first two columns show the projection of the data (variables) onto the first two principal components. The plots of Figure 3 show that all three variables have low vector loadings for the first component. In the second component, DT and FL have positive (high) vectors loadings, while SE have negative vector loadings.

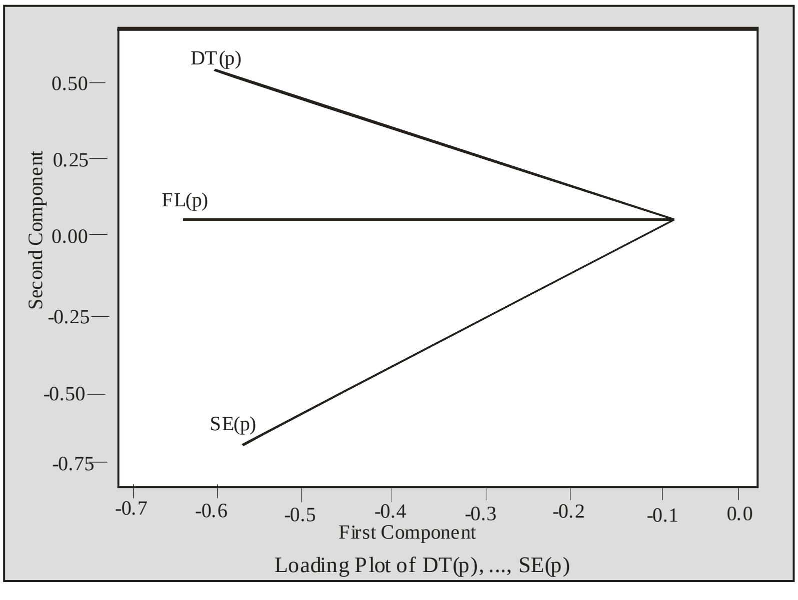

Figure 3: Loadings of DT, FL and SE, first component against second.

The analysis of this work was based on the merged pooled data of the products together with their corresponding downtime given in Table 2.

Table 2: Merged pooled data of product and their corresponding down time for 2004 - 2006

MONTH	DT(p)	FL(p)	SE(p)
Jan.	-2.4576	-3.0743	-37062
Feb.	-0.5751	-0.3724	-0.501
Mar.	0.6195	0.5745	-0.4949
Apr.	-0.488	0.7892	-04193
May	1.2774	0.7921	-0.4648
Jun.	0.3304	0.8821	0.107
Jul.	2.1638	1.2216	0.911
Aug.	-0.1234	0.6981	2.7033
Sep.	-1.2764	0.1695	1.13
Oct.	0.363	-1.5856	0.4796
Nov.	-0.2829	-0.3787	0.2499
Dec.	0.4492	0.2837	0.0056

CONCLUSION

This study was focused on determining the impact of down time on the production of wheat products, using the principles of principal component analysis. Results of the analysis have revealed that low Down Time corresponds to a high production of flour and low production of Somevita and vice versa. This implies that, on the average, there is no significant impact of Down Time on the production of wheat products.

REFERENCES

Jyh-cherng, J. and Samuel, K. (1999): On a relation between principal; components and regression analysis. The American Statistician, 53:256-286.

Raychaudhuri, S.; Stuart, J. M., and Altman, R. B. (2000). Principal Component Analysis to Summarize Micro-array experiments. Application to sporulation time series. Pacific Symposium on Biocomputation, 2: 311 – 321.

Richard, A. J. and Dean, W. W. (1992): Applied Multivariate Statistical Analysis. New Jersey: Prentice-Hall.

Yeung, K. Y. and Ruzzo, W. L. (2001): Principal Component Analysis for clustering gene expression data. Bioinformatics, 17: 763 – 774.