Available online at http://www.ajol.info/index.php/njbas/index
Nigerian Journal of Basic and Applied Science (June, 2015), 23(1): 75-79
DOI: http://dx.doi.org/10.4314/njbas.v23i1.11
ISSN 0794-5698

Short Communication

Principal Components as a Tool in Statistical Quality Control: A Case Study of Sokoto Portland Cement

***A.B. Zoramawa and Y. Musa**
Department of Mathematics, Statistics Unit, Usmanu Danfodiyo University, P.M.B. 2346, Sokoto, Nigeria
[* Corresponding author: aminubz@gmail.com : ☎: +2348064963620]

ABSTRACT: The principal components analysis and Hotelling's T^2 statistic are used in studying the clinker minerals of Portland Cement Company of Northern Nigeria (CCNN). The cement produced is considered to be a mixture of eight minerals, each being sensitive to the presence of fly ash. Effort was made in this paper to show that the dependence on only one mineral in detecting fault or diagnosing noise was not effective but rather using the entire components such that each component's contribution to variation is measured thereby making a valid interpretation about fault.
Keywords*: Hotelling's T^2 approximated model, Fault detection, PCA, Components treatments*

INTRODUCTION

Jackson (1991) reported the increasing needs for principal component analysis and Hotelling's T^2 approximated model, due to their efficiency in handling large number of highly correlated variables, measurement errors and missing data. Principal Components Analysis (PCA) is used to solve several tasks including, data rectification (Kramer and Mah, 1994), gross error detection (Tong and Crowe, 1995), disturbance detection and isolation (Ku et al., 1995), statistical process monitoring (Wise et al., 1990), and fault diagnosis (Dunia et al., 1996; MacGregor et al., 1994). If the measured variables are linearly related and are contaminated by errors, the first few components capture the relationship between the variables, and the remaining components are comprised only of the error. Thus, eliminating the less important components reduces the contribution of errors in the measured data and represents it in a compact manner. Applications of PCA rely on its ability to reduce the dimensionality of the data matrix while capturing the underlying variation and relationship between the variables (Jollife, 2002).

According to Kresta (1994) and Kourti (1996), faults detection is improved by making use of Hotelling's T^2 statistic and data dimensionality reduction technique of PCA and canonical variate analysis (CVA). The lower dimensional representations of the technique can be generalized without need for entire dimensionality. It is observed that quality of Portland cement clinker depends on its chemical and mineralogical composition. Clay contains basically three oxides: SiO, AlO and FeO. Limestone decomposes to CaO and CO_2 during firing. CO_2 is removed and CaO reacts to form alite ($3CaO.SiO_2$), belite ($2CaO.SiO_2$), celite ($3CaO.AlO_3$), and tetra calcium-alumino-ferrite ($4CaO.AlO_3$. FeO_3), abbreviated as C_3S, C_2S, C_3A and C_4AF respectively. The composition was reported to be: 45-65%C_3S, 15-35%C_2S, 4-14%C_3A, and 10-18%C_4AF (Komar, 1987).

MATERIALS AND METHODS

The data used in this study was a secondary dataset obtained from Quality Control Division of Cement Company of Northern Nigeria (CCNN). The data consisted of the clinker minerals of aluminum oxide (Al_2O_3), iron oxide (FeO_3)), magnesium oxide (MgO), sulphate (SO_3), phosphorous oxide (P_2O_5), calcium oxide (CaO), silica oxide (SiO_2), potassium oxide (K_2O) and loss on ignition (LOI).

Principal component analysis procedure

Given a data matrix X constructed by *m* observation of *n* variables, PCA projects it to a lower dimensional space that explains a large fraction of variability in the original data (Jackson, 1980, 1991 and Kresta *et al,* 1991). Each pair consists of a vector in *n* called the loadings, p_i, and a vector in *m* referred to as the scores, t_i Thus **X** can be written as:

$$X = t_1 p_1^T + t_2 p_2^T + \ldots + t_n p_n^T + E = \sum_{i=1}^{k} t_i p_i^T + E \qquad (1)$$

where E is the residual matrix.

The matrix of loadings vectors **P** forms a new orthogonal basis for the space spanned by **X** and the individual p_i are the eigenvectors of the covariance matrix of **X,** defined as:

$$\mathrm{cov(X)} = \frac{1}{m-1}\left(X^T X\right) \quad (2)$$

Thus

$$\mathrm{cov(X)}p_i = \lambda_i p_i \quad (3)$$

where λ_i is the eigenvalues associated with the eigenvector $\mathbf{p_i}$. The loadings vectors $\mathbf{P_i}$ are often referred to as principal components. Each of the $\mathbf{t_i}$ is simply the projection of **X** onto the new basis vector $\mathbf{p_i}$:

$$t_i = Xp_i \quad (4)$$

The value of each λ_i is an indicator of the covariance in the data set in the direction $\mathbf{p_i}$. In fact fraction variance in direction $p_i = \dfrac{\lambda_i}{\sum \lambda_i}$ (5)

In a data set scaled to have variables of zero mean and unit standard deviations

$$\sum \lambda_i = n \quad (6)$$

where *n* is the number of variables in the data set. In this case, each of the scores vectors $\mathbf{t_i}$ would have mean equals to zero and standard deviation equal to $(\sqrt{\lambda_i})$. closely related to Strang (1980) Singular Value Decomposition (SVD). **X** is decomposed as

$$X = USV^T \quad (7)$$

where **V** contains the eigenvectors (**p**i) and **S** is a diagonal matrix containing the square roots of the eigenvalues (the singular values) of the covariance matrix of **X**.

Once the eigenvectors have been determined using PCA or SVD, projections of the data onto the eigenvectors can be made. These projections are commonly referred to as "scores plots" and are often useful for showing the relationships between the samples (rows) in the data set. Plots can be done as the projections of the samples onto a single eigenvector versus sample number (or time) or onto the plane formed by two eigenvectors. A projection of the samples onto the two eigenvectors associated with the largest eigenvalues depicts the largest amount of information about the relationship between the samples that can be shown in two (linear) dimensions. It is for this reason that PCA is often used as a pattern recognition and sample classification technique.

Plots of the coefficients of the eigenvectors, known as "loadings plots", show the relationships between the original variables in the data set. Correlations between variables are shown. Hotelling (1933, 1947) and Jackson (1980) provided Hotelling's T^2 for identifying unusual variability within the normal subspace. The value of T^2 for one sample is equal to the sum of squares of the adjusted (unit variance) scores on each of the PCs in the model. That is:

$$T^2 = \sum_{i=1}^{k}\left(\frac{t_i}{\lambda_i}\right)^2 = \sum_{i=1}^{k}\left(\frac{t_i^2}{s_i^2}\right) \quad (8)$$

Here k is the number of principal components retained in the model. T^2 is the squared length of the projection of the current sample into the space spanned by the PCA model.

This square is a measure of how far the PCA estimate of the sample (as given by equation 9) is from the data containing the multivariate mean. The statistical confidence limits for T^2 can be calculated by using statistical F-distribution as follows:

$$T^2_{\alpha,m,k} = \frac{k(m-1)}{m-k} F_{\alpha,m-k,k} \quad (9)$$

where *m* is the number of samples in the data set used in the calculations involved in PCA model, *k* is the number of principal component vectors retained and α corresponds to the standard normal deviate.

T^2 statistic measures the variations inside the state space. Then process faults are detected, selecting a level of significance and using T^2_α to compute the appropriate threshold.

RESULTS AND DISCUSSIONS

With the eigenvectors as loadings of the principal components, spanning the new PCA coordinate system, a composition of principal components proportion reveals a clear picture of the variables that capture and contribute high variation in the data (see Table 1 and Figure 1 for the composition of PCA and scree plot of the eigenvalues for the clinker minerals). The result supports that component 1 has the highest variation and contains most of the mineralogical and chemical compositions. Three basic oxides were identified as overall strength of the cement and which also determine its quality (CaO, SiO_2 and Al_2O_3).

Table 1 reveals that Components 1 and 2 account for 27% and 18% respectively, with a corresponding values of 1.57 and 1.26. Even though components with lessthan 30% are not suppose to be retained, Kaiser

(1960) argue that components that displays an eigenvalue(s) greater than 1.00 are enough to account for a greater amount of variance than had been contributed by one variable. Therefore, in this study, only four components were considered.

Table 1: Composition of the Principal Components Proportion

Proportion	Std.dev	Propt.	Cumm
Comp.1	1.57	0.27	0.27
Comp.2	1.26	0.18	0.45
Comp.3	1.07	0.12	0.58
Comp.4	1.03	0.11	0.7
Comp.5	0.96	0.1	0.8
Comp.6	0.86	0.08	0.88
Comp.7	0.71	0.06	0.93
Comp.8	0.58	0.04	0.97
Comp.9	0.49	0.04	1

Cattell (1966) suggest the use of scree test in determining the number of PCs to retain, that is by making the use of the plot of eigenvalues associated with each component and look for a "break" between the components with relatively large eigenvalues and those with small eigenvalues. The components that appear before the break are assumed to be meaningful and are retained for rotation (Figure 1). Those that appeared after the break are assumed to be unimportant and are not retained. Outliers are detected with the use of biplot of the nine clinkers as shown in Figure 2. In this case, only Cao is considered because is the one that conforms to the negative correlation among the entire clinker variables. Jollife (2002) suggest that when there are more outliers and the *p* is not too large, turning to correlation coefficient is preferred. Table 2 shows the correlation matrix of the nine clinker variables.

Table 3 indicates the compositions of principal components. The first component that accounts for 27% of the total variance has a linear combination of:

$$PC1 = 0.132A - 0.957C + 0.146F + 0.018K + 0.077M + 0.015P + 0.19Si + 0.03S + 0.002L$$

We notice that the PC1 is a contrast between the calcium (C) and silica, iron and aluminium oxide.

Defining T^2 to be the sum of squares of the adjusted (unit variance) as in equation (8), it was obtained and compared with the values of F-ratio according to equation (9) as follows

$$T^2 = \sum_{i=1}^{k}\left(\frac{t_i}{\lambda_i}\right)^2 = \sum_{i=1}^{k}\left(\frac{t_i^2}{s_i^2}\right) = 2.1658^2 = 4.68$$

Taking α = 0.05, the F-statistic value is 4.77 which is the upper 100 α % critical point of the F-distribution with *k, m-k* degrees of freedom. Having the test value of T^2 calculated greater than F- statistic presumes that a fault has occurred.

Table 2 reveals the values of correlations and Table 3 the covariance matrix of the clinker variables

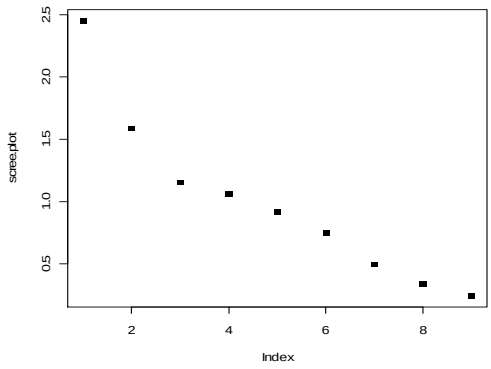

Figure 1: Scree-plot of the eigenvalues of the clinker minerals

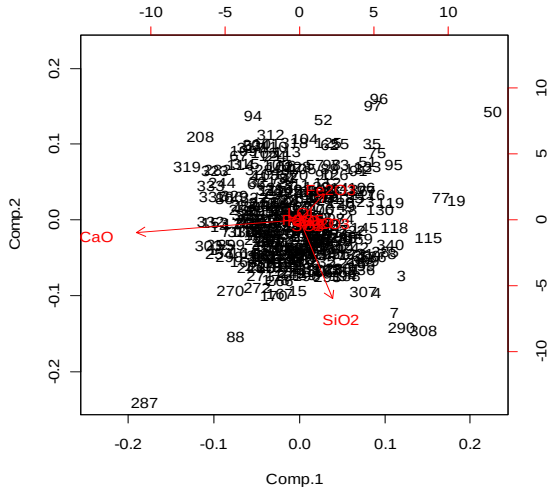

Figure 2: Biplot of the clinker minerals

Table 2: Correlation matrix of clinker variables

Variable	Al2O3	CaO	Fe2O3	K2O	MgO	P2O5	SiO2	SO3	LOI
Al_2O_3	1.00	-0.54	0.30	0.61	0.21	0.12	0.25	-0.06	-0.12
CaO	-0.54	1.00	-0.35	-0.32	-0.30	-0.17	-0.25	-0.19	-0.04
Fe_2O_3	0.30	-0.35	1.00	0.24	-0.04	0.49	-0.14	-0.11	0.04
K_2O	0.61	-0.32	0.24	1.00	0.06	0.17	-0.05	0.00	-0.29
MgO	0.21	-0.30	-0.04	0.06	1.00	-0.14	0.12	0.08	-0.04
P_2O_5	0.12	-0.17	0.49	0.17	-0.14	1.00	-0.01	-0.01	-0.02
SiO_2	0.25	-0.25	-0.14	-0.05	0.12	-0.01	1.00	0.11	-0.20
SO_3	-0.06	-0.19	-0.11	0.00	0.08	-0.01	0.11	1.00	-0.09
LOI	-0.12	-0.04	0.04	-0.29	-0.04	-0.02	-0.20	-0.09	1.00

Table 3: Eigen analysis of the covariance matrix of the clinker

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Al_2O_3	0.132	-0.018	-0.127	-0.295	0.186	0.561	-0.710	-0.000
CaO	-0.957	-0.146	-0.189	-0.113	-0.024	-0.019	-0.115	-0.003
Fe_2O_3	0.146	0.310	-0893	-0.090	-0.137	-0.216	0.008	0.107
K_2O	0.018	0.017	-0.019	-0.046	-0.026	0.144	-0.090	-0.085
MgO	0.077	-0.026	0.190	-0.899	-0.072	-0.375	0.037	-0.050
P_2O_5	0.015	0.024	-0.105	0.040	-0.033	-0.011	0.007	-0.989
SiO_2	0.192	-0.934	-0.273	0.031	0.041	-0.113	0.010	0.008
SO_3	0.039	-0.036	0.157	0.196	-0.758	-0.288	-0.525	0.017
LOI	0.002	0.088	0.062	0.204	0.602	-0.617	-0.444	-0.019

CONCLUSION

The analysis carried out reveals that about four components need to be retained in order to meet 70% total variance; this reflects the variation in overall quality of the Portland cement. The relationship between the aluminium and iron makes the process to be in control despite the presence of fault. Insufficient aluminium and Iron which may lead to difficulty in burning the clinker and excessive amounts also which may lead to low strength, due to dilution of the silicate by aluminates and ferrites, were avoided. The PCA exhibits higher sensitivity in detecting outliers than is the case with Hotelling's.

Acknowledgements: Gratitude is expressed to the Cement Company of Northern Nigeria (CCNN), Sokoto for being kind to make data available for studies.

REFERENCES

Cattell, R.B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research,* **1:** 245-276.

Dunia, R., Qin, S.J. Edgar, T.F. and McAvoy, T.J. (1996) Identification of Faulty Sensors Using Principal Component Analysis, *American Institute of Chemical Engineer's Journal*, **42(10):** 2797-2812.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components*, Journal of Educational Psychology,* **24(6):** 417-441.

Hotelling, H. (1947). Multivariate Quality Control Illustrated by air testing of sample bombsights," Techniques of Statis-tical Analysis, Mcgraw-Hill, New York, pp. 11–184.

Jackson, J.E. (1980). Principal Components and Factor Analysis, Part 1- Principal Components, *Journal of Quality Technology*, **12**: 201 - 213.

Jackson, J.E. (1991). *A User's Guide to Principal Components.* New York; John Wiley: 11-23.

Jolliffe I.T. (2002). *Principal Components Analysis*, 2nd edition, Springer Series, pp. 168

Komar A. (1987). *Building Materials and Components*. MIR Publishers, Moscow. Pp. 136

Kresta, J., MacGregor, J.F. and Marlin, T.E. (1991). Multivariate Statistical Monitoring of Processes. The Canadian Journal of Chemical Engineering, **69(1):** 35-47.

Kresta, J.V., Marlin, T.E., MacGregor, J.F. and Can, J. (1994). Development of inferential process models

using PLS. *Computers and Chemical Engineering* **18(7):** 597-611

Kramer, M.A. and Mah, R.S.H. (1994). Model-Based Monitoring, in *Proc. Second Int. Conf. on Foundations of Computer Aided Process Operations*, D. Rippin, J. Hale, J. Davis, eds. CACHE g 45

Ku, W., Storer, R.H. and Georgakis, C. (1995). Disturbance Detection and Isolation by Dynamic Principal Component Analysis, *Chemometrics and Intelligent Laboratory System*, **30:** 179-196

Kourti, T. and MacGregor, J.F. (1996). Recent developments in multivariate SPC methods for monitoring and diagnosing process and product performance. *Journal of Quality Technology.* **28:** 409 – 428

MacGregor, J.F., Jaeckle, C., Kiparissides, C. and Koutoudi, M. (1994). Process Monitoring and Diagnosis by Multiblock PLS Methods, *American Institute of Chemical Engineer's Journal*, **40:** 5, 827

Strang, G. (1980). *Linear Algebra and Its Applications*, Academic Press New York.

Tong, H. and Crowe, C.M. (1995). Detection of Gross Errors in Data Reconciliation by Principal Component Analysis, *American Institute of Chemical Engineer's Journal*, **4(7):** 1712-1722.

Wise, B.M., Ricker, N.L., Veltkamp, D.F. and Kowalski, B.R. (1990). A Theoretical Basis for the use of Principal Component Models for Monitoring Multivariate Processes, *Process of Quality Control.*, **1:** 41-51.