JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 2818 - 2831
https://publishoa.com
ISSN: 1309-3452

Improving the Efficiency Measurement Index Using Principal Component Analysis (PCA)

[1]Lecturer, Ahmed Husham Mohammed, [2]Prof. Dr. Marwan Abdul Hameed Ashour

[1] University of Basrah, College of Administration & Economics, PhD Student Department of Statistics, University of Baghdad ,Iraq , Email: ahmed.albasrai@uobasrah.edu.iq
[2] University of Baghdad, College of Administration & Economics, Department of Statistics,Iraq

ABSTRACT

This study aims at presenting an approach capable of improving the efficiency measurement index using the Principal Component Analysis. The main reason for the adoption of this approach in measurement process lies in the fact that during an analysis of data envelopment, there is often discrimination problems among measurement units, whether a measurement unit is efficient or not, especially there is a large number of variables (inputs and outputs). As to the size of a sample to be studied, and for addressing this problem, the Principal Component Analysis is adopted in this study, which is deemed one of the important statistical tools in decreasing original variable dimensions. For obtaining new variables representing principal components, factors for each variable have been identified, which works on determining its effect. Therefore, this study brings to light an integrated approach between two procedures: PCA-DEA, taking part in enhancing results of the efficiency measurement index. This approach has been applied to a problem related to the state financial management of the Iraqi Budget for the years (2005 – 2019). As per the decrease principal accomplished through the use of PCA, efficient years are identified.

Key Word: Multivariate, PCA, DEA, Efficiency

1. Introduction

First of all, the Principle Component Analysis is considered one of the statistical tools of multi-variables concerned aimed at treating and analyzing high-dimensioned factors through reducing dimensions for a phenomenon's variables to be studied. As it is well-known building up statistical and mathematical models via making use of high-dimensioned variables is a bit hard pertaining to analysis and interpretation of because there is no capability to follow up variables accurately. In addition, there are problems might appear at times of studying systems where its number of variables is large compared to the size of a sample studied. Also, in these cases, estimation and testing processes will be impossible or much more directly, results arrived at will be inaccurate or of no value. This could be seen when analyzing Data Envelopment Analysis (DEA), which deemed one of the fractional mathematical procedures caring measurement of a proficiency index for any phenomenon strongly. However, despite its advantages, there lies a significant problem that may face this procedure that brings about failure or unreliability of its outcomes.

Therefore, due to the importance of this topic, there have been attempts by a number of researchers to come up with studies, and attempts for the purpose of addressing the problem of high-dimensional phenomena in order to improve results. One of these studies a study conducted by Azadeh, Ghaderi, & Ebrahimipour (2007), which made a scientific contribution on the employment of data envelopment analysis, principal components and methods of numerical classification of manufacturing systems based on equipment performance indicators. The main factors were identified in the evaluation process, namely downtime, repair time, average time between failure and operating time, added value added and production value as shaping factors. This study was able to arrange the sectors and identify strengths and weaknesses for each of the sectors.

Besides, in a study carried out by Pin Fu & Ruey Ou, in 2013, a new method through was presented in which project performance can be accurately evaluated to enhance the efficiency measurement index. The study was based on energy projects affiliated to the Energy Office of the Ministry of Economic Affairs in Taiwan. Whereas, the results of evaluating the performance of energy projects through the integrative methodology between the PCA and DEA methods showed an improvement in the efficiency measurement index compared to the application of data envelope analysis DEA alone.

In addition, in 2014, both Põldaru & Roots conducted a study applied an integrative approach between the two PCA-DEA methods to assess quality of life scores in Estonian counties. The study included (15) Estonian provinces for the period from 2000 to 2011. The paper was able to reach the efficiency of the method used to identify and arrange the provinces within the specified period.

Further, a study implemented by Azadeh, Nasirian, Salehi and H. Kouzehchi (2016) presented an integrative approach between two methods: PCA-DEA with the aim of improving Six Sigma results for the auto industry, where the two methods PCA-DEA were adopted in determining the performance of subgroups for a number of employees according to the preparation of a standard form and then benefit from the results achieved in improving the results of Six Sigma.

Likewise, the study of Davoudabadi, Mousavi and Sharifi (2020) included evaluating a series of suppliers by adopting a number of criteria, the most important of which are cost and supply time, where two methodologies were adopted in the analysis process, namely PCA-DEA, in order to evaluate the performance of suppliers.

Also, Karami, Yaghin and Mousazadegan in (2020) conducted a study clarified the main role of business management policy decisions that depend mainly on product quality, popularity and company reputation and the promotion of rapid response and its reflection on the performance and satisfaction of suppliers of raw materials for the garment industry because of its importance in determining customer satisfaction. For the purpose of accomplishing this, three methodologies were adopted. The first methodology is to analyze the principal components of PCA. The second is to identify the competent suppliers by adopting the DEA method, and the third is to use the VIKOR method to arrange the supplier chain numerically.

And Layeb, Omrane, Siala, and Chaabani (2020) carried out a study contributed to the presentation of the distribution center (DC) assessment approach, which is a basis and foundation in evaluating the performance of operators in developing countries. The study was applied within a realistic case of a third-party logistics service provider (3PL). This third party logistics service provider works in Tunisia. The study showed that the main problem here is how to choose the appropriate efficiency indicators for these systems. This problem was addressed by adopting an integrative mechanism between the PCA and DEA method. The study was built through evaluation in two directions (warehousing activities and transportation activities). Indicators for each activity were diagnosed, and then the performance of these systems evaluated.

Finally, the study of Wu, Chung and Huang (2021) had a purpose to use several multivariate methods in order to conduct a comprehensive and systematic evaluation of the overall performance of 125 for the period (1997-2017) by highlighting the spatial differences, as well as changes and trends in the performance of energy security. The methods were PCA analysis and data envelopment analysis with the presence of the warranty area index, and then the adoption of the clustering method by adopting the results of the ESP classification.

1.1 Method and Material

1.1.1 Principal Component Analysis

Principal component analysis is known as the Karhunen-Loév transformation (KTL) or the Hotelling transformation. This method was first proposed by Pearson in (1901), who was concerned with the geometrical form to address the problem of higher dimensions. It was developed in (1933) by Hotelling through adopting the method Algebraic, that the principal components method transforms the coordinate system of the original data into composites so that the largest variance is along the first coordinate, and the second largest variation in the data is along the second coordinate, and thus these coordinates are called the principal components (Hochreiter, 2013, p. 55).

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 2818 - 2831
https://publishoa.com
ISSN: 1309-3452

Therefore, this method is one of the most important statistical methods, and a special case of factor analysis, and the most common in multivariate data processing, through the formation of a linear set of random variables that have certain characteristics in terms of variance conditions (Anderson, 2003, p. 459). Hence, the goal can be set. The main application of the PCA method is to define a new linear set of original variables that helps improve the discriminability of the model even though the model will lose some of its explanatory power.

1.2.2 Building Main Components Analysis Model

Suppose we have P of random variables $\underline{X}' = [X_1, X_2, \dots, X_p]$that have a multivariate normal distribution (MVN) (Johnson & Wichern, 2002, p. 149) as building a PCA model requires estimating the covariance matrix Σ, and calculating the roots and eigenvectors (Eigen Values and Eigen Vector), and accordingly, the information matrix that has the degree (n×p) can be written as follows:

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \quad \dots (1)$$

The covariance matrix Σ can be obtained by applying the covariance and common covariance formulas.

$$\boldsymbol{S_i} = \frac{1}{n-1}\left(x_{ij} - \bar{x}_i\right)^T\left(x_{ij} - \bar{x}_i\right) \quad \dots (2)$$

$$\boldsymbol{S_{ik}} = \frac{1}{n-1}\left(x_{ij} - \bar{x}_i\right)^T\left(x_{kj} - \bar{x}_k\right) \quad ; i \neq k \quad \dots (3)$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1k} \\ S_{21} & S_{22} & \cdots & S_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ S_{k1} & S_{k2} & \cdots & S_{kk} \end{bmatrix} \quad \dots (4)$$

After determining the covariance matrix according to formula (4), the eigenvalues and eigenvectors will be calculated as below:

$$\left|\boldsymbol{\Sigma} - \lambda_j \mathbf{I}\right| = \mathbf{0} \quad \dots (5)$$

$$\left|\boldsymbol{\Sigma} - \lambda_j \mathbf{I}\right| v = \mathbf{0} \quad \dots (6)$$

Through formulas (4, 5 and 6), the main compounds can be executed:

$$\mathbf{C}_{PC_j} = e_{1j}X_1 + e_{2j}X_2 + \cdots + e_{ij}X_j \quad \dots (7)$$

Whereas:

i: represents the observations i=1,2,…,n

j: represents the number of variables j=1,2,…,p

$\mathbf{C}_{PC_j}$**:** represents the main compounds. In other words, they are the new realized variables, so that each new variable is a linear combination of the original variables.

And the constraints associated with the main compounds achieved according to formula (7) are:

$$Var\left(\mathbf{C}_{PC_j}\right) = max \quad \dots (8)$$

$$e^T e = 1 \quad \dots (9)$$

Where formula (8) shows that the best component is the one that represents the highest variance, while formula (9) shows that the sum of the squares of the eigenvector values must be equal to one (Põldaru & Roots, 2014, p. 67), or by adopting the Kaiser Guttman (KG) test, proposed by Guttman in 1954. It disclosed that the identification of the principal important components is reliant on the eigenvalues verified from the matrix of correlation coefficients according to the following condition (Peres-Neto, Jackson, & Somers, 2005, p. 976)**:**

$$\mathbf{PC}_j = \begin{cases} \lambda_j > 1 \;; Retaind\ the\ PC \\ Trivial \qquad ; e/w \end{cases} \qquad \dots(\mathbf{10})$$

1.2.3 Data Envelopment Analysis

This was proposed by Charnes et al. in 1978 with the purpose of expanding the principle of measuring the efficiency index by adopting multi-dimensional variables in terms of inputs and outputs. It is also considered a non-parametric technique because it does not require strict assumptions that limit its flexibility (Zhu, Aparicio, Li, Wu, & Kou, 2022, p. 927).

Much attention has been paid to this method in addressing problems related to performance evaluation and identifying deficiencies and ways of treatment to reach full efficiency, but when referring to reality, there are many problems that may face DEA technology, the most important of which are the structural complexities of the target sample and data irregularities (Zhu & Cook, 2007, p. 1).

1.2.4 Data Envelopment Model Establishment

To establish a DEA model, we assume that we have DMU_j from the decision-making units that we have (x_{ij}) representing the system input variables and (y_{rj}) representing the system output. So, we can get the formula for measuring efficiency (Banker, Charnes, & Cooper, 1984, p. 1078):

$$\boldsymbol{Ef_o} = \frac{\sum_{r=1}^{s} \boldsymbol{u_r y_{ro}}}{\sum_{i=1}^{m} \boldsymbol{v_i x_{io}}} \qquad \dots(\mathbf{11})$$

where $\boldsymbol{Ef_o}$ is the degree of efficiency for the hypothetical decision-making unit

The Variance Return Scale (VRS) model developed by Banker (1984) is considered the best in determining the degree of efficiency because it is based on an important assumption that the decision-making units DMU_j work within different environments and achieve efficiency under three directions of return for volume (decreasing, constant, and high) and within these parameters, the VRS model (Paradi, Sherman, & Tam, 2018, p. 9) can be established according to two orientations**:**

First: Input Oriented VRS

The input orientation in the DEA model is based on the principle of reducing inputs and achieving the same actual amount of output, so an objective function will be of type *Min*, as shown below:

$Min\ \theta$

Subject to:

$$\begin{aligned} &\sum_{j=1}^{n} \propto_j x_{ij} \le \theta x_o \\ &\sum_{j=1}^{n} \propto_j y_{rj} \ge y_{ro} \qquad \dots(\mathbf{12}) \\ &\sum_{j=1}^{n} \propto_j = 1 \\ &\propto_j \ge \mathbf{0} \end{aligned}$$

Second: Output Oriented VRS

The output orientation in the DEA model is based on the principle of maximizing outputs by using what is available from the inputs. So, an objective function will be of type *Max*, as shown below:

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 2818 - 2831
https://publishoa.com
ISSN: 1309-3452

$Max\ \phi$

Subject to:

$$\sum_{j=1}^{n} \propto_j y_{rj} \geq \phi y_{ro}$$
$$\sum_{j=1}^{n} \propto_j x_{ij} \leq x_{io} \qquad \dots (\mathbf{13})$$
$$\sum_{j=1}^{n} \propto_j = 1$$
$$\propto_j \geq \mathbf{0}$$

(Johnes, 2006, p. 276)

whereas:
x_{ij}: represents the model's input to variable (i) for the decision-making unit (j).
y_{rj}: represents the model's output of the variable (r) for the decision-making unit (j).
x_{io}: represents the energy available from the target unit input.
y_{ro}: represents the power available from the target unit output.
$j = 1,2, \dots, m$: represents the number of decision-making units.
$i = 1,2, \dots, k$: represents the number of system entries.
$r = 1,2, \dots, b$: represents the number of system outputs.

1.2.5 Proposed Methodology PCA – DEA

One of the most important conditions for applying the data envelopment analysis model to reach reliable results is that reality $\boldsymbol{m} \geq \mathbf{2}(\boldsymbol{k}+\boldsymbol{b})$ is achieved **(Layeb & etal, 2020, p. 296**). This reality is often violated as a result of the presence of variables that may exceed the sample size. Several scenarios aim to reduce the number of variables, including Principal Compound Analysis (PCA) and Hierarchical Analysis Processes (AHP). Therefore, in this study, there is an intention to enhance DEA results through resorting to the Principal Components Analysis (PCA).

2.1 Experimental Methodology

The study proposal was applied to the public financial management data to determine the financial reform policies in the Iraqi financial budget, as financial policies are one of the important pillars in managing the state's financial resources and drawing government spending policies on its various sectors. The efficiency of spending policies was studied through spending on the infrastructure, education, and the health sectors through some criteria related to expenditure. Four criteria were defined for each of the sectors referred to for the period from (2005-2019) and the proven data were used in the source (Al-taee, 2021). It is worth mentioning that the variable (the number of published research) has been removed from the education sector due to its incompatibility with the sector's variables (the enrollment coefficient in primary education, the literacy coefficient, and the secondary education enrollment coefficient):

2.2 Estimation for Mean Vector and Covariance Matrix for Each Sector

Block	Mean Vector	Cov. Matrix
Infrastructure	$\begin{bmatrix} \mathbf{20.62} \\ \mathbf{1.98} \\ \mathbf{85.72} \\ \mathbf{91.788} \end{bmatrix}$	$\begin{bmatrix} \mathbf{1.79171} & \mathbf{0.469} & \mathbf{5.50227} & \mathbf{3.08511} \\ \mathbf{0.469} & \mathbf{0.37171} & \mathbf{4.29659} & \mathbf{2.44553} \\ \mathbf{5.50227} & \mathbf{4.29659} & \mathbf{51.7947} & \mathbf{29.4955} \\ \mathbf{3.08511} & \mathbf{2.44553} & \mathbf{29.4955} & \mathbf{16.8009} \end{bmatrix}$

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 2818 - 2831
https://publishoa.com
ISSN: 1309-3452

Education	$\begin{bmatrix} 91.43 \\ 406.71 \\ 49 \end{bmatrix}$	$\begin{bmatrix} 15.90 & 65.48 & 21.18 \\ 65.48 & 1624599.42 & 2961.09 \\ 21.18 & 2961.09 & 61.04 \end{bmatrix}$
Healthy	$\begin{bmatrix} 29.81 \\ 68.37 \\ 69.27 \\ 1.33 \end{bmatrix}$	$\begin{bmatrix} 97.65 & -46.51 & -71.55 & -0.20 \\ -46.51 & 35.29 & 45.77 & 0.11 \\ -71.55 & 45.77 & 174.35 & 0.01 \\ -0.20 & 0.11 & 0.01 & 0.002 \end{bmatrix}$

Source: Set by the Researcher as per results of SPSS V.26

2.3 Testing Quality and Suitability of Analysis Data

This test requires estimating the correlation coefficients matrix to determine the values of KMO and Bartlett Shpericity. The test results will be shown as in the following tables:

Table 1: Correlation Coefficients Matrix for Infrastructure Sector along with Tests of Data Suitability for Analysis

Corr. Matrix for Infrastructure

$$\begin{bmatrix} 1 & 0.5747 & 0.5712 & 0.5623 \\ 0.5747 & 1 & 0.9792 & 0.9786 \\ 0.5712 & 0.9792 & 1 & 0.9999 \\ 0.5623 & 0.9786 & 0.9999 & 1 \end{bmatrix}$$

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		**.662**
Bartlett's Test of Sphericity	**Approx. Chi-Square**	**148.691**
	Df	**6**
	Sig.	**.000**

Source: Set by the Researcher as per results of SPSS V.26

Results shown in table (1) brought to light that the data is appropriately suited to the analysis, as the KMO test achieved = 0.662, and the Bartlett Shpericity test achieved $\chi^2 = 148.691$ with a level of significance Sig = 0.000. This is evidence of the significance of the analysis and its consistency with the hypothesis setting out that the matrix of correlation coefficients is not isolated alone.

Table 2: Correlation Coefficients Matrix for Education Sector along with Tests of Data Suitability for Analysis:

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 2818 - 2831
https://publishoa.com
ISSN: 1309-3452

Corr. Matrix for Education

$$\begin{bmatrix} 1 & 0.0129 & 0.6799 \\ 0.0129 & 1 & 0.2974 \\ 0.6799 & 0.2974 & 1 \end{bmatrix}$$

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.430
Bartlett's Test of Sphericity	Approx. Chi-Square	9.598
	Df	3
	Sig.	.022

Source: Set by the Researcher as per results of SPSS V.26

Results given in table (2) made it clear that the data related to the education sector are poorly suited to the analysis, as the KMO test achieved = 0.430, and the Bartlett Shpericity test achieved $\chi^2 = 21.785$ with a level of significance Sig = 0.022. This is evidence of the significance of the analysis and its consistency with the hypothesis setting out that the matrix of correlation coefficients is not isolated alone, Besides, the reason for the weakness of the education sector data can be explained from an analytical point of view, where it is possible to choose or enhance variables with more influential ones (number of teachers, buildings, classrooms, etc.) as well as inclusion aspects related to higher education, etc.).

Table 3: Correlation Coefficients Matrix for Health Sector along with Tests of Data Suitability for Analysis:

Corr. Matrix for Healthy

$$\begin{bmatrix} 1 & -0.7923 & -0.5483 & -0.4383 \\ -0.7923 & 1 & 0.5835 & 0.4212 \\ -0.5483 & 0.5835 & 1 & 0.0228 \\ -0.4383 & 0.4212 & 0.0228 & 1 \end{bmatrix}$$

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.679
Bartlett's Test of Sphericity	Approx. Chi-Square	21.236
	Df	6
	Sig.	.002

Source: Set by the Researcher as per results of SPSS V.26

Results shown in table (3) stating that data are appropriately suited to the analysis, as the KMO test achieved = 0.679, and the Bartlett Shpericity test achieved $\chi^2 = 21.236$ with a level of significance Sig = 0.002. This is evidence of the significance of the analysis and its consistency with the hypothesis setting out that the matrix of correlation coefficients is not isolated alone.

2.4 Identification of Principal Components PC_J

At this phase, the principal components for each sector will be identified, so that each component is a linear combination of all target variables according to the variances, as follows:

- Identification of Principal Components for Infrastructure Sector

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 2818 - 2831
https://publishoa.com
ISSN: 1309-3452

Table 4: Eigenvalues and its Variances for Infrastructure Sector

Total Variance Explained						
	Initial Eigenvalues			**Extraction Sums of Squared Loadings**		
Component	**Eg. Value**	**% of Variance**	**Cumulative %**	**Total**	**% of Variance**	**Cumulative %**
1	**3.380**	**84.510**	**84.510**	**3.380**	**84.510**	**84.510**
2	**.592**	**14.791**	**99.301**			
3	**.028**	**.698**	**99.998**			
4	**6.253E-5**	**.002**	**100.000**			
Extraction Method: Principal Component Analysis.						

Source: Set by the Researcher as per results of SPSS V.26

Table (4) summarizes that the number of the achieved principal compounds is (4) compounds, and that the best component is the one corresponding to $\lambda_1$which accomplished an explanatory variance (84.5%). This means that the first component was able to represent the data well and to reflect the significant and influential variables.

As for the number of important components that can be selected, it was one component based on the KG standard, which fulfills the requirement according to the formula (10) where $\lambda_1 = 3.380 > 1$. The figure below shows the distribution of the principal components according to the eigenvalues as per the Scree Plot method:

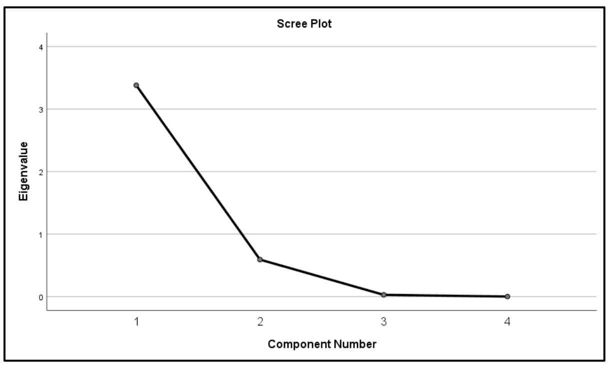

Figure (1): Number of Principal Components for Infrastructure Sector
Source: Set by the Researcher as per results of SPSS V.26

Figure (1) depicts that the number of the significant influential principal components are the ones whose value is greater than one. There was one component that corresponds to $\lambda_1 = 3.380$, and this goes consistent with the KG test.

Table 5: Factors of Infrastructure Component

Component Coefficient	**Agricultural lands**	**An individual's share of electricity**	**Individuals using sewage**	**Individuals consuming drinkable water**
	.704	**.978**	**.983**	**.981**
Extraction Method: Principal Component Analysis. a. 1 components extracted.				

As enumerated in table (5), all variables coefficients are $a_{ij} \geq 0.70$. This is evidence of significance of infrastructure variables and necessity for including them in the analysis.

- Identification of Principal Components for Education Sector

Table 6: Eigenvalues and its Variances for Education Sector

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 2818 - 2831
https://publishoa.com
ISSN: 1309-3452

Total Variance Explained						
	Initial Eigenvalues			**Extraction Sums of Squared Loadings**		
Com.	**Eg. value**	**% of Variance**	**Cum. %**	**Total**	**% of Variance**	**Cum. %**
1	**1.747**	**58.229**	**58.229**	**1.747**	**58.229**	**58.229**
2	**.991**	**33.018**	**91.247**			
3	**.263**	**8.753**	**100.000**			
Extraction Method: Principal Component Analysis.						

Source: Set by the Researcher as per results of SPSS V.26

As table (6) classified, the number of the realized principal components are (3), and that the best component is the one corresponding to $\lambda_1$which achieved an explanatory variance (58.229%), and the component corresponding to λ_2 which achieved an explanatory variance (33.018%). This means that both components were able to explain together a percentage of (91.247%), which was able to represent the data well, and able to indicate the important and influential variables, as the KG indicator showed that the value of $\lambda_1 > 1$and the value of λ_2 was very close to (1) as shown in the table above. Besides, figure below reflects the distribution of the principal components according to the eigenvalues according to the Scree Plot procedure:

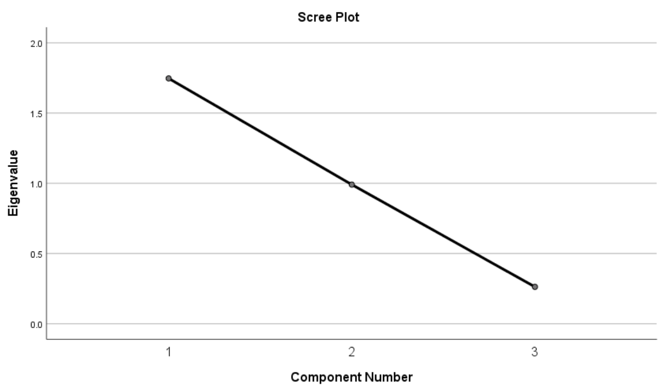

Figure 2: Number of Principal Components for Education Sector

Source: Set by the Researchers as per results of SPSS V.26

Figure (1) illustrates that the number of the principal and influential components are the ones whose value is greater than one. There was one component that corresponds to $\lambda_1 = \mathbf{1.747}$, and this is consistent with the KG test.

Table (7): Coefficients of Principal Components within Education Sector

Component Coefficient	**Average of enrollment in primary education**	**Average of acquaintance with reading and writing**	**Average of enrollment in secondary education**
PC1	**0.855**	**0.386**	**0.932**
Extraction Method: Principal Component Analysis. **a. 1 components extracted.**			

Source: As per results of SPSS V.26 Program

Having a look at results of table (7), it is clear through the values of the variables coefficients that there are important variables in influence, and others that are not important. The results of PC1 showed that the literacy average coefficient is very weak, there it is possible to delete this variable.

- Identification of Principal Components for Health Sector

Table (8): Eigenvalues and its Variances for Health Sector

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 2818 - 2831
https://publishoa.com
ISSN: 1309-3452

Total Variance Explained						
	Initial Eigenvalues			**Extraction Sums of Squared Loadings**		
Component	**Eg. Value**	**% of Variance**	**Cumulative %**	**Total**	**% of Variance**	**Cumulative %**
1	**2.483**	**62.083**	**62.083**	**2.483**	**62.083**	**62.083**
2	**0.979**	**24.470**	**86.553**			
3	**0.332**	**8.296**	**94.849**			
4	**0.206**	**5.151**	**100.000**			
Extraction Method: Principal Component Analysis.						

Source: Set by Researcher as per results of SPSS V.26

As far as table (8) is concerned, the number of the realized principal components is (4), and the best component is the one corresponding to λ_1, which gained an explanatory variance (62.08%), and the component corresponding to λ_2, gaining an explanatory variance (24.47%). This made it clear that both components were able to interpret together a percentage of (86.55%), which in return means that both components were able to represent the data well, and able to indicate the important and influential variables. Also, as per the KG test, the value of $\lambda_1 > 1$ and $\lambda_2 < 1$, therefore it is possible to exclude the PC2 component. Figure below shows the distribution of the principal components according to eigenvalues as per the Scree Plot procedure:

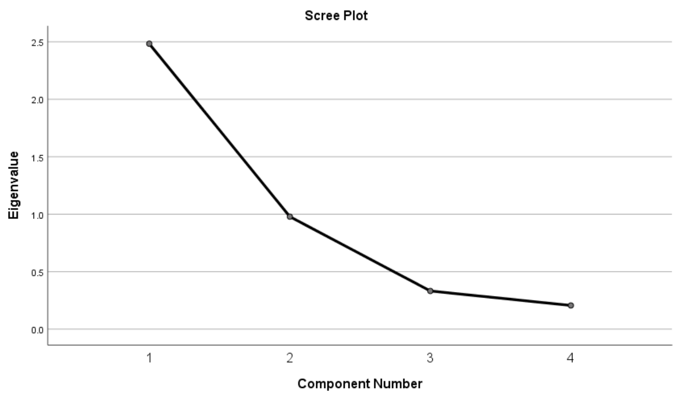

Figure 3: Number of Principal Components for Health Sector along with Identification of Number of Influential Components

Source: Set by Researchers as per results of SPSS V.26

Figure (3) portrays that the number of the principal and influential components are the ones whose value is greater than one. There was one component that corresponds to $\lambda_1 = \mathbf{2.483}$, and this is consistent with the KG test.

Table (9): PC1 Coefficients for Health Sector

Component Coefficient	**Average of child death under the age of 5**	**Life expectancy**	**Percentage of people who get health services**	**Number of beds in hospitals**
	-0.915	**0.922**	**0.709**	**0.543**
Extraction Method: Principal Component Analysis. a. 1 components extracted.				

As it can be elicited from table (9), all the coefficients of the variables are $a_{ij} > 60\%$ except for the variable number of beds. The coefficient of this variable was of average importance (0.543), and this is logical because the number of families is

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 2818 - 2831
https://publishoa.com
ISSN: 1309-3452

related to the number of buildings. Hence, any increase in expenditure on this sector without development or increase in the number of buildings will be of no effect on this variable. Therefore, it is possible to remove the effect of this variable.

2.4 Application of DEA – PCA Model for Efficiency Measurement

In this part, the oriented extrinsic approach to the variable volume return model will be conducted, assuming that the same amount of inputs is maintained and outputs enhanced for the public financial management of the Iraq budget, where the DEA model was estimated in the normal case, and then employing the results of the analysis of the principal components within the cases of reducing the number of variables for the sector of Education (DEA-PC1-Ed) and the sector of health (DEA-PC1-H). Then the common state of the two sectors (DEA-PC1-EDH), while retaining all the variables of the infrastructure sector due to the importance of the influence of their coefficients as identified in table (5). Table below disclosed the results of the analysis:

Table (10): Comparison of Models toward Budget Outputs

Year	DEA	DEA-PC1-ED	DEA-PC1-H	DEA-PC1-EDH
2005	1.0000	1.0000	1.0000	1.0000
2006	0.9863	0.9907	0.9776	0.9776
2007	0.9822	0.9913	0.9615	0.9116
2008	0.9815	0.9283	0.8373	0.5655
2009	1.0000	0.9563	1.0000	0.7098
2010	1.0000	0.9232	1.0000	0.5503
2011	1.0000	0.9356	1.0000	0.6514
2012	1.0000	0.9428	1.0000	0.5266
2013	1.0000	0.9525	1.0000	0.2997
2014	1.0000	0.9621	0.7492	0.4421
2015	1.0000	0.9717	1.0000	0.6317
2016	1.0000	0.9811	1.0000	0.7968
2017	1.0000	1.0000	1.0000	1.0000
2018	1.0000	1.0000	1.0000	1.0000
2019	1.0000	1.0000	1.0000	1.0000

Source: Set by Researchers as per Results of XL-DEA V2

Table (10) shed light on results and made it clear that the DEA models that were hybridized with the analysis of the principal components were able to grant a more credible image, especially the (DEA-PC1-EDH) model than the normal case DEA.

Through examination of models, it was evident that the years (2005, 2019, 2018 & 2019) are the efficient years in financial management within the specific criteria for the sectors. Moreover, the results also pinpointed that the best years for which it is possible to adopt a financial policy in addressing the financial management defect are (2005, 2018 & 2019).

2.5 Conclusions

Through findings obtained, the study highlighted that the analysis of principal components has played a significant role in in responding to the problem of the number of variables that exceed the number of targeted cases in samples of data envelopment analysis (DEA). The impact of the average coefficient of being capable of reading and writing has been dismissed from the education sector, and impact of number of beds' coefficient from the health sector has been dismissed for the above mentioned reasons. Also, it is concluded that the samples of a data envelopment analysis are considered sensitive toward any change as to addition or deletion of any variable. The study has been able to pinpoint years presenting Peer Units in treatment of a defect. The study has featured and come with the fact that the best years, as far as budget administration as per adopted standards is concerned, are (2005, 2017, 2018 & 2019), whereas peer years that can be made use of in enhancing other budgets' performance are (2005, 2018 & 2019).

References

1. Al-taee, O. S. (2021). *Public Financial Managment and Buget Reform Scenarios in Iraq.* University of Basrah, Ph.D in financial and banking sciences. university of Basrah; college of adminstration and economics; pp(1-219).

2. Anderson, T. (2003). *An Introduction to Multivariate Statistical Analysis* (3 ed.). (Anderson, Ed.) John Wiley & Sons, John Wiley & Suns, Inc. Huboken, New Jersey.

3. Azadeh, A., Ghaderi, S. F., & Ebrahimipour, V. (2007). An integrated PCA DEA framework for assessment and ranking of manufacturing systems based on equipment performance. *Engineering Computations: International Journal for Computer-Aided Engineering and Software, 24*(4), 347-372.

4. Azadeh, A., Nasirian, B., Salehi, V., & H. Kouzehchi, H. (2016). Integration of PCA and DEA for identifying and improving the impact of Six Sigma implementation on job characteristics in an automotive industry. *Quality Engineering, Taylor & Francis, 29*(2), 273-290.

5. Banker, R. D., Charnes, A., & Cooper, W. W. (1984, Sep 1). Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis. *management Science, 30*(9), 1078-1092.

6. Davoudabadi, R., Mousavi, S., & Sharifi, E. (2020, 2). A new integrated weighting and ranking model based on entropy, DEA and PCA considering two aggregation approaches for resilient supplier selection problem. *Journal of Computational Science, 40,* 1-32.

7. Dunteman, G. H. (1989). *Principal Components Analysis (Quantitative Applications in the Social Sciences).* (S. Q. Series, Ed.) Sara Miller,McCune, Sage Publications, Inc.

8. Dziuban, C. D., & Shirkiy, E. C. (1974). When is a correlation matrix appropriate for factor analysis? Some decision rules. *Psychological Bulletin, 81*(6), 358-361.

9. Hochreiter, S. (2013). *Basic Methods of Data Analysis.* Institute of Bioinformatics, Johannes Kepler University Linz, statistics. Austria: Johannes Kepler University Linz.

10. Javadi, S., Rameez, M., Dahl, M., & Pettersson, M. I. (2018). Vehicle Classification Based on Multiple Fuzzy C-Means Clustering Using Dimensions and Speed Features. (ScienceDirect, Ed.) *Procedia Computer Science, 126*, 1344–1350.

11. Johnes, J. (2006, june). Data envelopment analysis and its application to the measurement of efficiency in higher education. *Economics of Education Review, 25*(3), 273-288.

12. Johnson, R. A., & Wichern, D. W. (2002). *Applied Multivariate Statistical Analysis* (5 ed.). Prentice Hall, upper saddle river, New Jersey.

13. Karami, S., Yaghin, G. R., & Mousazadegan, F. (2020, 5). Supplier selection and evaluation in the garment supply chain: an integrated DEA–PCA–VIKOR approach. *The Journal of The Textile Institute, 112*(4), 578-595.

14. Layeb, S. B., Omrane, N. A., Siala, J. C., & Chaabani, D. (2020). Toward a PCA-DEA based Decision Support System: A case study of a third-party logistics provider from Tunisia. *2020 4th International Conference on Advanced Systems and Emergent Technologies (IC_ASET)* (pp. 294-299). Hammamet, Tunisia: IEEE.

15. Lee, G. M., & Gao, X. (2021). A Hybrid Approach Combining Fuzzy c-Means-Based Genetic Algorithm and Machine Learning for Predicting Job Cycle Times for Semiconductor Manufacturing. (F. Bianconi, Ed.) *Applied Sciences*(7428), 1-21.

16. Li, Y., Bao, T., Shu, X., Chen, Z., Gao, Z., & Zhang, K. (2020, September). A Hybrid Model Integrating Principal Component Analysis, Fuzzy C-Means, and Gaussian Process Regression for Dam Deformation Prediction. *Arabian Journal for Science and Engineering, 46*, 4293–4306.

17. Morrison, D. F. (2005). *Multivariate Statistical Methods* (4 ed.). United States of America: Duxbury Press.

18. Mustaqeem, M., & Saqib2, M. (2021, 4 16). Principal component based support vector machine (PC-SVM): a hybrid technique for software defect detection. (Springer, Ed.) *Cluster Computing, 24*, 2581–2595.

19. Paradi, J. C., Sherman, H. D., & Tam, K. (2018, july 9). *Data Envelopment Analysis in the Financial Services Industry* (1 ed.). Springer International Publishing AG.

20. Peres-Neto, P. R., Jackson, D. A., & Somers, K. M. (2005, June). How many principal components? stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis, 49*(4), 974-997.

21. Pin Fu, H., & Ruey Ou, J. (2013, February). Combining PCA With DEA to Improve the Evaluation of Project Performance Data: A Taiwanese Bureau of Energy case study. *Project Management Journal, 44*(1), 94-106.

22. Põldaru, R., & Roots, ü. (2014, 3). A PCA-DEA approach to measure the quality of life in Estonian counties. *Socio-Economic Planning Sciences, 48*(1), 65-73.

23. Rahman, A. S., & Rahman, A. (2020, march 12). Application of Principal Component Analysis and Cluster Analysis in Regional Flood Frequency Analysis: A Case Study in New South Wales, Australia. *Water, 12*(3), 1-26.

24. Shrestha, N. (2021). Factor Analysis as a Tool for Survey Analysis. *American Journal of Applied Mathematics and Statistics, 9*(1), 4-11.

25. Vidal, R., Ma, Y., & Sastry, S. S. (2016). *Generalized Principal Component Analysis* (1 ed., Vol. 40). Verlag New York: Springer-Verlag New York.

26. Wu, T.-H., Chung, Y.-F., & Huang, S.-W. (2021, 6). Evaluating Global Energy Security Performances Using an Integrated PCA/DEA-AR Technique. *Sustainable Energy Technologies and Assessments, 45*, 1-17.

27. Zhu, J., & Cook, W. D. (2007). *Modeling Data Irregularities and Structural Complexities in Data Envelopment Analysis* (1 ed.). New York, USA: Springer Science, Business Media, LLC.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 2818 - 2831
https://publishoa.com
ISSN: 1309-3452

28. Zhu, Q., Aparicio, J., Li, F., Wu, J., & Kou, G. (2022, February 1). Determining closest targets on the extended facet production possibility set in data envelopment analysis: Modeling and computational aspects. *European Journal of Operational Research, 296*(3), 927-939.