JBAAR
Journal of Basic And Applied Research
ISSN 2413-7014

Stochastic Programming for Selection Variables in Cluster Analysis

Safia Mahmoud Ezzat[1], Ramadan Hamid Mohamed[2], Elham Abdul-Razik Ismail[3], Mahmoud Mostafa Rashwan[4]

[1]Faculty of Commerce, Al-Azhar University (Girls' Branch), Cairo, Egypt
[2]Faculty of Economics and Political Science, Cairo University, Cairo, Egypt
[3]Faculty of Commerce, Al-Azhar University (Girls' Branch), Cairo, Egypt
[4]Faculty of Economics and Political Science, Cairo University, Cairo, Egypt
*corresponding author: safiahamed@azhar.edu.eg

Received: 30-4-2016
Revised: 22-5-2016
Published: 9-6-2016

Keywords:
Cluster analysis;
Variable selection;
Goal programming;
stochastic programming

Abstract: Cluster analysis is one of the most important techniques in the exploratory data analysis; it is goal to discover a natural grouping in a set of observations without knowledge of any class labels. Variable selection has been very important for a lot of research in several areas of application. The study suggested a stochastic programming approach which selects the most important variables in clustering a set of data. The study evaluates the performance of the stochastic programming suggested approach for selection variables in cluster analysis used numerical example. The suggested stochastic programming approach selects the most important variable in cluster analysis simultaneously and the results are satisfied.

Cite this article as: Ezzat, S.M., Mohamed, R.H., Ismail, E.A., Rashwan, M.M. (2016). Stochastic Programming for Selection Variables in Cluster Analysis. J. basic appl. Res 2(3): 373-378
Like us on Facebook – click here Join us on academia – click here Be co-author with JBAAR on Google Scholar – click here

1- INTRODUCTION

Data clustering is a common technique for statistical data analysis; it is defined as lass of statistical techniques for classifying a set of observations into completely different groups Webb [2002], Hair [2009]. Cluster analysis seeks to minimize group variance and maximize between group variance.

Stochastic programming is a framework for modeling optimization problems that involve uncertainty. Whereas deterministic optimization problems are formulated with known parameters, real world problems almost invariably include some unknown parameters Kall [1994]. The idea of stochastic programming is to convert the probabilistic problem into an equivalent deterministic situation.

Recently variable selection becomes more important for a lot of research in several areas of application, since datasets with tens or hundreds of variables are available and may be unequally useful; some may be just noise, thus not contributing to the process.

Variable selection plays an important role in classification. Before beginning designing a classification method, when many variables are involved, only those variables that are really required should be selected; that is, the first step is to eliminate the less significant variables from the analysis Brusco [2005]. There can be many reasons for selecting only a subset of the variables instead of the whole set of candidate variables: (1) It is cheaper to measure only a reduced set of variables, (2) Prediction accuracy may be improved through exclusion of redundant and irrelevant variables, (3) The predictor to be built is usually simpler and potentially faster when fewer input variables are used and (4) Knowing which variables are relevant can give insight into the nature of the prediction problem and allows a better understanding of the final classification model. Research in variable selection started in the early 1960s. Over the past four decades, extensive research into feature selection has been conducted. Much of the work is related to medicine and biology. The selection of the best subset of variables for building the predictor is not a trivial question, because the number of subsets to be considered grows exponentially with the number of candidate variables.

There have been many trials for stochastic programming and variable selection in cluster analysis, Jeeva el at. [2004] detailed study of recruitment based on cluster analysis technique with an application of Linear Stochastic Programming Problem is considered. The problem under consideration is a Linear Stochastic programming problem (LSPP), where the parameters follow certain empirical distributions. In our case it is assumed to follow Weibull distribution. The model developed can also be modified with more probabilistic constraints.

Khan el at. [2012] studied the recruitment of personnel to various jobs when the job completion times follow identical Gamma distributions. The best groups of persons based on their efficiency in completion of jobs are determined by using cluster analysis technique. The Stochastic formulation of the problem has been brought down to a deterministic NLPP through Chance constrained programming technique. The model developed has also been extended to the case of non-identically distributed time variables. The illustrative examples are given for both, identical and non-identical distributions. It is seen that the solutions with minimum completion times are easily obtained.

Boutsidis and Ismail [2013] presented the deterministic variable selection algorithm for K-means clustering with relative error guarantees. Their result improves upon this in two ways. First, their algorithms are deterministic; second, by using their deterministic algorithms in combination with this randomized algorithm. They can select features and obtain a competitive theoretical guarantee.

Beraldi and Bruni [2014] addressed the scenario tree reduction problem and proposed a new method relying on the clustering analysis. In particular, two variants of the basic scheme were implemented and tested on a multistage stochastic programming formulation of a portfolio optimization problem. Extensive computational experiments were carried out to evaluate the performance of the proposed approach, both in terms of computational efficiency and solution efficacy. Different quality measures have been introduced and widely tested. The results show that the clustering approach exhibits good performance and viewed as an alternative method to the other existing reduction approaches.

Benati and Garcia [2014] focused on binary data where proposed a combinatorial model for clustering that selects simultaneously the best set of variables, the best set of median and the optimal data partition when the criterion used is the minimization of the total distance inside the cluster between the median of the cluster and the units that belong to the cluster. Instead of developing new solution tools for this nonlinear model, this approach study two different mixed integer linearization and to determine which one is the most efficient. The first is a direct linearization formulation of the initial quadratic model and the second is based on the so-called radius formulation of the p median problem.

This study suggested stochastic programming approach which select the most important variables in cluster analysis. The idea of the suggested approach depends on clustering data by minimizing the distance between observations within groups. Indicator variables are used to select the most important variables in the cluster analysis.

The organization of the study is as follows: In Section 2 the study described cluster analysis by nonlinear goal programming. In Section 3 the suggested stochastic programming approach which used in this study. The study applied data analysis to evaluate the performance of the suggested approach under consideration in Section 4. Finally, concluding remarks are provided in Section 5.

2- Goal programming approach for selection variables in cluster analysis

Ezzat el at. [2016] presented a nonlinear goal mathematical programming approach to select the most important variable in cluster analysis. The idea of the approach is to minimize the total distance between observations within groups.

Cluster analysis a nonlinear goal mathematical programming approach described as follows:

Let i = 1,2,…,n be the set of observations that are to be clustered into *m* clusters (groups).

For each observation $i \in N$, we have a vector of observations $yi = \{yi1\ ,yi2\ ,\ldots\ldots,\ yip\} \in R^p$, where *p* is the number of variables.

If the data is standardized using the formula $Z_k = \frac{Y_k - \bar{Y}_k}{S_k}$, Then we have the corresponding vector of observations $zi = \{zi1,\ zi2\ ,\ldots\ldots,\ zip\} \in R^p$.

Since we aim to construct m clusters, we start by defining *n* clusters fictitiously, *n-m* of which will be empty.

Therefore we define *nxn* (0-1) variables x_{ij} such that

$$x_{ij} = \begin{cases} 1 & \text{if the } i^{th} \text{ element belongs to the } j^{th} \text{ cluster} \\ 0 & \text{otherwise} \end{cases}$$

Where cluster j is non empty if x_{jj} =1 *j*=1,…,n.

These variables need to satisfy the following conditions Subhash (1996):

1- In order to insure that each element belongs only to one non empty cluster, then the following constraint is needed:

$$\sum_{j=1}^{n} x_{ij} = 1 \qquad i = 1,\ldots,n. \qquad (1)$$

2- In order to insure that jth cluster is non empty only if x_{jj} =1, then this can be represented as follows:

$$x_{jj} \geq x_{ij} \qquad i = 1,\ldots,n. \quad j = 1,\ldots,n. \qquad (2)$$

3- In order to insure that the number of non-empty clusters is exactly m, then this can be written as:

$$\sum_{j=1}^{n} x_{jj} = m \qquad (3)$$

For example if x_{77}=1, then cluster 7 is non empty and if it includes element number 1 and 3 in addition to element number 7, then we have (1,3,7) as cluster 7 and $x_{17}= x_{37}=x_{77}=1$, while $x_{71}=x_{73}=x_{13}=x_{31}=0$ and also $x_{11}=x_{33}=0$.

Note that summing (2) with respect to *i* results in the following set of constraints

$$nx_{jj} \geq \sum_{i=1}^{n} x_{ij} \qquad j = 1,\ldots,n \qquad (4)$$

This can be written as

$$nx_{jj} - \sum_{i=1}^{n} x_{ij} \geq 0 \qquad j= 1,\dots,n \qquad (5)$$

Thus it reduces the number of constraints from n^2 to n as suggested in Arthanari and Dodge (1993).

To achieve the aim of the study we define a set of variables as follows:

We define p variables V_s such that:

$$v_s = \begin{cases} 1 & \text{if the } s^{\text{th}} \text{ variable is important} \\ 0 & \text{otherwise} \end{cases}$$

For s= 1,……,p

These variables need to satisfy the following condition:

In order to insure that the selected number of important variables is exactly r, then:

$$\sum_{s=1}^{p} V_s = r \qquad (6)$$

To obtain the most important variables, the suggested version to achieve this aim is the minimization of the total sum of square deviations within groups by minimizing the weighted total sum of squares of distance between all observations within each cluster. The suggested weights are the indicator variables V_s.

This objective may be written as

$$Min \sum_{s=1}^{p} (\sum_{i=1}^{n} \sum_{j=1}^{n} (z_{is} - z_{js})^2 x_{ij}) V_s \qquad (7)$$

where z_{is} is the standardized i^{th} observations of the s^{th} variable.

The corresponding fourth goal, in which d^+ and d^- represent the non-negative deviational variables and for which d^+ needs to be minimized, can be written as:

$$\sum_{s=1}^{p} (\sum_{j=1}^{n} \sum_{i=1}^{n} (z_{is} - z_{js})^2 x_{ij}) V_s + d^- - d^+ = 0 \qquad (8)$$

Since the model aims to select the important variables in cluster analysis (8) with respect to the structural constraints (1, 3, 5 and 6), the above discussion the objective function F to take the formula given in x_{ij}, V_s, d^+ (9) and the goal nonlinear programming model takes the form:

Find the values of i,j=1,2,…,n and s=1,2,…,p . Which Minimize:

$$f = d^+ \qquad (9)$$

Subject to

$$\sum_{s=1}^{p} V_s = r \qquad (10)$$

$$\sum_{s=1}^{p} (\sum_{j=1}^{n} \sum_{i=1}^{n} (z_{is} - z_{js})^2 x_{ij}) V_s + d^- - d^+ = 0 \qquad (11)$$

$$\sum_{j=1}^{n} x_{ij} = 1 \qquad i = 1,2,...,n \qquad (12)$$

$$\sum_{j=1}^{n} x_{jj} = m \qquad (13)$$

$$nx_{jj} - \sum_{i=1}^{n} x_{ij} \geq 0 \qquad j = 1,2,...,n \qquad (14)$$

each x_{ij}, V_s is either 0 or 1

The previously approach objectives in the deterministic model (8), is reformulated but in real life applications some of the criterion variables that are used in clustering a set of data could be random so that the suggested following approach is an extension of work presented by Ezzat el at. [2016].

3- Stochastic programming approach for selection variables in cluster analysis

The suggested stochastic programming approach used to select the most important variable in cluster analysis. The following section describe the suggested approach:

The previously goal programming objective in the deterministic model (11) are reformulated. In what follows the normal distribution is assumed for the criterion variables. A commonly used normality test is applied. On the other hand the chance-constraint method is used as a technique to convert the stochastic model to its equivalent deterministic one.

In reality, the difference between observations ***(i)*** and ***(j)*** may be random, hence we can construct a set of goals based on the idea of minimizing the probability of getting very close objects in the same cluster.

Refering to previously the deterministic model. It is clear the random variable may appear in the objectives defined by the equations (11). Therefore this objectives need to be replaced by the following stochastic constraint:

$$pr\left[\sum_{s=1}^{p} (\sum_{j=1}^{n} \sum_{i=1}^{n} (z_{is} - z_{js})^2 x_{ij}) V_s \leq \text{some very small value} \right] \geq (1 - \alpha) \qquad (15)$$

so corresponding deterministic set of constraints can be written as

$$f_p \left[\sum_{s=1}^{p} (\sum_{j=1}^{n} \sum_{i=1}^{n} (z_{is} - z_{js})^2 x_{ij}) V_s \right] \geq (1 - \alpha) \qquad (16)$$

where F_p stands for the χ^2 cumulative distribution function with (p) degrees of freedom.

Hence

$$\left[\sum_{s=1}^{p} (\sum_{j=1}^{n} \sum_{i=1}^{n} (z_{is} - z_{js})^2 x_{ij}) V_s \right] \geq F_p^{-1}(1 - \alpha) \qquad (17)$$

The corresponding set of goals in which d^+ and d^- represent the non-negative deviational variables and for which d^+ needs to be minimized, can be written as:

$$\left[\left[\sum_{s=1}^{p} (\sum_{j=1}^{n} \sum_{i=1}^{n} (z_{is} - z_{js})^2 x_{ij}) V_s \right] + d^- - d^+ \geq F_p^{-1}(1 - \alpha) \right] \qquad (18)$$

From the above discussion, since the model aims to select the important variables in cluster analysis (18) with respect to the structural constraints (1, 3, 5 and 6), the stochastic programming model suggested for the selection of variables in cluster problem takes the form:

Find the values x_{ij}, V_s, d^+ of *i,j*=1,2,…,*n* and *s*=1,2,…,*p* . Which Minimize:

$$f = d^+ \qquad (19)$$

Subject to

$$\sum_{s=1}^{p} V_s = r \qquad (20)$$

$$\left[\sum_{s=1}^{p}\left(\sum_{j=1}^{n}\sum_{i=1}^{n}(z_{is}-z_{js})^2 x_{ij}\right)V_s\right] + d^- - d^+ \geq F_p^{-1}(1-\alpha) \qquad (21)$$

$$\sum_{j=1}^{n} x_{ij} = 1 \qquad i = 1,2,...,n \qquad (22)$$

$$n x_{jj} - \sum_{i=1}^{n} x_{ij} \geq 0 \qquad j = 1,2,...,n \qquad (23)$$

$$\sum_{j=1}^{n} x_{jj} = m \qquad (24)$$

each x_{ij}, V_s is either 0 or 1

4- Data analysis and results

This section discusses comparison between suggested stochastic programming model and goal nonlinear mathematical programming model by the three different real data to evaluate the performance for the suggested approach. Fisher Iris data set, Leaf data set and Education data set which used to evaluate the performance of the stochastic programming suggested approach for selection variables in cluster analysis. The steps which applied to evaluate the performance of the stochastic programming suggested approach for selection variables in cluster analysis as follows:

1- The real data which used in this study has different number of variables which category between small and fairly large.

2- The real data which used sometimes the actual cluster is known and maybe not.

3- The Easy fit package used to determine the distribution of the real data under consideration.

4- The real data under consideration are follow normal distribution with different parameters so the data is standardized to follow standard normal distribution.

5- The results are based on the most three commonly validity measures to evaluate the suggested cluster model, as follows:

- The correct classification percent.
- The adjusted rand index Yeung and Ruzzo [2001] is external criterion, which evaluates the results of a clustering method using a pre-specified structure imposed on a data set. The value of this index is proved to lie between zero and one. The largest value of this index mean that the model performance is good.
- Davies-Bouldin Index Webb [2002] is internal criterion, which evaluates the clustering results in terms of quantities obtained from the data set itself. It is being more close to zero indicates a better clustering.

The validity measures correct classification and the adjusted rand index are used when the actual clustering is known. While Davies-Bouldin Index used when the actual clustering is not known.

The solution steps

The stochastic programming model presented above is a nonlinear stochastic programming model. The following steps are suggested as a technique to solve this problem:

Step 1

Specify the number of observations (*n*) and number of variables (*p*), then enter the real or the standardized values of each variable. The variables are standardized using this formula

$$Z_k = \frac{Y_k - \overline{Y}_k}{S_k}$$

where $\overline{Y}_k$ and S_k are the mean and standard deviation values respectively.

Step 2

The Lingo software used for solving the stochastic programming problems with too many variables, constraints or both to solve the final model.

Step 3

Obtain the values of decision variables x_{ij}, V_s and hence state the most important variables and the clustering results.

Fisher Iris data set* (1936)*:

Fisher data applied to evaluate the performance of the goal programming suggested approach for selection variables in cluster analysis. It considered a set of 150 objects to illustrate linear discriminate analysis. The data set describes three species (clusters) of Iris flowers: Setosa, Versicolor and Virginica on four variables on each plant (lengths (l) and widths (w) of sepals(s) and petals (p)). i.e. we have four variables s.l, s.w, p.l and p.w. The data set includes 50 plants in each cluster. In this study, a random sample of 30 observations is chosen. Ten observations are drawn from each of the three clusters. According to the actual clustering: the first cluster contains objects (1,2, 3,4,5,6,7,8,9,10), the second contains(11,12,13,14,15,16,17,18, 19,20) and finally the third cluster contains (21, 22,23,24,25,26,27,28,29,30). The results are summarized in the following table (1).

Table (1) shows the suggested stochastic programming mode and deterministic model give approximately similar results in clustering and selected the same variables. In these data selection variables is not necessary because the number of variables is small but it used to clarify cluster analysis.

Table (1): The clustering results for fisher data set

	Deterministic model	stochastic model
Variables	V_1 , V_3 , V_4	V_1 , V_3 , V_4
Cluster 1	(1,2,3,4,5,6,7,8,9,10)	(1,2,3,4,5,6,7,8,9,10)
Cluster 2	(11,12,13,14,15,16,17,18,19,20)	(11,12,13,14,15,16,17,18,19,20, 27,30)
Cluster 3	(21,22,23,24,25,26,27,28, 29,30)	(21,22,23,24,25,26,28, 29)
% Correct classification	100%	93.33%
Adjusted Rand index	1	0.808

Table (2): The clustering results for leaf data set

	Deterministic model	Stochastic model
Variables	V_1, V_2 , V_3 , V_4,V_5, V_6 , V_7 , V_8, V_{12}	V_1, V_2 , V_3 , V_4,V_5, V_6 , V_7 , V_8, V_{12}
Cluster 1	(2,3,4,5,6,7,8,9,11,12,16,19,22,35, 38)	(2,3,4,5,6,7,8,9,11,12,16,19,22,35, 38)
Cluster 2	(10,13,14,15, 17,18,20,21)	(10,13,14,15, 17,18,20,21)
Cluster 3	(23,25,27,28, 29,30,31,32)	(23,25,27,28, 29,30,31,32)
Cluster 4	(1,24,26,33,34,36,37,39,40)	(1,24,26,33,34,36,37,39,40)
Cluster 5	(41,42,43,44,45, 46,47,48,49,50,51, 52)	(41,42,43,44,45, 46,47,48,49,50,51, 52)
% Correct classification	82.69%	82.69%
Adjusted Rand index	0.566	0.566

Table (3): The clustering results for Edu data set

	Deterministic model	stochastic model
Variables	V_1, V_7 , V_8,V_9,V_{10} , V_{11} , V_{12},V_{13} ,V_{14} ,V_{16}	V_7 , V_8,V_9,V_{10} , V_{11} , V_{12} ,V_{14},V_{16},V_{17} , V_{18}
Cluster 1	(1,9,11,12,13,22, 25,26)	(1,9,11,12,13,22)
Cluster 2	(2,6,7,8,14,15)	(2,6,7,8,14,15,21)
Cluster 3	(3,4,17,18,19)	(3,4,5,10,16,17,18,19,20)
Cluster 4	(5,10,16,20,21)	(25)
Cluster 5	(23)	(23)
Cluster 6	(24,27)	(24,26,27)
D. B. index	0.287	0.235

Leaf data set:

According to Silva et al. (2013) presented database comprises 40 different plant species (clusters) considered a set of 340 observations on 13 variables (Eccentricity, Aspect Ratio, Elongation, Solidity, Stochastic Convexity, Isoperimetric Factor, Maximal Indentation Depth, Lobedness, Average Intensity, Average Contrast, Smoothness, Third moment, Uniformity, Entropy). This data applied to evaluate the performance of the goal programming suggested approach for selection variables in cluster analysis. In this study, a random sample of 52 observations is chosen, 12 observations from the 1st cluster, 10 from the 2nd cluster, 10 from the 3th cluster, 8 the 4th cluster and 12 the 5th cluster. The listed results are summarized in the following table (2).

In table (2) when leaf data set was used with fairly large variables. The results for the suggested model give similar results in clustering and the same selected variables by deterministic model.

Education data set:

Education data applied to evaluate the performance of the goal programming suggested approach for selection variables in cluster analysis. The annual statistical report of Ministry of Education 2007/2008 [12] includes many variables about education. The suggested model used to group the Egyptian governorates in six clusters together with selecting ten variables from the set of 18 variables related to basic education. The values of eighteen variables defined as follows:

1) The failure percentage in the primary certificate ***(v_1)***.

2) The failure percentage in the preparatory certificate ***(v_2)***.

(The failure percentage is the number of failing pupils divided by the total number of pupils who already attended final exams for each certificate).

3) The percentage of teachers having intermediate qualifications in the primary stage ***(v_3)***.

4) The percentage of teachers having intermediate qualifications in the preparatory stage ***(v_4)***.

(Teachers having intermediate qualifications are those who have a 3 or 5 years diploma certificate .i.e. not having a university level qualification certificate).

5) The percentage of non educational teachers in the primary stage ***(v_5)***.

6) The percentage of non educational teachers in the preparatory stage ***(v_6)***.

(Teachers who are not graduates of the faculty of education are called non educational).

7) The average class density in the primary stage ***(v_7)***.

8) The average class density in the preparatory stage ***(v_8)***.

(The average class density is the total number of pupils divided by the number of classes for a given stage).

9) The percentage of non full day schools in the primary stage ***(v_9)***.

10) The percentage of non full day schools in the preparatory stage ***(v_{10})***.

11) The percentage of multi-shift schools in the primary stage ***(v_{11})***.

12) The percentage of multi-shift schools in the preparatory stage ***(v_{12})***.

(The multi-shift schools are those schools having two or three periods per day).

13) The percentage of defective school buildings ***(v_{13})***.

(Defective school buildings are those that have too old buildings or those that are about to collapse in both stages).

14) The illiteracy rate in the governorate ***(v_{14})***. (The illiteracy rate is defined by the number of illiterate persons in the age group (10+) as a ratio of population in the same age group).

15) The females ratio of pupils in the primary stage ***(v_{15})***.

16) The females ratio of pupils in the preparatory stage ***(v_{16})***.

17) The percentage of special education schools in the primary stage ***(v_{17})***.

18) The percentage of special education schools in the preparatory stage ***(v_{18})***.

(Special education schools are those specified for disabled children).

The data set includes 27 objects (areas) which are the 27 governorates. The results are summarized in table (3). For education data set the study used the Davies-Bouldin index as a performance measurement. Table (3), shows the results for the suggested model is satisfied, where the suggested stochastic programming mode and deterministic model give approximately similar results in clustering and variables selected.

5- CONCLUSION

From the previous results when the study used three different types of real data the suggested have the following advantages:

- The suggested stochastic programming approach and nonlinear goal programming model selects the most important variable in cluster analysis simultaneously.
- Stochastic programming suggested when real world problems are used because almost invariably include some unknown parameters. The idea of stochastic programming is to convert the probabilistic problem in real data and uncertainty.
- The suggested stochastic programming approach used when three different types of real data are used and the results are satisfied.

REFERENCES

Webb, A.R., *Statistical Pattern Recognition*, Second Edition, John Wiley & Sons Ltd, 2002.

Yeung, K.Y and Ruzzo W.L., "Details of the Adjusted Rand Index and Clustering algorithms supplement to the paper "An empirical study on Principal Component Analysis for Clustering gene expression data", *Bioinformatics*, pp. 1-6, 2001.

Boutsidis, C. and Magdon-Ismail M., "Deterministic Feature Selection for k-Means Clustering", *IEEE Transactions on Information Theory*, Vol. 59, Issue 9, pp. 6099 – 6110, 2013.

Benati, S. and Garcia, S., "A mixed integer linear model for clustering with variable selection", *Computers & operations research*, Vol. 43, ISSN 0305-0548, pp. 280 – 285, 2014.

Subhash, S., *Applied Multivariate Techniques*, John Wiley & Sons, Inc., New York, 1996.

Arthanari, T.S. and Dodge Y., *Mathematical Programming in Statistics*, A Wiley Interscience Publication, New York, 1993.

Fisher, R.A, "The use of Multiple Measurements in Taxonomic Problems", *Annals of Eugenics*, Vol.7, pp.179-188, 1936.

Silva, P. F. B., Marcal, A. R. S. and da Silva, R. A., "Evaluation of Features for Leaf Discrimination", *Springer Lecture Notes in Computer Science*, Vol.7950,197-204, 2013.

Ministry of Education, statbook, *http://services.moe.gov.eg/books/A_0708/main_book0708.html,5/3/2010*.

Hair, J.F., Black W.C., Babin B.J. and Anderso R.E., *Multivariate Data Analysis*, Seventh Edition, Prentice Hall, 2009.

Kall, P.and Wallace, S.W., *Stochastic Programming*, John Wiley & Sons, New York, 1994.

Brusco, M.J. and Stahl S., *Branch-and-Bound Application in Combinatorial Data Analysis*, Springer Science Business Media, Inc, U.S.A, 2005.

Ezzat, S. M., Hamid, R. M., Ismail, E. A. and Rashwan, M. M., "Variables Selection In Cluster Analysis Using Goal Programming", *Journal of Multidisciplinary Engineering Science and Technology*, ISSN: 2458-9403, Vol. 3 Issue 3, pp. 4396-4403, March 2016.

Beraldi, P. and Bruni, M. E., "A clustering approach for scenario tree reduction: an application to a stochastic programming portfolio optimization problem", *TOP*, Volume 22, Issue 3, pp 934-949, October 2014.

Khan, M. F., Anwar, Z. and Ahmad, Q. S., " Assignment of Personnels when Job Completion time follows Gamma distribution using Stochastic Programming Technique", *International Journal of Scientific & Engineering Research*, Volume 3, Issue 3, 1 ISSN 2229-5518, March 2012.

Jeeva, M., Rajagopal, R., Charles, V. and Yadavalli, V. S. S., "An Application of Stochastic Programming with Weibull Distribution-Cluster Based Optimum Allocation of Recruitment in Manpower Planning", *stochastic analysis and applications*, Vol. 22, No. 3, pp. 801–812, 2004