

Socio-Economic Differences among Districts of the Punjab: A Cluster Analysis Approach Based on Multiple Indicator Cluster Survey

¹Shahla Ramzan, ²M. Inayat Khan, ¹Faisal Maqbool Zahid, ³Sajid Rasul and ³Shamim Rafiq

¹Department of Statistics, Government College University, Faisalabad, Pakistan

²Department of Mathematics and Statistics, University of Agriculture, Faisalabad, Pakistan

³Department of Planning & Development, Bureau of Statistics, Punjab, Lahore, Pakistan

Abstract: In this paper internally homogenous groups of the districts in the Punjab province of Pakistan are explored on the basis of their socio-economic indicators. The methodological approach is based on cluster analysis. The government of Punjab is dedicated to achieve Millennium Development Goals (MDG) for poverty, health, education and water & sanitation. The districts are clustered using the data of Multiple Indicator Cluster Survey (MICS) 2007-08. On the basis of the analysis, districts are classified into five clusters. The socio-economic indicators like poverty, health, education and water & sanitation showed marked differences in these groupings. The government can benefit from this study while planning and preparing annual development plans for future and can allocate resources according to the development indicators.

Key words: Cluster analysis . multiple indicator cluster survey . millennium development goals . hierarchical method

INTRODUCTION

In a developing country like Pakistan, socio-economic progress is very essential in order to achieve a stronger economy. For any chance of success in achieving targets for improvement in indigenous socio-economic outcomes, policy makers need to understand where relative and absolute need is highest. Punjab is the most populous province of Pakistan and its contribution to Gross Domestic Product (GDP) is more than 50 percent. The government of Punjab is committed to attain the Millennium Development Goals (MDGs) regarding education, health, poverty and water & sanitation. Towards this end, the government has started to conduct a periodic survey to assess the gains of the efforts and to point out where the further improvements are required. So, in the present study we focus on the districts the Punjab province. Our aim is to examine the socio-economic differences among these districts and to classify them into relatively homogenous groups.

At present, Punjab is divided into 35 districts. According to the reports of Federal Bureau of Statistics Pakistan about MICS data 2007-08, ninety-two percent of the population has access to improved drinking water sources within dwelling and five percent within the distance of half an hour (hand/motorized pump 71 percent; piped water 20 per cent; others 9 per cent). Only 57 percent of household population disposes of waste water properly including 96 percent in major cities, 88 percent in other urban areas but only 41 percent in rural areas. About 14 percent of households dispose of solid waste properly with over three-quarters population using open fields. The source of drinking water for the population does not vary greatly by area of residence and district. In rural areas, 97 percent use water from an improved source, mainly hand pump and motorized pump. Major cities have 95 percent usage of piped water and motorized pump, while in other urban areas more than half use motorized pumps, 14 percent use hand pumps and 25 percent use piped water. In most districts, over 95 percent of the population has access to improved sources of drinking water.

The methodology used in this work includes multivariate statistical method-cluster analysis, to explore the groups of homogeneous districts. It is a standard approach for analyzing socio-economic disparities among the territories. Similar analyses have done for some other countries (see for example Ozimek [1] for US, Openshaw [2] for UK and Soares *et al.* [3] for Portugal).

Corresponding Author: Shahla Ramzan, Department of Statistics, Government College University, Faisalabad, Pakistan

Baum [5] used various social and demographic factors to differentiate between Australian metropolitan areas using cluster analysis. Since socioeconomic and demographic characteristics are important determinants of population health, Odoi *et al.* [6] used these characteristics to find the clusters for health planning. Multivariate statistical analysis has been applied to socioeconomic problems and particularly to the classification of different types of administrative divisions (municipalities, counties or regions) in the literature [3, 7, 8, 10-12]. These studies are restricted to a smaller area inside Europe, specifically Croatia, the Midi-Pyrénées Region, Tenerife Island, Portugal, the Baltic Sea countries, a Swedish county, Slovenia and the Spanish region of Galicia in their respective cases. There are other contributions outside of Europe e.g., Stimson *et al.* (2001) focused on Australia. [13, 14].

When the number of factors (socio-economic indicators) is large, the factor analysis may be used prior to cluster analysis [4]. In such case, factor analysis is first used to summarize the information contained in a wide range of observed variables. Cluster analysis is then performed on the basis of such formed factors. However, in our case of district data, we have only a small number of suitable variables, so there is no need for factor creation and we also do not want to lose the information.

The rest of the paper is organized as follows: we describe and characterize the data used in this work in Section 2. Section 3 is dedicated to finding clusters of regions presenting similar characteristics of development. Results are presented in Section 4 with some discussion about the findings. The conclusions reached are stated in Section 5.

DATA DESCRIPTION

Multiple Indicator Cluster Survey (MICS) is an international household survey program developed by UNICEF. It has been conducted in more than 100 countries of the world. First MICS was conducted in Baluchistan province of Pakistan during 2003-04 at district level with the technical and financial assistance of UNICEF. In Punjab, first such survey was conducted in 2003-04. The second survey was conducted with expanding the number of indicators and the sample size. The socioeconomic variables considered in this text (Table 1) are selected from Punjab Multiple Indicators Survey 2007-08.

On the basis of the variables described in Table 1, we conducted cluster analysis to identify several groups of the districts in Punjab. For the purpose of this analysis, all considered variables have been standardized. The method employed in this paper does not make any distributional assumptions, so no other transformation of the data has been performed. Ward's hierarchical procedure is first used to define the number of clusters, whereas the K-means non-hierarchical cluster procedure using the cluster centers obtained with the Ward's method as the initial seed points is used to improve the results.

Table 1: Regional indicators considered in the study

Indicator	Description
Literacy rate 10+ years	Number of household members age 10 years or older who are able, with understanding, to both read and write in any language divided by Total household members age 10 years or older surveyed
Infant Mortality Rate	Probability of dying by exact age 1 year
Care provided by Lady Health Worker (LHW)	Number of women aged 15-49 years that were visited by a Lady Health Worker (LHW) in the last month divided by Total number of women surveyed aged 15-49 years
Skilled attendant at delivery	Number of women aged 15-49 years with a birth in the 2 years preceding the survey that were attended during childbirth by skilled health personnel divided by the total number of women surveyed aged 15-49 years with a birth in the 2 years preceding the survey
Use of contraceptives (any method)	Number of women currently married aged 15-49 years that are using (or whose partner is using) a contraceptive method (either modern or traditional) divided by Total number of women aged 15-49 years that are currently married
Reported tuberculosis	Number of household members who reported that they were diagnosed with tuberculosis in the past year divided by Total household members surveyed
Physical access to drinking water (within dwelling)	Number of household members living in households using improved sources of drinking water divided by Total number of household members in households surveyed
Use of sanitary means of excreta disposal	Number of household members using improved sanitation facilities divided by Total number of household members in households surveyed

ANALYSIS

Descriptive statistics: The descriptive statistics shown in Table 2 reflect some huge asymmetries among the districts of Punjab. The most prominent one is the percent of household members who were diagnosed with tuberculosis in the past year (1:7, the ratio between lowest and highest number of tuberculosis). Also, the value of coefficient of variation is highest for this variable. Another variable with great inconsistency is number of women aged 15-49 years with a birth in the two years preceding the survey who were attended by skilled health personnel (medical doctor, Nurse/midwife or Lady Health Visitor) during childbirth divided by the total number of women surveyed aged 15-49 years with a birth in two years preceding the survey (almost 1:6, also indicated by the value of C.V.). The access to the Lady Health Worker (LHW) and the use of contraceptive methods, have also large discrepancies among the districts (with 32.29% and 31.71% C.V.). Finally, it should be noted that skewness is not present in the data in almost all the variables, so normality can be assumed for our data with sufficiently large sample size.

Ward’s hierarchical method: In the first step we have used Ward’s hierarchical method. The graphical presentation of results with dendrogram given in Fig. 1 shows a fairly clear picture. On the vertical axis we can easily notify two big jumps of the values of the between-group sum of squares-namely at two-group and at five-group level. Hierarchical methods have constraint that once a cluster is formed, it cannot be split, whereas a non-hierarchical methods are more flexible, allowing cases to separate from the clusters that they previously integrated.

Consequently, following the procedure suggested by several authors e.g., Punj and Stewart [16], a non-hierarchical k-means clustering procedure has been performed. Empirical evidence suggests that we come very close to global optimum if we take centroids from hierarchical methods as initial seed-points for the K-means method. In our case centroids from the Ward’s method have been used.

K-means method: Ward’s hierarchical method has given us the initial number of groups (clusters) and the group centroids. In the second step we have used K-means method to improve the results of Ward’s method. The main deficiency of the Ward’s method (and also of all other hierarchical method) is that the allocation of units is final with no possibility of reassignment to another (more appropriate) group during the procedure. The results of the Ward’s method will be used as an input for the K-means method and at this stage we don’t comment on them in more details.

RESULTS AND DISCUSSION

To search for groups of districts of Punjab possessing similar socio-economic indicators, different agglomerative hierarchical clustering procedures were carried out. The objective of this first step was to analyze the agglomeration schedules and dendrograms in order to choose the number of clusters. A dendrogram is a two-dimension diagram that illustrates the fusions made at each successive stage of the process. The observations (in this case, the districts) are listed on the horizontal axis and the vertical axis represents the successive steps. The best interpretative cluster solution can be illustrated by the dendrogram shown in Fig. 1 corresponding to Ward’s method and squared Euclidean distances (Everitt [4, 17], Punj and Stewart [16], Millingan [18], Niknam [19].

Table 2: Descriptive statistics of the regional indicators

Item	Min.	Max.	Mean	Median	Mode	S.D.	C.V.	Skewness	Kurtosis	MDG
Literacy rate 10+ years	33.0	81.00	57.74	56.0	45.0	11.32	19.61	0.215	-0.537	88
Infant Mortality Rate	40.0	110.00	76.86	78.0	78.0	17.93	23.33	0.022	-0.176	40
Care provided by Lady Health Worker	18.3	83.10	56.67	60.5	31.0	18.30	32.29	-0.622	-0.672	100
Skilled attendant at delivery	12.0	68.00	41.06	39.0	35.0	14.52	35.36	0.277	-0.696	>90
Use of contraceptives	13.0	50.00	30.03	29.0	23.0	9.52	31.71	0.141	-1.000	55
Reported tuberculosis	0.1	0.70	0.33	0.3	0.4	0.14	41.64	0.380	0.211	.0045
Physical access to drinking water	76.0	100.00	92.40	94.0	99.0	6.54	7.08	-0.731	-0.570	93
Use of sanitary means of excreta disposal	32.0	95.00	66.49	66.0	54.0	14.59	21.95	-0.070	0.061	90

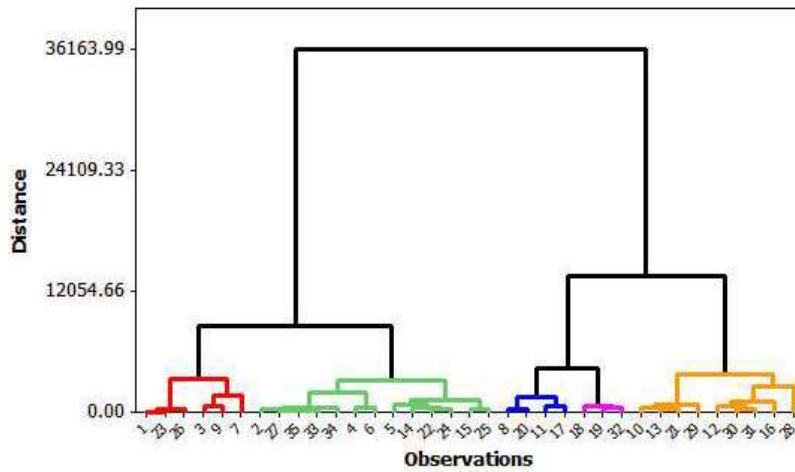


Fig. 1: Dendrogram-ward’s method

Table 3: Clusters of the districts of Punjab (MICS 2007-08)

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Attok	Faisalabad	Bhakkar	Layyah	Bahawalnagar
Chakwal	Gujranwala	Hafizabad	Muzaffargarh	Bahawalpur
Gujrat	Lahore	khanelwal		D.G.khan
Jhelam	Nankana	Mandi Baha-ud-din		jhang
Rawalpindi	Sargodha	Mianwali		Kasur
Sialkot	Sheikhpura	Multan		Khushab
Toba Tek Singh		Narowal		Lodhran
		Vehari		Okara
		Sahiwal		Pakpattan
				R.Y.kahn
				Rajanpur

Table 4: Cluster-wise Summary statistics for education

Literacy rate 10+ years	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Minimum	63.00	55.00	50.00	45.00	33.00
Maximum	81.00	74.00	69.00	53.00	58.00
Mean	71.86	64.50	57.22	49.00	47.09
Median	73.00	63.00	56.00	49.00	48.00
S.D	6.74	7.48	6.24	5.66	6.27
C.V	9.38	11.59	10.91	11.54	13.31

We focus on the solution with five groups of districts, because it offers a more detailed picture about socio-economic differences among districts and the results are logical and very informative with this setting. The list of districts with five groups is presented in Table 3.

In general, the differences between groups (clusters) are characterized by their unique combination of means of socio-economic indicators. For six out of eight variables, the mean values are monotonically increasing or decreasing from the first to fifth group and at the same time most of these variables are indicators of socio-economic development. For that reason the given groups can clearly be ranked with respect to the socio-economic development.

Table 5: Cluster-wise summary statistics for child mortality (Number/Thousand)

Infant mortality rate	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Minimum	40.00	53.00	54.00	72.00	75.00
Maximum	70.00	81.00	92.00	86.00	110.00
Mean	54.57	71.00	78.22	79.00	92.73
Median	52.00	73.00	82.00	79.00	88.00
S.D	10.64	10.20	11.52	9.90	14.47
C.V	19.50	14.36	14.73	12.53	15.61

Table 6: Cluster-wise summary statistics for adult health

Indicator	Statistic	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Care provided by Lady Health Worker (LHW)	Minimum	44.00	18.00	57.00	65.00	31.00
	Maximum	82.00	42.00	83.00	77.00	73.00
	Mean	67.00	28.83	68.67	71.00	52.91
	Median	70.00	26.50	66.00	71.00	59.00
	S.D.	12.75	9.72	8.25	8.49	14.31
	C.V.	19.04	33.73	12.01	11.95	27.05
Reported tuberculosis	Minimum	0.10	0.30	0.10	0.60	0.30
	Maximum	0.30	0.50	0.40	0.70	0.50
	Mean	0.21	0.40	0.24	0.65	0.39
	Median	0.20	0.40	0.20	0.65	0.40
	S.D.	0.07	0.06	0.11	0.07	0.08
	C.V.	32.20	15.81	46.24	10.88	21.26

First cluster contains northern areas of the Punjab and three districts from central Punjab (Gujrat, Sialkot and Toba Tek Singh). Second cluster presents the eastern districts of the Punjab. Two western districts Layyah and Muzaffargarh form a separate cluster. The number of districts in each group is different. To have a detailed look into these clusters, we describe them with respect to each indicator separately in the following subsection.

Cluster-wise summary statistics of the socio-economic indicators: The cluster analysis has classified the districts of the Punjab into five different groups. These groups are heterogeneous with respect to the indicators used for analysis. The summary statistics for the education are presented in Table 4.

The average literacy rate is highest in the first cluster while cluster 5 has the lowest average literacy rate among the children aged ten years and above. The value of literacy rate in cluster 1 ranges from 63 to 81, whereas in cluster 5 this range is from 33 to 58. These values indicate large differences among the literacy rate among districts. The average literacy rate for cluster 1 is 71.86 and for cluster 5 it is 47.09 percent with about 1:2 ratio. The MDG set for the literacy rate is 88 percent and none of the districts in Punjab has achieved this goal.

Table 5 shows cluster wise summary of child mortality. The minimum value of infant mortality rate in cluster 1 is 40 whereas maximum value is 70 having a ratio 1:2, whereas same values for cluster 5 are 75 and 110 respectively. The values of means and medians are almost similar and monotonically increasing from cluster 1 to cluster 5. Although cluster 1 contains district Rawalpindi which achieved the value of MDG (that is 40) but this cluster cannot be regarded as developed regarding infant mortality rate because C.V. has value 19.50 in this cluster which is maximum among all clusters.

The average rate of infant mortality is least in cluster 1. It is because the average literacy rate for this cluster is high and more educated people tend to take more care of their children as compared to less educated people. Cluster 5 contains the districts with very high values of infant mortality rate.

The minimum and maximum values for the care provided by LHW are again in 1:2 ratio in almost all the clusters as shown in Table 6. The average value for the coverage of LHW for cluster 2 is least. It is justified as this cluster contains major cities whereas the LHW program was launched by the government for providing the health facilities to the rural areas. The MDG for LHW coverage is 100 but still no district can achieve this goal. The main reasons for high infant mortality rate in these areas are less education and non-coverage of lady health workers.

Table 7: Cluster-wise summary statistics for reproductive health

Indicator	Statistic	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Skilled attendant at delivery	Minimum	44.00	44.00	31.00	20.00	12.00
	Maximum	68.00	66.00	50.00	26.00	42.00
	Mean	58.00	53.83	38.33	23.00	28.82
	Median	59.00	54.50	36.00	23.00	29.00
	S.D.	9.43	8.42	6.24	4.24	7.69
	C.V.	16.27	15.65	16.29	18.45	26.69
Use of contraceptives (any method)	Minimum	29.00	33.00	20.00	23.00	13.00
	Maximum	50.00	45.00	41.00	27.00	38.00
	Mean	38.43	39.50	28.44	25.00	21.73
	Median	37.00	40.00	26.00	25.00	21.00
	S.D.	6.27	4.76	7.20	2.83	6.69
	C.V.	16.31	12.06	25.30	11.31	30.81

Table 8: Cluster-wise summary statistics for water and sanitation

Indicator	Statistic	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Physical access to drinking water (within dwelling)	Minimum	76.00	84.00	92.00	99.00	82.00
	Maximum	99.00	99.00	99.00	100.00	97.00
	Mean	88.00	93.33	97.33	99.50	89.36
	Median	87.00	96.50	98.00	99.50	91.00
	s.d	7.64	6.62	2.24	0.71	5.33
	c.v	8.68	7.10	2.30	0.71	5.97
Use of sanitary means of excreta disposal	Minimum	73.00	67.00	61.00	42.00	32.00
	Maximum	89.00	95.00	69.00	54.00	65.00
	Mean	79.71	82.00	65.89	48.00	53.45
	Median	80.00	83.00	66.00	48.00	55.00
	s.d	5.85	12.12	2.47	8.49	9.77
	c.v	7.34	14.78	3.75	17.68	18.28

Tuberculosis is one of the major health issues in Pakistan. About 1 in 333 (0.3 per cent) of the surveyed population reported a diagnosis of tuberculosis in the past year (MICS 2007-08). The MDG for the reported tuberculosis cases is 45 per one million population. But the current situation regarding the reported cases of tuberculosis is very alarming. Layyah (0.6 per cent) and Muzaffargarh (0.7 per cent) districts have the largest population diagnosed with tuberculosis. In cluster 1 and 3, there is large variation in the cases reported for tuberculosis.

The summary statistics for the reproductive health presented in Table 7 are also disappointing. Three-quarters of all maternal deaths occur during delivery and the immediate postpartum period. The single most critical intervention for safe motherhood is to ensure the presence of a competent health worker with midwifery skills at birth and transport is available to a referral facility in case of emergency. The average number of skilled attendants at delivery is 58 in cluster 1 and 23 in

cluster 4. The minimum and maximum values in cluster 5 are in a ratio 1:3.5. In Rawalpindi and Gujrat districts, 68 percent of women are assisted during delivery by skilled personnel; medical doctors have the highest percentage of assistance in these districts. In Rajanpur district, only 12 percent women had deliveries assisted by skilled personnel (12 percent). About 38 percent of women aged 15-49 with a birth in the two years preceding the survey delivered in a health facility (institutional delivery) with a higher percentage in urban (57 per cent) than in rural areas (31 per cent). Institutional delivery increases markedly as mother's education and wealth index increases. This indicator is over 50 per cent in Gujranwala, Chakwal, Gujrat, Lahore, Faisalabad and Rawalpindi.

Appropriate family planning is important to the health of women and children by preventing early or late pregnancies, extending birth intervals and limiting the number of children. Differentials exist in the current use of contraception across 35 districts.

Women in Sialkot district have the highest contraceptive prevalence (50 percent) while women in Rajanpur district have the least (13 percent). Like other indicators, MDGs are still not achieved regarding reproductive health.

The results regarding water and sanitation given in Table 8 are in favour of fourth cluster with respect to the physical access to drinking water. The average number of households in first cluster with proper drinking water facility is smallest among all clusters but it has better situation regarding the use of sanitary means of excreta disposal. Although none of the clusters could meet the MDG value for proper sanitary usage, a special attention is needed for cluster 4 and 5 than other clusters to improve the means of sanitary usage facilities.

CONCLUSION AND RECOMMENDATIONS

The building of clusters and cluster-wise summary statistics of the socio-economic indicators reveal that northern and central parts of the Punjab province have better indicators than the other parts. Cluster 1 and 2 can be regarded as the developed group of districts. These clusters contain districts having maximum literacy rate, births by skilled birth attendants and the use of contraceptive methods. The districts with minimum infant mortality rate and prevalence of tuberculosis lie in these clusters. Cluster 5 has worst situation with respect to the availability of facilities related to health and education. It is the most deprived cluster in the Punjab. As an ideal situation all the districts should form a single cluster depicting the efficient use of resources and a fair system of distribution of the facilities but the reality is different. The policy makers should pay attention to these areas. There is a need to improve and purify the system for equalization in the distribution of resources.

REFERENCES

1. Ozimek, J., 1993. Targeting For Success: A Guide to New Techniques for Measurement and Analysis in Database and Direct Response Market. Berkshire: McGraw-Hill.
2. Openshaw, S., 1995. Census Users' Handbook. Geo-information International and Cambridge: John Wiley & Sons.
3. Soares, J.O., M.L. Marquês and C.F. Monteiro, 2003. A Multivariate Methodology to Uncover Regional Disparities: A Contribution to Improve European Union and Governmental Decisions. European Journal of Operational Research, 145: 121-135.
4. Everitt, B.S., 1993. Cluster Analysis. New York: Wiley.
5. Baum, S., 2004. The socio-spatial structure of Australia's metropolitan regions. Australasian Journal of Regional Studies, 10 (2): 157-179.
6. Odoi, A., R. Wray, M. Emo, S. Birch, B. Hutchinson, J. Eyles and T. Abernathy, 2005. Inequalities in neighborhood socioeconomic characteristics: Potential evidence-base for neighborhood health planning. Journal of Health Geographics, 4 (20): 1-15.
7. Cziráky, D., J. Sambt, J. Rován and J. Puljiz, 2005. Regional development assessment: A structural equation approach. European Journal of Operational Research, 174 (1): 427-442.
8. Aragon, Y., D. Haughton, J. Haughton, E. Leconte, E. Malin, A. Ruiz-Gaen and C. Thomas-Agnan, 2003. Explaining the Pattern of Regional Unemployment: The Case of Midi-Pyrenees Region. Papers in Regional Science, 82: 155-174.
9. González, J.I. and S. Morini, 2000. Posicionamiento socioeconómico y empresarial de los municipios de la Isla de Tenerife (Socioeconomic and business affairs situation in Tenerife Island). Working Paper 2000-07. Universidad de La Laguna (In Spanish).
10. Peschel, K., 1998. Perspectives of Regional Development around the Baltic Sea. The Annals of Regional Science, 32: 299-320.
11. Pettersson, O., 2001. Microregional Fragmentation in a Swedish County. Papers in Regional Science, 80: 389-409.
12. Rován, J. and J. Sambt, 2003. Socio-economic Differences among Slovenian Municipalities: A Cluster Analysis Approach. In Development in Applied Statistics. Ferligoj, A. and A. Mrvar (Eds.), pp: 265-278.
13. Rahman, M.M., A. Hussain, M.A. Syed, A. Ansari and M.A.A. Mahmud, 2011. Comparison among clustering in multivariate analysis of rice using morphological traits, physiological traits and simple sequence repeat markers. American-Eurasian J. Agric. & Environ. Sci., 11 (6): 876-882.
14. Nourani, Gh., A.A.A. Jeddi and M.B. Moghadam, 2011. Determining the Structural Parameters and Yarn Type Affecting Tensile Strength and Abrasion of Weft Knitted Fabrics Using Cluster Analysis. Middle-East Journal of Scientific Research, 8 (6): 1008-1017.
15. Stimson, R., S. Baum, P. Mullins and K. O'Connor, 2001. A Typology of Community Opportunity and Vulnerability in Metropolitan Australia. Papers in Regional Science, 80: 45-66.

16. Punj, G. and D. Stewart, 1983. Cluster Analysis in Marketing Research; A Review and Suggestions for Application. *Journal of Marketing Research*, 20: 134-148.
17. Everitt, B.S., S. Landau and M. Leese, 2001. *Cluster Analysis*. London: Edward Arnold.
18. Milligan, G., 1980. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, 45: 325-342.
19. Niknam, T., B. Bahmani Firouzi and M. Nayeripour, 2008a. An efficient hybrid evolutionary algorithm for cluster analysis. *World Appl. Sci. J.*, 4(2): 300-307.