Clustering of Bangladeshi Female Immigrants in West Bengal

Pranati Datta
Population Studies Unit
Indian Statistical Institute
203 B.T Road, Kolkata 700108
West Bengal, India
pranatidatta@hotmail.com
pranati@isical.ac.in

Abstract

Cluster analysis is typically used to identify relatively homogeneous subgroups within a more heterogeneous population. This study is devoted to find out whether useful grouping or cluster exists among Bangladeshi female immigrants in West Bengal, a state in India. Volume of female immigrants from Bangladesh by districts of West Bengal, an Indian state, has been obtained from Census of India , 2001. Agglomerative hierarchical average limkage within group clustering techniques have been used in the present study to identify homogeneous sub groups of Bangladeshi immigrants in districts of West Bengal. Clustering is grouping of objects that are homogeneous in terms of standard similarity judgement. Similarity judgement may be made in terms of a similarity measure or a distance measure. In our present data we have used special case of Minkowski distance measure. After preparing similarity matrix average linkage(within group) clustering has been applied to similarity matrix based on our univariate data. It joins the two clusters for which the average distance between members of the resulting cluster will be smallest to create homogeneity within groups . Proximity matrix, agglomeration schedule and dendrogram have been prepared to identify different clusters. *Dendrograms*, also called *hierarchical tree diagrams*, show the relative size of the proximity coefficients at which districts are combined . The first cluster is formed by districts Bankura and Birbhum at .060 distance coefficient. Darjiling, Murshidabad, Hugli, Bardhaman, Maldah, South 24 Parganas, Malda, Medinipur, Darjiling, Howrah, and other districts are merged gradually and different sub-clusters are formed at different distance levels. Districts Nadia and North 24 Parganas are merged at distance coefficients 4.867 and 7.345 respectively. These two districts are mostly dissimilar member and hence they are combined at highest distance level. This dissimilarity is supported by the fact that North 24 parganas receives highest Bangladeshi immigrants followed by districts Nadia.

Clustering of Bangladeshi Female Immigrants in West Bengal

Pranati Datta
Population Studies Unit
Indian Statistical Institute
203 B.T Road, Kolkata 700108
West Bengal, India
pranatidatta@hotmail.com
pranati@isical.ac.in

Full Paper

Accepted for submission in the European Population Conference 2010

Cluster analysis is typically used to identify relatively homogeneous subgroups within a more heterogeneous population. It is classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait, often proximity according to some defined distance measure. The result of cluster analysis (Everitt,. Sabine, & Morven 2001), is a number of heterogeneous groups with homogeneous contents: There are substantial differences between the groups, but the individuals within a single group are similar. In other words, homogeneity is maximized within clusters and heterogeneity is maximized between them.

Objective:

This study is devoted to find out whether useful grouping (Kaufman, & Rousseeuw, 2005). or cluster exists among Bangladeshi female immigrants in West Bengal, a state in India. Agglomerative hierarchical (Rao, 1969) average limkage within group clustering techniques (Everitt, 1980) have been used in the present study to identify homogeneous sub groups of Bangladeshi immigrants in districts of West Bengal.

Sources of Data : Volume of female immigrants by districts of West Bengal, an Indian state has been obtained from Census of India , 2001.

Methodology : Clustering is grouping of objects that are homogeneous in terms of standard similarity judgement. Similarity judgement may be made in terms of a similarity measure or a distance measure. The most sophisticated of the distance functions are those called metrics (Hand, 1981). Minkowski distance is the generalized distance function. The pth root of the sum of the absolute differences to the pth power between the values for the items.

$$D_{ij} = [\text{sum}(x_{ik} - x_{jk})^p]^{(1/p)}$$

In our present data we have used special case of Minkowski metric which may be described as follows

$$D_p (X_j, X_k) = [\sum_{i=1}^{n} | X_{ij} - X_{ik} |^p]^{1/p}$$

where p ≥ 1. By choosing various values of p many different metric distance function can be obtained. The familiar Euclidean distance or D_2 metric which is applied for our study may be obtained by taking p = 2.

$$D_2(X_j , X_K) = [\sum_{i=1}^{n} | X_{ij} - X_{ik} |^2]^{1/2}]$$

In case of our univariate data on Bangladeshi female immigrants, D_2 reduces to

$$D_2(X_j , X_K) = [| X_j - X_k |^2]^{1/2}$$

In our study we symbolise the similarity matrix as s_{ij} and assume similarity matrix to be symmetric ($s_{ij} = s_{ji}$). The complete schedule of similarities for all possible pairwise combination of entities may be arrayed in upper traingular similarity matrix as follows :

$$S = \begin{array}{cccccc}
S_{12} & S_{13} & S_{14} & \ldots & \ldots\ldots\ldots & S_{1n} \\
 & S_{23} & S_{24} & \ldots & \ldots\ldots\ldots & S_{2n} \\
 & & S_{34} & \ldots & \ldots\ldots\ldots & S_{3n} \\
 & & & \ldots & \ldots\ldots\ldots & \\
 & & & \ldots & \ldots\ldots\ldots & \\
 & & & & & S_{n(n-1)}
\end{array}$$

Amalgamation & Linkage Procedures
Average Linkage (within group) Clustering

After preparing similarity matrix stated above average linkage(within group) clustering has been applied to similarity matrix based on our univariate data. It joins the two clusters for which the average distance between members of the resulting cluster will be smallest to create homogeneity within groups . More specifically the average distance between all pairs in the resulting cluster is made to be as small as possibile. This method is therefore appropriate when the research purpose is homogeneity within clusters.

Here the distance between two clusters is defined as the average of distances between all pairs of objects, where each pair is made up of one object from each group.

In the average linkage method, ***D(r,s)*** is computed as

$$\boldsymbol{D(r,s) = T_{rs} / (N_r * N_s)}$$

Where $\boldsymbol{T_{rs}}$ is the sum of all pairwise distances between *cluster r and cluster s*. $\boldsymbol{N_r}$ and $\boldsymbol{N_s}$ are the sizes of the clusters ***r*** and ***s*** respectively**.**

At each stage of hierarchical clustering, the clusters ***r*** and ***s*** , for which ***D(r,s)*** is the minimum, are merged.

Mathematically the linkage function, the distance between clusters X and Y, is described by the following expression :

$$D(X,Y) = \frac{1}{N_X \times N_Y} \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} d(x_i, y_j);$$
$$x_i \in X,\ y_i \in Y,$$

where

- $d(x,y)$is the distance between objects $x \in X$and $y \in Y$;
- $X$and Y are two sets of objects (clusters);
- $N_X$and $N_Y$are the numbers of objects in clusters $X$and Yrespectively.

Average group Linkage

With this method, groups once formed are represented by their mean values for each variable, that is, their mean vector, and inter - group distance is now defined in terms of distance between two such mean vectors.

In the average group linkage method, the two clusters **r** and **s** are merged such that, after merger, the average pairwise distance within the newly formed cluster, is minimum. Suppose we label the new cluster formed by merging clusters **r** and **s**, as **t**. Then **D(r,s)** , the distance between clusters **r** and **s** is computed as

D(r,s) = Average{ d(i,j) : Where observations i and j are in cluster **t**, the cluster r formed by merging clusters **r** and **s** }

At each stage of hierarchical clustering, the clusters **r** and **s** , for which **D(r,s)** is minimum, are merged. In this case, those two clusters are merged such that the newly formed cluster, on average, will have minimum pairwise distances between the points in it.

Results

Using agglomerative hierarchical average linkage(within group) clustering techniques on percentage data of Bangladeshi Female immigrants into West Bengal by districts (univariate data) we obtain proximity matrix, agglomeration schedule and dendrogram. Basic data on volume and percentage of Bangladeshi migrants are presented in table 1. We would like to search how the districts form cluster in terms of standard similarity judgement.

Table1: Volume and Percentage of Female Immigrants from Bangladesh to West Bengal :2001

District Number	District name with short name in parenthesis	Volume of female immigrant from Bangladesh	percent
1	Kooch Bihar (KB)	90702	6.37
2	Jalpaiguri (Jal)	105237	7.39
3	Darjiling (Dar)	26545	1.86
4	Dinaj Pur (Dina P)	125939	8.84
5	Maldah (Mal)	43592	3.06
6	Murshidabad (Mur)	24493	1.72
7	Nadia (Nad)	272503	19.12
8	North 24 parganas (N 24Pgs)	431655	30.29
9	South 24 Parganas (S 24 Pgs)	46827	3.28
10	Kolkata (Kol)	84051	5.9
11	Haora (Hao)	17916	1.26
12	Hugli (Hug)	67167	4.72
13	Medinipur (Med)	8888	0.62
14	Bankura (Ban)	3950	0.28
15	Puruliya (Pur)	644	0.04
16	Barddha (Bar)	69916	4.91
17	Birbhum (Bir)	4847	0.34

The first step in cluster analysis is establishment of the similarity or distance matrix. This matrix is a table in which both the rows and columns are the units of analysis and the cell entries are a measure of similarity or distance for any pair of cases. The district numbers with short names are presented in rows and in columns in Table 2 and this is the first proximity matrix

Table: 2 Female migrants from Bangladesh to Districts of West Bengal:2001

Proximity Matrix

District no	Minkowski (2) Distance																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
	KB	Jal	Dar	Dina P	Mal	Mur	Nad	N 24 pgs	S 24 pgs	Kol	Hao	Hug	Med	Ban	Pur	Bar	Bir
1KB	.00	1.02	4.51	2.47	3.31	4.65	12.75	23.92	3.09	.470	5.11	1.65	5.75	6.09	6.33	1.46	6.03
2Jal		.000	5.53	1.45	4.33	5.67	11.73	22.9	4.11	1.49	6.13	2.67	6.77	7.11	7.35	2.48	7.05
3Dar			.000	6.98	1.20	.140	17.26	28.43	1.42	4.04	.600	2.86	1.24	1.58	1.82	3.05	1.52
4Din P				.000	5.78	7.12	10.28	21.45	5.56	2.94	7.58	4.12	8.22	8.56	8.80	3.93	8.50
5Mal					.000	1.34	16.06	27.23	.220	2.84	1.800	1.66	2.44	2.78	3.02	1.85	2.72
6Mur						.000	17.4	28.57	1.56	4.18	.460	3.00	1.10	1.44	1.68	3.19	1.38
7Nad							.000	11.17	15.84	13.22	17.86	14.4	18.5	18.84	19.08	14.21	18.78
8N24pgs								.000	27.01	24.39	29.03	25.57	29.67	30.01	30.25	25.38	29.95
9S24pgs									.000	2.62	2.02	1.44	2.66	3.00	3.24	1.63	2.94
10kol										.000	4.64	1.18	5.28	5.62	5.86	.990	5.56
11Hao											.000	3.46	.640	.980	1.22	3.65	.920
12Hug												.000	4.10	4.44	4.68	.190	4.38
13Med													.000	.340	.580	4.29	.280
14Ban														.000	.240	4.63	.060
15Pur															.000	4.87	.300
16Bar																.000	4.57
17Bir																	.000

Table 2 shows the first proximity matrix. The underlined data in upper triangular proximity matrix show the level at which districts are merged. The other distance coefficients or level of merging will be observed in next consecutive proximity tables which are not shown here After each merging and forming one sub-cluster the entries of next proximity matrix are updated in order to reflect similarity between cluster. Performing the process of merging of sub-cluster n-1 times one single cluster is obtained . At each merging level distance coefficient is recorded. Consecutive distance matrix reducing the number of cluster are not shown due to economy of space.

Table : 3 Clustering of Bangladeshi Female Immigrants by Districts of West Bengal at different Distance level using Average Linkage (Within Groups)

Agglomeration Schedule

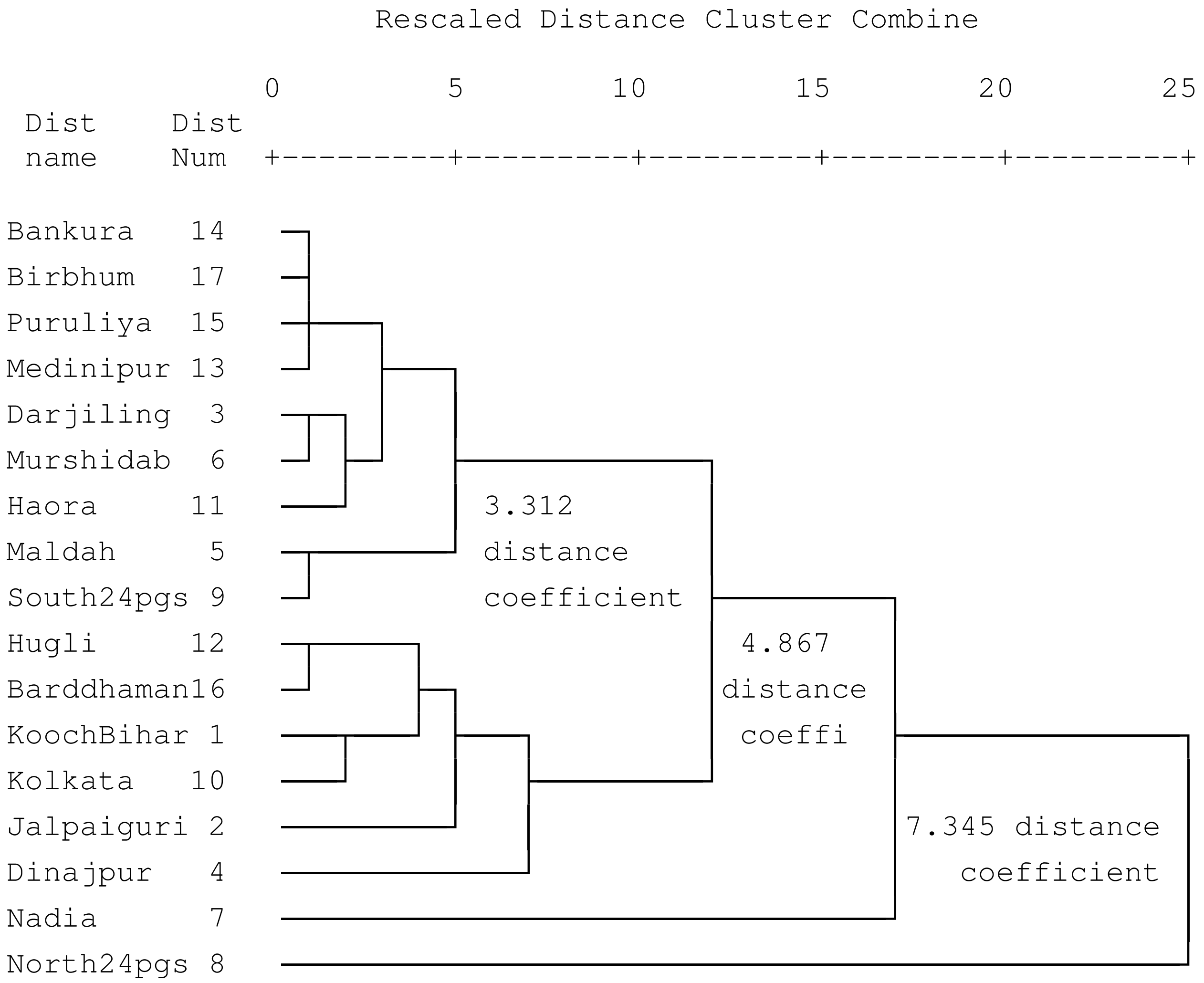
Stage	Cluster Combined with serial no. of districts and short name within bracket		Distance Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	14 (Ban)	17(Bir)	.060	0	0	4
2	3 (Dar)	6 (Mur)	.140	0	0	7
3	12 (Hug)	16(Bar)	.190	0	0	10
4	14 (Ban)	15(Pur)	.200	1	0	6
5	5 (Mal)	9 (S 24pgs)	.220	0	0	12
6	13 (Med)	14(Ban)	.300	0	4	9
7	3(Dar)	11(How)	.400	2	0	9
8	1(KB)	10(Kol)	.470	0	0	10
9	3 (Dar)	13(Med)	.882	7	6	12
10	1(KB)	12(Hug)	.990	8	3	11
11	1(KB)	2(Jal)	1.360	10	0	13
12	3(Dar)	5 Mal)	1.413	9	5	14
13	1(KB)	4(Dina)	1.901	11	0	14
14	1(KB)	3(Dar)	3.312	13	12	15
15	1(KB)	7(Nad)	4.867	14	0	16
16	1(KB)	8(N 24pgs)	7.345	15	0	0

Table 3 shows agglomeration schedule of districts of West Bengal, by cluster combined, distance level, stage cluster, and next stage cluster. The information given in the stage clusters first appears column just indicates when an object is joining an existing cluster or when two existing clusters are being combined.

From Table 3 and Fig 1 it is observed that first cluster is formed by districts Bankura and Birbhum at .060 distance coefficient. Darjiling, Murshidabad, Hugli, Bardhaman, Maldah, South 24 Parganas, Malda, Medinipur, Darjiling, Howrah, Kooch Bihar are merged gradually and different subclusters are formed at different distance levels which are shown in table 3. Districts Nadia and North 24 Parganas are merged at distance coefficients 4.867 and 7.345 respectively. These two districts are mostly dissimilar member in terms of volume of Bangladeshi female migrants and hence they are combined at highest distance level. From

Table 1 it is also revealed that North 24 parganas receives highest Bangladeshi immigrants followed by district Nadia.

Fig:1 Dendrogram of Female Bangladeshi Immigrants using Average Linkage (Within Group)cluster analysis, Census :2001



Dendrograms, also called *hierarchical tree diagrams or plots*, show the relative size of the proximity coefficients at which cases were combined. Trees are usually depicted horizontally, not vertically, with each row representing a case on the Y axis, while the X axis is a rescaled version of the proximity coefficients. Cases with low distance/high similarity are close together. Cases showing low distance are close, with a line linking them a short distance from the left of the dendrogram, indicating that they are agglomerated into a cluster at a low distance coefficient, indicating alikeness. Here Bankura, Birbhum, Purulia, Medinipur, are merged at very low distance levels. Murshidabad, Haora and Darjiling have formed another cluster at small distance level. Districts Nadia and North 24 Parganas are merged at relatively higher distance levels. i.e 4.867 and 7.345 respectively. These two districts are most dissimilar in terms of volume of migrants .

Determinants of Clustering :

From this cluster analysis it is clear that districts Nadia and North 24 Parganas having relatively higher distance levels emerge as dissimilar member of the districts. This fact can be explained by existence of variation in the volume of Bangladeshi migrants(Table1) in West Bengal. West Bengal has 9 border districts with Bangladesh, these are Kooch Bihar, Jalpaiguri, Dinaj Pur, Maldah, Murshidabad, Nadia, Kolkata, 24 Parganas (North and South). Bankura, Birbhum, Purulia, Medinipur are not border districts. These districts experience very very low volume of Bangladeshi migrants. Hence they are merged at very very low distance level. Murshidabad, Darjiling, Kooch Bihar, Dinaj pur, Kolkata being border districts with moderate volume of migrants are agglomerated relatively higher distance coefficients. But among the border districts of West Bengal, North 24 parganas, Nadia stand as mostly dissimilar member. They are agglomerated at 4.867 and 7.345 levels respectively since these two districts attract large volume of Bangladeshi female migrants. In this context one of the important laws of migration explained by Ravenstein(1885, 1886) can be cited. It states that most migrants travel short distance and with increasing distance the numbers of migrants decrease. This law is based upon the assumptions that the higher travel costs and a lack of knowledge of more distant places acts against large volumes of migration. The main reason of large volume of female migrants in border districts might be marriage migration or movement with house hold. There may be distress migration in the informal sector of West Bengal. Census does not provide any information on it. Large scale survey enquiring reason for female labour migraton is recommended.

References

Everitt, Brain., 1980. **Cluster Analysis,** New York, Halsted.

Everitt, Brian S., Sabine Landau, & Morven Leese (2001). **Cluster Analysis***, 4th Edition*. London: Edward Arnold Publishers Ltd..

Hand, D.J., 1981. **Discrimination and Classification,** New York, John Wiley.

Kaufman, Leonard & Rousseeuw Peter J. (2005). **Finding groups in data: An introduction to cluster analysis** NY: Wiley-Interscience

Rao, M.R., 1969 **Cluster Analysis and Mathematical Programming.** , Carnegie-Mellon University.

Ravenstein, E.G., 1885. The Laws of Migration, **Journal of the Royal Statistical Society,** June, Vol. 48, No.2.

Ravenstein, E.G., 1989. The laws of Migration , **Journal of Royal Statistical Society,** June, Vol. 52, No.2.