Reliability of the Model for Clustering of Longitudinal datasets of Infant Mortality Rate in India

Ajay Kumar Bansal* S D. Sharma†

*University College of Medical Sciences, University of Delhi, akchbansal@yahoo.com
†Meerut College, Meerut, Uttar Pradesh, India, sd_sharma10@yahoo.co.in

Reliability of the Model for Clustering of Longitudinal datasets of Infant Mortality Rate in India

Ajay Kumar Bansal and S D. Sharma

Abstract

Because of the natural tendency of human beings and heavenly bodies to form groups, the technique of cluster analysis or segmentation analysis find its importance and applications in many fields of study. A model for clustering of time trends was proposed by authors whose beauty is that 2-way dimensions that is the horizontal flow of the trend and vertical distance of the trend from a common base are considered to obtain the natural clusters. In the present paper, the reliability of this model is studied in two steps namely (i) by repeating the analysis but using different interval distance measures and (ii) by repeating the analysis but using different hierarchical clustering techniques. Dissimilarity coefficients were calculated for the time trends of infant mortality rates in India using this model. In SPSSv17.0, four different clustering methods were applied using generalized power function. Agglomeration schedules were obtained and elbow criterion diagrams were made for each trend. Five stable clusters were suggested by these methods. K-means clustering technique was applied to obtain the actual members of these five clusters.

1. Introduction

Reliability of a test or a model is generally considered as if we get the same results by performing the test or by using the model, again and again. In this paper, the reliability of the model proposed by Bansal and Sharma (2003) is studied, where the authors have suggested a method for clustering the time trends. Cluster analysis is a technique by which a set of observations with similar characteristics is classified into mutually exclusive groups or sets. These groups are called clusters (Anderberg 1973; Copley 1971; Devijver 1982; Fukunaga 1972; Hartigan 1975; Jain 1988 and Zupan 1982). This technique minimizes the within group variations and maximizes the between group variations. Sometimes the cluster analysis is also called segmentation analysis, automatic classification, numerical taxonomy and typological analysis. Because of the natural tendency of human beings and heavenly bodies to form groups, this technique finds application in many fields of study, such as machine learning, data mining, pattern recognition, image analysis, bioinformatics, space sciences, earth sciences, engineering, life sciences, behavioral sciences, medicine, social sciences, etc.

The model proposed by the author was to obtain the clusters of time trends as there have been very few studies to cluster longitudinal datasets. The authors stated in their previous paper (Bansal and Sharma 2003) that: Dunn and Landwehr(1980) obtained the changes in cluster characteristics across two successive time periods. Symon et al.(1983) divided countries into high and low-risk categories on the basis of the ordered rates. The applicability of staged clustering and canonical analysis to classification was studied by Ishii et al (1981). Stanfel(1986) used location theory to cluster the different States of the US for cancer mortality data over the period 1950 to 1967.Ulm(1984) considered a model in which the measurements follow exponential decay curves which are described by an autoregressive stochastic process of the first order. The discriminant function was estimated by the expected values and covariance matrices of the variables. Bhattacharya(1945) gave the measure of divergence between two multinomial populations. Wallenstein(1980) showed whether the data points in a data set tend to cluster or not. He used scan statistics to test for clustering in time. Kafadar and Karon(1993) used a log-linear model to estimate the scale factors and the common trend for the longitudinal data. Bansal and Indrayan (1993) used hierarchical clustering methods to cluster mortality indicators up to the age of one year. Browdy(1982) used Bayes procedures for the classification of multiple trends with dependent residuals. The model they produced is not realistic, however, because for each variable the temporal trends of all subjects in a given universe are represented by a common regression function, and the trends observed in the subjects in the training data are deviations from the common trend. The dependence among the successive residuals of all the variables follows the same pattern.

2. Material and Methods

We have considered the model proposed by Bansal and Sharma(2003) where the coefficient of dissimilarity was obtained for longitudinal datasets measured over a period of time and use it to cluster the infant mortality rate (IMR) trends for 14 major States of India from 1972 to 1998. In the present paper, the idea is not to cluster the IMR trends but to study the reliability of the model proposed. The dataset was taken from the Sample Registration System (SRS) of Registrar General of India(1972-2000). In this model each State was represented as n^{th} degree polynomial by fitting a curve with the help of curvilinear regression method using SPSSv10.0. The total difference in rate of change from time t_1 to t_n (where n=2,3,4,…….,N) for each State was obtained by summing the differences in velocity between two adjacent time points i.e. $\sum_{n=2}^{N} [(f_p'(t_n) - f_p'(t_{n-1})]$ where p=1,2,3,….,P and P are the number of objects or states in this case, to be clustered. The distance of the trend from the base was calculated using the formula $sqrt[(x_p t_1 + x_p t_n + \sum_{z=1}^{Z-2} x_p t_{[1+(\frac{n-1}{z-1})*Z]})^2]$ by dividing the trend objectively in to the optimum number of divisions Z. The Z was postulated as 3 if the (degree of the trend^2/number of time points)≤ 3 and otherwise round(degree of the trend^2/number of time points). By adding these two, Bansal and Sharma(2003) proposed to calculate the dissimilarity coefficient (D) which is given by:

$$D = \sum_{n=2}^{N} [(f_p'(t_n) - f_p'(t_{n-1})] + sqrt[(x_p t_1 + x_p t_n + \sum_{z=1}^{Z-2} x_p t_{[1+(\frac{n-1}{z-1})*Z]})^2]$$

Based on this model, final dissimilarity coefficients were calculated for major 14 States of India, are given in table 1. In this paper we have used this dissimilarity coefficient to establish the reliability of the model proposed by Bansal and Sharma(2003). Reliability refers to the consistency of results which is done in two steps i.e. (i) by repeating the analysis but using different interval distance measures and (ii) by repeating the analysis but using different hierarchical clustering techniques. This is done by taking different distance measures for each of the hierarchical clustering methods available in SPSSv17.0.Generalized power function is

applied and found that Euclidean distance measure, Chebychew interval measure, City block distance, Minkowski-1, Minkowski-2, Minkowski-3 and Minkowski-4 interval measure gives the same results as given by the generalized power function with power 1 and n^{th} root 1. We denoted it by Power(1,1). Square of Euclidean distance, Power(2,1) and Power(4,2) gives the same results, Power(1,2) and Power(2,4) also gives the same results. It is also noted that Centroid linkage method, Median linkage method and Ward's method gives stable results only with square of Euclidean distance measure. Because of this limitation, only four methods namely between group linkage method, within group linkage method, single linkage method (nearest neighbor) and complete linkage method (furthest neighbor) were employed to obtain the agglomeration schedule. The Elbow rule diagrams were also made to decide the number of clusters. Although the diagrams were obtained for all the four methods of clustering but for the brevity of the results we are presenting diagrams only for one method. After obtaining the number of clusters, k-means clustering technique is used to identify the actual members of different clusters.

Table 1: The dissimilarity coefficients calculated using the model

State	Dissimilarity Coefficient (D)
Andhra Pradesh (AP)	244.29
Assam (AS)	308.32
Gujarat (GJ)	287.24
Haryana (HR)	246.85
Himachal Pradesh (HP)	247.79
Karnataka (KT)	236.17
Kerala (KL)	108.18
Madhya Pradesh (MP)	386.59
Maharashtra (MH)	236.63
Orissa (OR)	356.08
Punjab (PJ)	250.87
Rajasthan (RJ)	284.35
Tamil Nadu (TN)	258.86
Uttar Pradesh (UP)	449.51

3. Results

With the help of SPSSv17.0, the agglomeration schedules were obtained by repeating the analysis on dissimilarity coefficients given in table 1 using the different interval distance measures and different methods of clustering. The coefficients calculated at different stages of clustering are given in Table 2.

Cluster analysis presents the problem of how many factors, or dimensions, or clusters to keep. One rule of thumb for this is to choose a place where the cluster structure remains stable for a long distance. Also at the clustering state, where there occurs a sudden change in this coefficient, the clusters are taken as the optimum number of clusters (SPSSv10.0 Base Manual). Alternatively, one can choose a number of clusters so that adding another cluster doesn't give much better modeling of the data. More precisely, if we graph the coefficients against the number of cluster stages, the first clusters will add information (explain a lot of variance), but at some point the marginal gain will drop, giving an angle in the graph, which looks like an elbow. The number of clusters, are chosen at this point, hence the "elbow criterion" (available at http://biocomp.bioen.uiuc.edu/oscar/tools/Hierarchical Clustering.html). In Table 2, where there is not much change after adding a new cluster is taken as the point of optimum number of clusters. The cell for such a point is filled with grey color in the table.

Table 2: The coefficients calculated at different stages of clustering

Clustering method			B-AVERAGE		W-AVEARGE		SINGAL-LINKAGE		COMPLETE-LINKAGE	
Distance measure	Clustering stage	No.of clusters	Coefficients	ratio	Coefficients	ratio	Coefficients	ratio	Coefficients	ratio
Power(1,1)	7	8	15.093	0.670	9.529	0.596	7.990	0.379	22.690	0.947
	8	7	22.525	0.738	15.980	0.579	21.080	0.827	23.970	0.786
	9	6	30.510	**0.644**	27.623	0.905	25.490	0.835	30.510	**0.423**
	10	5	47.380	0.606	30.510	**0.607**	30.510	**0.639**	72.150	0.772
	11	4	78.175	0.570	50.229	0.806	47.760	0.759	93.430	0.467
	12	3	137.256	0.747	62.287	0.714	62.920	0.492	200.140	0.586
	13	2	183.632		87.195		127.990		341.330	
Power(1,2)	7	8	3.819	0.805	2.895	0.776	2.827	0.616	4.763	0.973
	8	7	4.744	0.859	3.729	0.780	4.591	0.909	4.896	0.886
	9	6	5.524	0.810	4.782	**0.866**	5.049	0.914	5.524	0.650
	10	5	6.818	**0.775**	5.524	0.899	5.524	**0.799**	8.494	**0.879**
	11	4	8.799	0.762	6.147	0.821	6.911	0.871	9.666	0.683
	12	3	11.545	0.863	7.483	0.898	7.932	0.701	14.147	0.766
	13	2	13.372		8.335		11.313		18.475	
Power(1,3)	7	8	2.434	0.862	1.995	0.846	1.999	0.724	2.831	0.982
	8	7	2.823	0.903	2.357	0.875	2.762	0.939	2.883	0.923
	9	6	3.125	0.871	2.694	0.873	2.943	0.942	3.125	0.751
	10	5	3.588	**0.843**	3.086	**0.914**	3.125	**0.861**	4.163	**0.917**
	11	4	4.257	0.836	3.376	0.912	3.628	0.912	4.538	0.776
	12	3	5.091	0.906	3.701	0.931	3.977	0.789	5.849	0.837
	13	2	5.618		3.975		5.040		6.989	
Power(1,4)	7	8	1.946	0.893	1.666	0.883	1.681	0.785	2.183	0.986
	8	7	2.178	0.927	1.886	0.916	2.143	0.954	2.213	0.941
	9	6	2.350	0.902	2.059	0.896	2.247	0.956	2.350	0.806
	10	5	2.605	**0.879**	2.299	**0.935**	2.350	**0.894**	2.914	**0.937**
	11	4	2.963	0.875	2.457	0.934	2.629	0.933	3.109	0.827
	12	3	3.385	0.929	2.630	0.948	2.816	0.837	3.761	0.875
	13	2	3.645		2.775		3.364		4.298	
Power(2,1)	7	8	258.653	0.508	126.777	0.370	63.840	0.144	514.836	0.896
	8	7	509.464	0.547	342.426	0.368	444.366	0.684	574.561	0.617
	9	6	930.860	0.386	930.860	0.770	649.740	0.698	930.860	0.179
	10	5	2,413.385	**0.380**	1,208.237	**0.266**	930.860	**0.408**	5,205.623	**0.596**
	11	4	6,344.046	0.304	4,539.650	0.859	2,281.018	0.576	8,729.165	0.218
	12	3	20,896.225	0.552	5,285.764	0.405	3,958.926	0.242	40,056.020	0.344
	13	2	37,832.236		13,040.609		16,381.440		116,506.169	
Power(2,3)	7	8	6.016	0.755	4.260	0.711	3.997	0.524	8.015	0.964
	8	7	7.972	0.816	5.991	0.709	7.631	0.881	8.313	0.851
	9	6	9.764	**0.752**	8.447	**0.865**	8.661	0.887	9.764	**0.563**
	10	5	12.984	0.713	9.764	0.809	9.764	**0.742**	17.331	0.842
	11	4	18.205	0.693	12.076	0.785	13.164	0.832	20.590	0.602
	12	3	26.269	0.823	15.391	0.860	15.819	0.623	34.215	0.701
	13	2	31.922		17.905		25.397		48.841	

Table 2 *continued…*

Clustering method			B-AVERAGE		W-AVEARGE		SINGAL-LINKAGE		COMPLETE-LINKAGE	
Distance measure	Clustering stage	No.of clusters	Coefficients	ratio	Coefficients	ratio	Coefficients	ratio	Coefficients	ratio
Power(3,1)	7	8	4,893.177	0.423	2,004.390	0.260	510.082	0.054	11,681.631	0.848
	8	7	11,569.734	0.407	7,721.202	0.272	9,367.244	0.566	13,772.225	0.485
	9	6	28,400.542	0.217	28,400.542	0.454	16,561.875	0.583	28,400.542	0.076
	10	5	131,047.522	**0.246**	62,605.639	**0.172**	28,400.542	**0.261**	375,585.663	**0.461**
	11	4	5.323E+05	0.155	3.644E+05	0.495	1.089E+05	0.437	8.156E+05	0.102
	12	3	3.445E+06	0.392	7.363E+05	0.302	2.491E+05	0.119	8.017E+06	0.202
	13	2	8.789E+06		2.434E+06		2.097E+06		3.977E+07	
Power(3,2)	7	8	61.591	0.575	33.841	0.463	22.585	0.233	108.082	0.921
	8	7	107.070	0.635	73.018	0.433	96.785	0.752	117.355	0.696
	9	6	168.525	0.503	168.525	0.951	128.693	0.764	168.525	0.275
	10	5	335.266	**0.478**	177.118	**0.363**	168.525	**0.511**	612.850	**0.679**
	11	4	701.091	0.419	487.294	0.638	330.063	0.661	903.087	0.319
	12	3	1,674.463	0.645	764.155	0.744	499.095	0.345	2,831.397	0.449
	13	2	2,596.062		1,027.733		1,447.985		6,306.112	
Power(3,4)	7	8	7.561	0.732	5.191	0.680	4.752	0.483	10.396	0.960
	8	7	10.335	0.796	7.629	0.675	9.838	0.867	10.833	0.834
	9	6	12.982	0.724	11.299	**0.870**	11.344	0.874	12.982	0.524
	10	5	17.932	**0.685**	12.982	0.759	12.982	**0.715**	24.756	**0.824**
	11	4	26.196	0.660	17.098	0.785	18.168	0.813	30.051	0.565
	12	3	39.671	0.803	21.791	0.824	22.340	0.587	53.211	0.670
	13	2	49.380		26.439		38.052		79.411	
Power(4,1)	7	8	99,037.660	0.375	35,710.106	0.203	4,075.558	0.021	265,056.210	0.803
	8	7	263,790.863	0.304	175,883.828	0.203	197,461.497	0.468	330,120.228	0.381
	9	6	866,500.526	0.115	866,500.526	0.245	422,162.198	0.487	866,500.526	0.032
	10	5	7.511E+06	**0.164**	3.534E+06	**0.114**	8.665E+05	0.167	2.710E+07	0.356
	11	4	4.594E+07	0.076	3.091E+07	0.269	5.203E+06	**0.332**	7.620E+07	**0.047**
	12	3	6.032E+08	0.266	1.149E+08	0.218	1.567E+07	0.058	1.604E+09	0.118
	13	2	2.271E+09		5.260E+08		2.684E+08		1.357E+10	
Power(4,3)	7	8	38.413	0.603	22.032	0.503	15.973	0.274	64.236	0.929
	8	7	63.672	0.668	43.820	0.464	58.232	0.776	69.113	0.725
	9	6	95.336	0.547	94.539	**0.992**	75.017	0.787	95.336	**0.317**
	10	5	174.292	**0.517**	95.336	0.423	95.336	**0.550**	300.364	0.708
	11	4	337.103	0.465	225.537	0.879	173.282	0.692	423.950	0.362
	12	3	725.441	0.678	256.514	0.574	250.256	0.388	1,170.699	0.491
	13	2	1,070.070		447.230		645.012		2,385.426	

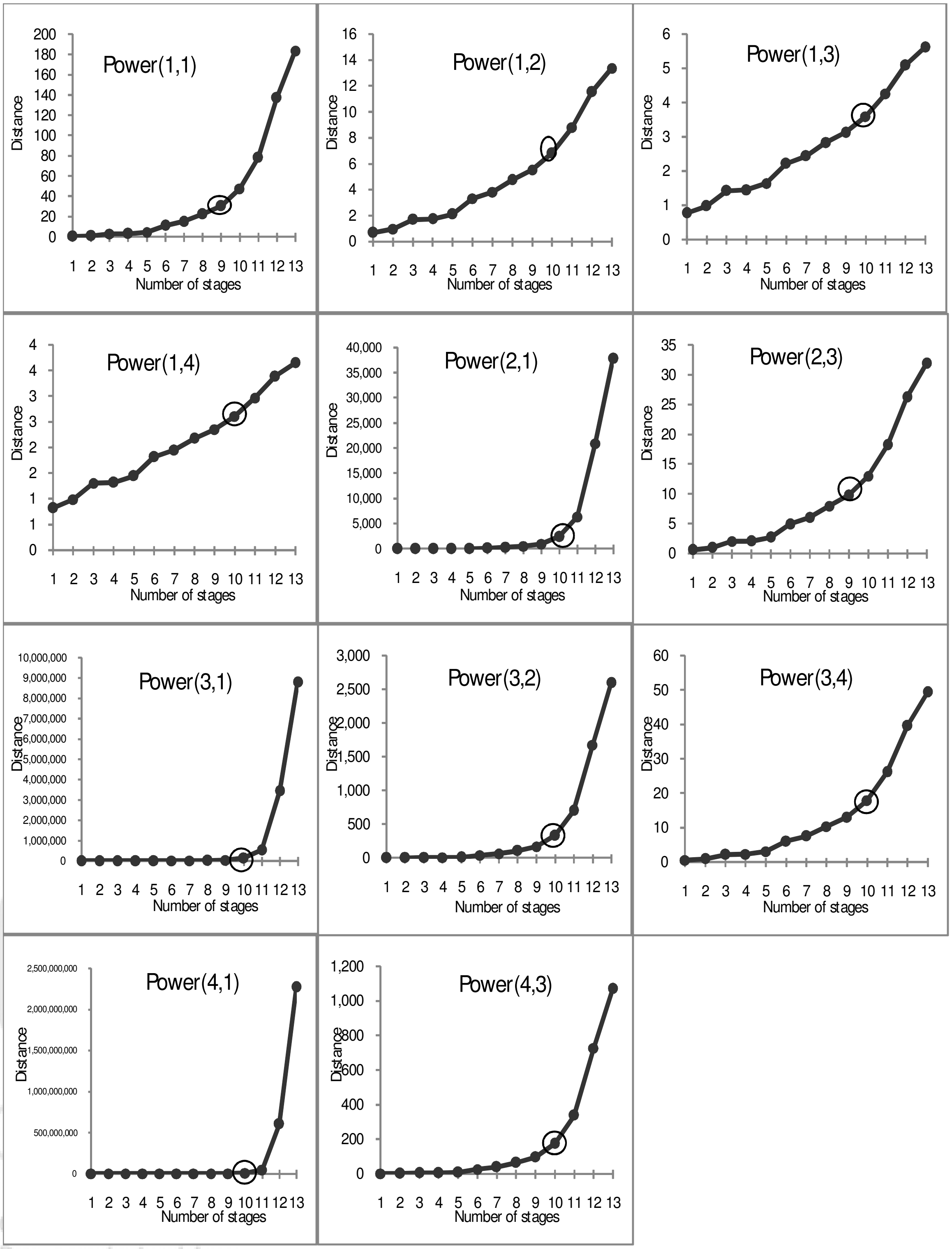
It is not advisable to go for too many or too few clusters. Simultaneously the graphs for all the four clustering methods are also made to obtain the point of elbow to verify the number of clusters. If there was any discrepancy arises in deciding the number of clusters based on the agglomeration schedule, final number of clusters taken as are suggested by the elbow criterion diagram. Summary of the number of clusters obtained for each of the four clustering methods and for each interval measure of distance is given in table 3.

Table 3: The number of clusters obtained for different clustering methods and interval measure of distance

	NUMBER OF CLUSTERS			
	B-AVERAGE	W-AVEARGE	SINGAL-LINKAGE	COMPLETE-LINKAGE
POWER(1,1)=MINKO-1, POWER(2,2)=MINKO-2, POWER(3,3)=MINKO-3, POWER(4,4)=MINKO-4, EUD, CHEBYCHEW AND CITY BLOCK GIVE THE SAME RESULTS	6	5	4	4
POWER(1,2) & POWER(2,4) GIVE THE SAME RESULTS	5	6	5	5
POWER(1,3)	5	5	5	5
POWER(1,4)	5	5	5	5
POWER(2,1), EUD^2 & POWER(4,2) GIVE THE SAME RESULTS	5	5	5	5
POWER(2,3)	6	6	5	6
POWER(3,1)	5	5	5	5
POWER(3,2)	5	5	5	5
POWER(3,4)	5	6	5	5
POWER(4,1)	5	5	4	4
POWER(4,3)	5	6	5	5

As mentioned earlier, graphs are presented only for one method for the brevity of the results. It is observed that elbow criterion diagram is a better alternative than to decide alone on the basis of agglomeration schedule because even a very small twist in the graph is

Figure 1: Elbow criterion diagrams for between the group linkage method



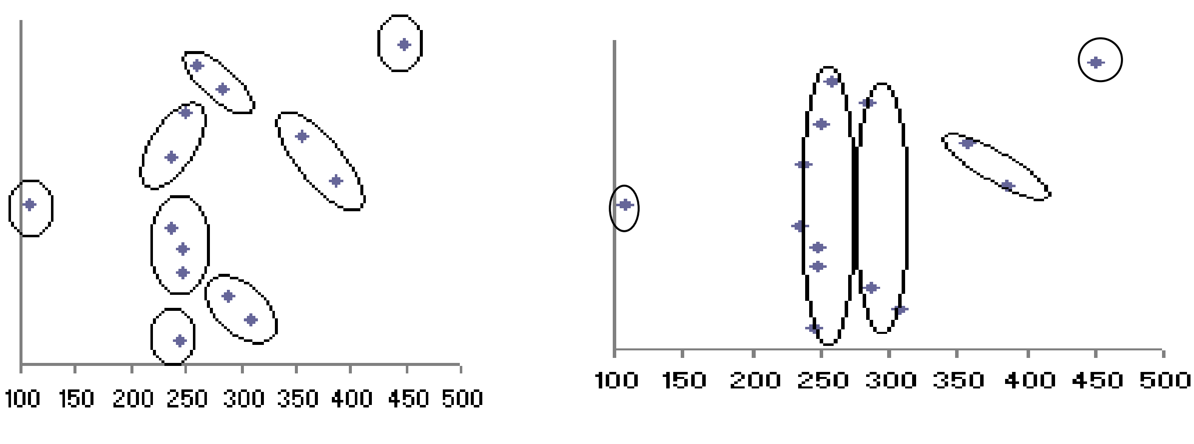
clearly visible and hence gives more confidence. If we obtain a consensus on the number of clusters among all the four clustering techniques for each of the measure, it is seen that in
most of the cases, 5 clusters are suggested. k-means clustering technique for k=5 is applied to get the actual member of the clusters. The k-means clustering method gives:

Cluster centers (k-means clustering)	1	2	3	4	5
Initial	108.18	356.08	287.24	236.17	449.51
Final	108.18	371.34	293.30	245.92	449.51

Cluster Membership

State	Cluster	Distance
AP	4	1.633
AS	3	15.017
GJ	3	6.063
HR	4	0.927
HP	4	1.867
KT	4	9.753
KL	1	0.000
MP	2	15.255
MH	4	9.293
OR	2	15.255
PJ	4	4.947
RJ	3	8.953
TN	4	12.937
UP	5	0.000

Figure 2: Showing (a) arbitrary clusters and (b) natural clusters



(a) (b)

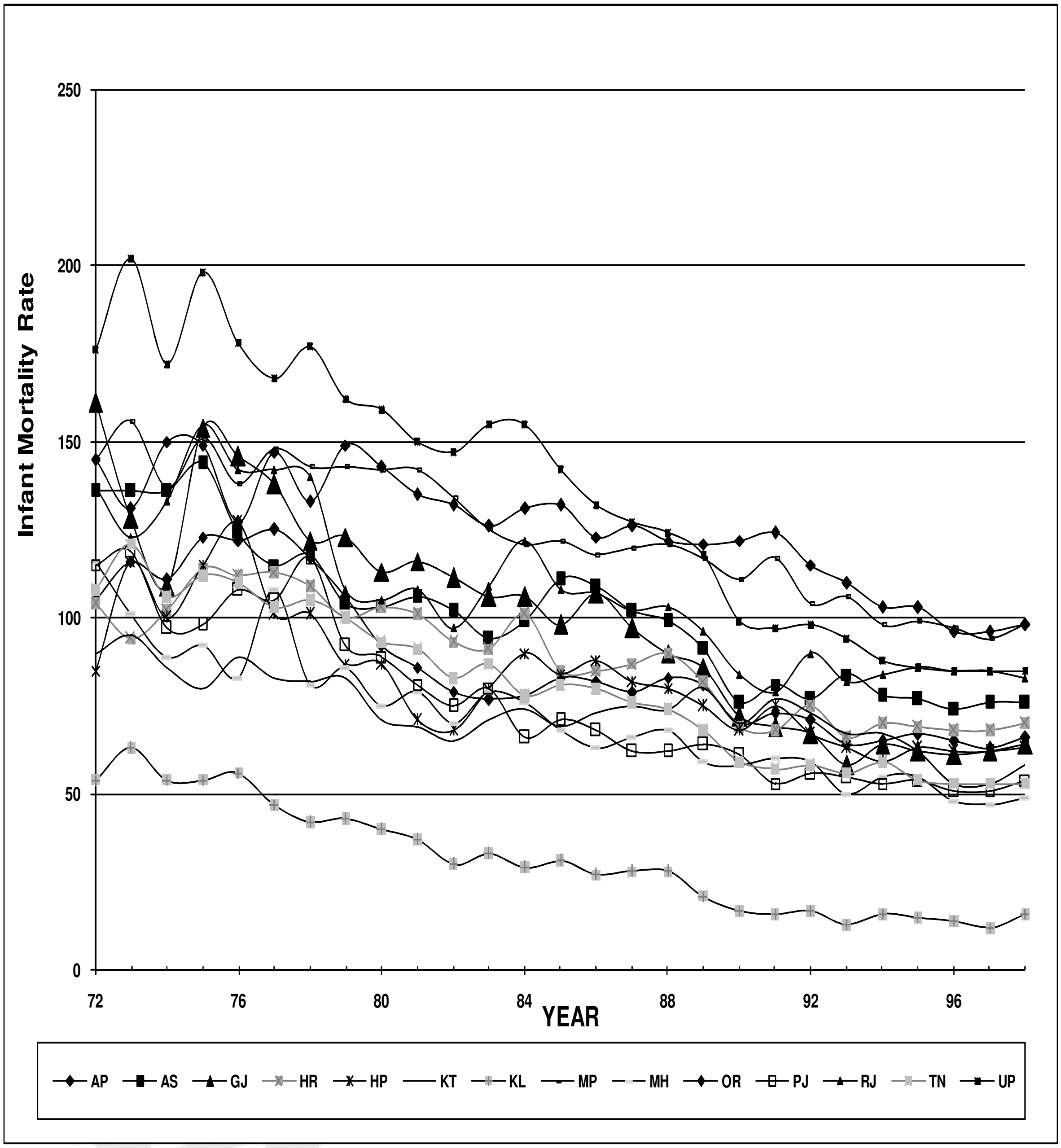
These members are: Cluster I: KL; Cluster II: MP, OR; Cluster III: AS, GJ, RJ: Cluster IV: AP, HR, HP, KT, MH, PJ, TN: Cluster V: UP

4. Discussion

Reliability is nothing but the repetition of the same result. After an extensive search of literature on internet and in journals, it is found that there is paucity of studies on the reliability of the methods for cluster analysis of time trends. In this paper the reliability of the model proposed by Bansal and Sharma (2003) is studied by applying the different clustering techniques by changing different interval distance measures one by one available in SPSSv17.0. Kerr and Churchill (2001) utilizes an analysis of variance model to achieve normalization and estimate differential expression of genes across multiple conditions. They applied bootstrapping to assess the stability of results from a cluster analysis. Tarpey (2007) showed that clustering the raw data would often give results similar to clustering regression coefficients, obtained using an orthogonal design matrix. Clustering functional data using an L^2 metric on function space can be achieved by clustering a suitable linear transformation of the regression coefficients. Que and Tsui (2008) obtained a multi-level spatial clustering algorithm for detection of disease outbreaks by using Kulldorff's spatial scan statistic and Bayesian spatial scan statistic. Richards et al. (2008) compared four clustering methods for brain expression micro array data. Mun et al. (2008) used the model-based cluster analysis to investigate population heterogeneity utilizing finite mixture multivariate normal densities and accordingly to classify subpopulations using more rigorous statistical procedures for the comparison of alternative models. Johnson et al. (2007) used trajectory cluster analysis to characterize and identify the trends in average ambient ozone and fine particulate matter levels. Monda and Popkin (2005) used cross sectional samples of children from the longitudinal data sets to correlate the activity and BMI status through clustering techniques. Sacchi et al. (2005) described a new technique of clustering through temporal abstraction based on a qualitative representation of profiles. They visualized the TA-clustering algorithm as a three-level hierarchical tree of qualitative representations which is easy to interpret and better than the standard hierarchical clustering techniques. Longstreth et al. (2001) applied the cluster analysis and studied the pattern on the findings on cranial magnetic resonance imaging of the elderly: the cardiovascular health study, a longitudinal study. Most of the studies are done on cross sectional data at a single time point. Stanfel (1986), Wallenstein (1980), Kafadar & Karon (1993) and Browdy (1982) clustered the time trends, but in the model given by Bansal and Sharma (2003), the divisions of the trend are decided objectively by the degree of the trend and number of time points, which is not seen in any of the previous studies listed.

It is clear from table 3 that there are five stable clusters. Single linkage and complete linkage method gives the same results and are in one to one correspondence. Square of Euclidean distance is the appropriate distance measure for such type of data. Few methods and measures have suggested 4 or 6 clusters. But if we take 4 clusters then the Cluster III States AS, GJ and RJ are merged with Cluster IV states. By looking at figure 3 we observe that these 3 states are more close to each other than the Cluster IV states,
hence five cluster solution is better. Similarly in case of 6 clusters solution, MP and OR are moving side by side till the last year except at for a period of 3 years from 1989 to 1991 which may be attributed to chance or errors as quite evident from figure 3. Since there was no gold standard available to compare our results, we took printouts of all the trends on

Figure 3: Time trends of Infant Mortality Rate of 14 major states of India



separate transparencies and super imposed them one by one over each other and found that there are five natural clusters. Although it was not required to study the reliability of the model proposed. But to have more confidence to suggest that this model can be reliably used to study the clustering of such type of time trends. The beauty of this model is that 2-

way dimensions of the trend i.e. horizontal flow of the trend and vertical distance of the trend from a common base are considered. The divisions of the trend are decided objectively by a formula which minimizes the subjectivity. Clustering of time trends is really more important than to cluster at a single time point because it gives more strength to the planners to predict the future trend based on their past behavior. Better strategies can be devised and policies can be implemented to combat the adversities in future. By this model, differences and similarities among the clusters can be studied at more ease than to study the individual clustering items especially in case of longitudinal datasets. It also becomes easier to study the homogeneity and heterogeneity in dynamics of a disease or phenomenon over a period of time.

5. References

- Bhattacharya, On a measure of divergence between two multinomial populations. Sankhyaa, The Ind J of Stat 7(1) (1945) 401-406.
- A.K.Bansal, A.Indrayan, Computer based statistical study of cartography in mortality upto age of one year. Ind Pediatr, 30 (1993) 1251-1258.
- A.K.Bansal, S.D.Sharma, A model for clustering of longitudinal data sets of infant mortality rates in India. Med Sci Monit 9(4) (2003) PH1-6.
- A.K.Jain, R.C.Dubes, Algoritm for Clustering Data. Prentice Hall Eaglewood Cliffs, 1988.
- A.L. Richards, P.Holmans, M.C. O'Donovan, M. J. Owen, L. Jones, A comparison of four clustering methods for brain expression microarray data. BMC Bioinformatics 9 (2008) 490.
- B.L.Browdy, P.C.Chang, Bayes procedures for the classification of multiple polynomial trends with dependent residuals. J Am Stat Assoc 77 (1982) 483-487.
- D. Johnson, D.Mignacca, D.Herod, D.Jutzi, H.Miller, Characterization and identification of trends in average ambient ozone and fine particulate matter levels through trajectory cluster analysis in eastern Canada. J Air Waste Manag Assoc 57(8) (2007) 907-18.
- D.M.Dunn, J.M. Landwehr, Analysing clustering effects across time. J Am Stat Assoc 75(369)(1980)8-15.
- E.Y. Mun, M.Windle, L.M.Schainker, A model-based cluster analysis approach to adolescent problem behaviors and young adult outcomes. Dev Psychopathol 20(1) (2008) 291-318.
- http://biocomp.bioen.uiuc.edu/oscar/tools/Hierarchical Clustering.html - Accessed on 18/4/2009.
- J Hartigan, Clustering Algorithms. Wiley, 1975.
- J. Zupan, Clustering of large datasets. John Wiley and Sons, 1982.
- J.Que, F.C. Tsui, A Multi-level Spatial Clustering Algorithm for Detection of Disease Outbreaks. AMIA Annu Symp Proc (2008) 611–615.
- K. Fukunaga, Introduction to Statistical Pattern Recognition. Academic Press, 1972.
- K. Kafadar, J.M.Karon, An analysis of AIDS incidence data by clustering trends. Statistics in Medicine 12 (1993) 311-326.

- K.L.Monda, B.M.Popkin, Cluster analysis methods help to clarify the activity-BMI relationship of Chinese youth. Obes Res 13(6) (2005) 1042-51.
- K.Ulm, Classification on the basis of successive observations. Biometrics 40 (1984) 1131-1136.
- L.E.Stanfel, Application of clustering theory to cancer mortality data. Comp and Biomed Res 19 (1986) 117-141.
- L.Sacchi, R.Bellazzi,C. Larizza, P.Magni, T.Curk, U.Petrovic, B.T.A. Zupan, clustering: cluster analysis of gene expression profiles through Temporal Abstractions. Int J Med Inform 74(7-8) (2005) 505-17.
- M.J.Symons, R.C.Grimson, Y.C.Yuan, Clustering of rare events. Biometrics 39 (1983) 193-205.
- M.K. Kerr, G.A. Churchill, Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. Proc Natl Acad Sci USA 98(16) (2001) 8961–8965.
- M.R.Anderberg, Cluster analysis for applications. Academic Press Inc., New York 1973.
- N. Ishii, K. Katano, et.al., Classification of time series by distance measure. Systems Computers Controls 12 (1981) 55-63.
- P.A. Devijver, J.Kittler, Pattern Recognition: A Statistical Approach. Prentice Hall, Englewood cliffs, 1982.
- S.Wallenstein, A test of detection of clustering over time. Am J Epidemio 111(3) (1980) 367-372.
- Sample Registration System, Registrar General Of India 1972 to 2000.
- SPSS Base Ver 10.0, Application Guide. Statistical Software for Social Sciences, SPSS Inc.
- SPSS Base Ver 17.0, Application Guide. Statistical Software for Social Sciences, SPSS Inc.
- T.Tarpey, Linear Transformations and the *k*-Means Clustering Algorithm: Applications to Clustering Curves. Am Stat 61(1) (2007) 34–40.
- W.T.Jr.Longstreth, P.Diehr, T.A.Manolio, N.J.Beauchamp, C.A.Jungreis,D. Lefkowitz, Cluster analysis and patterns of findings on cranial magnetic resonance imaging of the elderly: the Cardiovascular Health Study. Arch Neurol 58(4) (2001) 635-40.
- W.W.Copley, P.R.Lohens, Multi-variate Data Analysis. Wiley, 1971.