Date of Publication: 30 April 2018 | Pages: 3-8 | Volume 6, Issue 2

A Case Study: Natural Clustering Among Indian States

Varghese F, National Institute of Mental Health and Neuroscience, Bangalore,India,febavarghese82@gmail.com

ABSTRACT: India with a population of 1.34 billion stands as the second populous country in the world. In India about 51 births takes place in a minute. Child health plays a vital role in the development of a country. Health of the population significantly affects both social development and economic progress. Given the relevance of health for human well-being and social welfare, it is important to ensure equitable access to health care services by identifying priority areas and ensuring improvements in quality of healthcare services. Recent studies had reflected that the neighbourhood plays a crucial role in the health status. Socio- economic status of the neighbourhood has linked with the mortality, general health status, disability, birth-rate, chronic condition, health behavior and other risk factors for chronic disease, as well as mental health, injuries, violence's and other indicators of health [4]. This study aims to determine whether on the basis of maternal and children health status, there could be any natural clustering among the different districts of India. K mean clustering was used to find the number of clusters among Indian states. According to the majority rule, 2 would be the best number of clusters in the data set. In fact, 10 among 27 indices select 2 as the optimal number of cluster. Hence, the majority rule seems to be a more reliable solution for selecting the best number of clusters. Hence the different districts are grouped together to form two natural clusters. This implies that the health status of children in these district are interdependent. Not only the factors within one district are responsible for the health status of the children, there is also a great influence from the neighbouring districts. In order to mould a better future generation, the focuses should be made in the entire country.

Keywords: Infant mortality, India, clustering, Principal component K-mean clustering, Nb-clust.

1. INTRODUCTION

India with a population of 1.34 billion stands as the second populous country in the world. In India about 51 births takes place in a minute. Child health plays a vital role in the development of a country. In order to secure the future generation, India's major focus were on the health status of the children. Currently, projects and programmes were established to improve the health status of the country.

Health of the population significantly affects both social development and economic progress. Given the relevance of health for human well-being and social welfare, it is important to ensure equitable access to health care services by identifying priority areas and ensuring improvements in quality of healthcare services. Recent years have therefore witnessed a surge in demand from various quarters for generating timely and reliable statistics at the district level to enable informed decision making in the health sector. Timely and systematic estimates of the magnitude and changes in health indicators can play a crucial role in creating and assessing policies which aim to eliminate the disproportionate burden of health deprivations among disadvantaged populations.

Recent studies had reflected that the neighbourhood plays a crucial role in the health status. Socio- economic status of the neighbourhood has linked with the mortality, general health status, disability, birth-rate, chronic condition, health behaviour and other risk factors for chronic disease, as well as mental health, injuries, violence's and other indicators of health [4].

2. OBJECTIVE

On the basis of maternal and children health status, there could be any natural clustering among the different districts of India.

3. MATERIALS AND METHODS

The data is an aggregate of the Annual Health Survey (AHS) and the District Level Household Survey (DLHS) of 2012-2013. The data contains information regarding the health status of the 28 states in India [16].

Although it is a multivariate study, it is difficult to analyses this data. So in order to reduce the dimension of the data we may employ the principal component method.

The study focused on the health status of children, so the maternal and childhood factor were considered for the analysis [6][11][12][18]. For the principal component analysis the following variables were considered;

Date of Publication: 30 April 2018 | Pages: 3-8 | Volume 6, Issue 2

STATES	DISTRICTS
Andaman and Nicobar Islands	3
Andhra Pradesh	13
Arunachal Pradesh	16
Assam	23
Bihar	37
Chhattisgarh	16
Goa	2
Haryana	21
Himachal Pradesh	12
Jharkhand	19
Karnataka	30
Kerala	12
Madhya Pradesh	45
Maharashtra	35
Manipur	9
Meghalaya	7
Mizoram	8
Nagaland	11
Odisha	30
Puducherry	4
Punjab	20
Rajasthan	32
Sikkim	4
Tamil Nadu	32
Telangana	10
Tripura	4
Uttar Pradesh	70
Uttarakhand	13
West Bengal	19
TOTAL	558

1) Children aged 12-23 months who have received 3 doses of polio vaccine (%) – total.
2) Children aged 12-23 months who have received DPT vaccine (%) – total.
3) Children aged 12-23 months who have received measles vaccine (%) – total.
4) Children aged 12-23 months who have received vitamin a (%) – total.
5) Children with birth weight less than 2.5 kg. (%) – total.
6) Children (aged 6-35 months) exclusively breastfed for at least six months (%) – total.
7) Mothers who received any Antenatal Check-up (%) - Total
8) Mothers who consumed IFA for 100 days or more (%) - Total
9) Children aged 12-23 months Fully Immunized (%) - Total
10) Mothers who received at least one Tetanus Toxoid (TT) injection (%) – Total

4. RESULTS

Principal component analysis is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated components. The central idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of variables, the principal components (PCs), which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all of the original variables [1][8].

Now the percentage of variance explained by each principal component is explained in Table 1

Table 1: Eigen value

	eigenvalue	Variance percent
Dim.1	4.684565	52.05072
Dim.2	1.741802	19.35335
Dim.3	1.015025	11.27806
Dim.4	0.760797	8.453299
Dim.5	0.313406	3.482292
Dim.6	0.207199	2.302211
Dim.7	0.184719	2.052431
Dim.8	0.075868	0.842979
Dim.9	0.016619	0.18466

Here we can see that the first principal component contains 52% of the total variation in the observed variables. The second principal component contains 19% of the total variation in the observed variables and so on. The first, second and third principal component explains the most variations. The visual representation of the variation can be represented by a scree plot.

Scree plot is a simple line segment plot that shows the fraction of total variance in the data as explained by each principal component. The figure shown below gives the scree plot drawn using the package R. In this graph, the Eigen values or variances are plotted against the principal components [10].

Date of Publication: 30 April 2018 | Pages: 3-8 | Volume 6, Issue 2

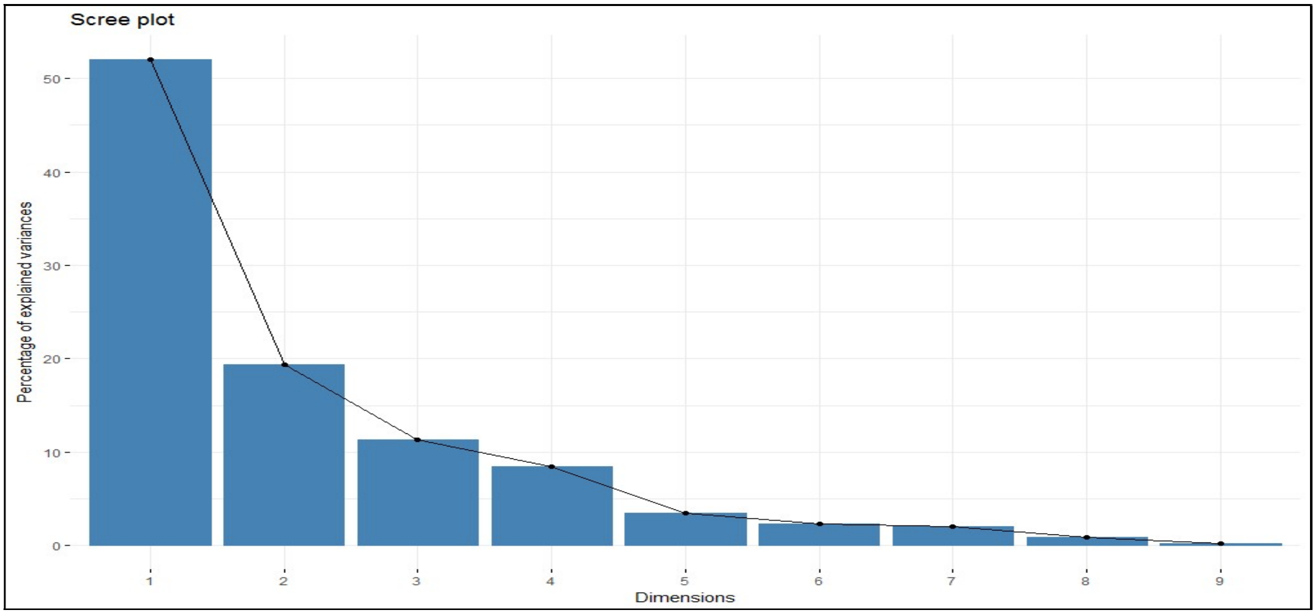


Fig-1: scree plot

Now we want to find whether there is any natural clustering among all the districts in India. For that purpose we are doing cluster analysis. The goal of cluster analysis is to partition the observations into groups ("clusters") so that the pair wise dissimilarities between those assigned to, the same cluster tends to be smaller than those in different clusters. The attempt is to maximize the homogeneity of objects within the clusters while maximizing the heterogeneity between the clusters. There are different algorithms for finding clusters, but here we are employing the k-means algorithm for finding the number of clusters.

Before moving to cluster analysis we should make a rough idea about how may clusters should be possible. For this purpose a Bi-plot is used. From the Bi-plot figure (fig-2), one can understand that there could be a possibility of 2 clusters.

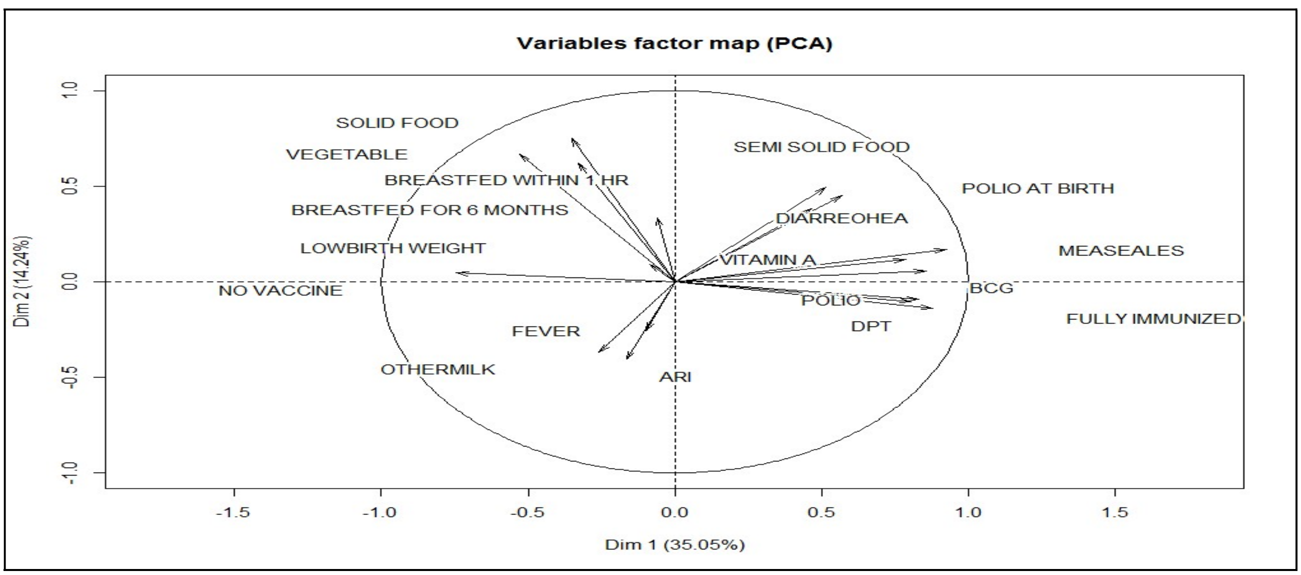


Fig- 2 Bi-plot

Now we can proceed with k means clustering method. K Means Clustering [3] is an **unsupervised learning algorithm** that tries to cluster data based on their similarity. **Unsupervised learning** means that there is no outcome to be predicted, and the algorithm just tries to find patterns in the data. In 'k' means clustering, we have specify the number of clusters we want the data to be grouped into. The algorithm randomly assigns each observation to a cluster, and finds the centroid of each cluster. Then, the algorithm iterates through two steps.

Date of Publication: 30 April 2018 | Pages: 3-8 | Volume 6, Issue 2

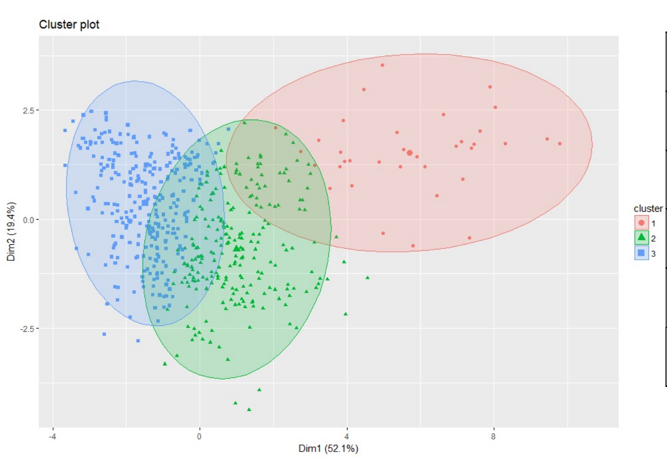
The algorithm of k mean clustering is as follows;

1. ***Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.***
2. ***Assign each object to the group that has the closest centroid.***
3. ***When all objects have been assigned, recalculate the positions of the K centroids.***
4. ***Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.***

The figures shown below gives the k means clusters drawn using R package;

a) If we recommend for 2 cluster ,the k means method will provide the 2 clusters as follows;

b) If we recommend for 3 cluster the k means method will provide the 3 clusters as follows;



c) If we are recommending for 5 clusters the k-means will provide the clusters as follows;

But from the above figure we can understand that cluster's 2, 3, 4 and 5 are overlapping each other. So it is best to have either 2 or 3 clusters. Now there arises a serious question that how many natural clusters can be formed by these districts in India? For that we should determine the optimal number of clusters. There are different methods for determining optimal number of clusters. The R package NbClust [2] is very effective for this purpose. It provides 30 indices which determine the number of clusters in a data set and it offers also the best clustering scheme from different results to the user. In addition, it provides a function to perform k-means and hierarchical clustering with different distance measures and aggregation methods. Any combination of validation indices and clustering methods can be requested in a single function call. This enables the user to simultaneously evaluate several clustering schemes while varying the number of clusters, to help determining the most appropriate number of clusters for the data set of interest.

Table 2: The 30 indices used in NbClust package

CH index	Duda index	Pseudot 2 index	C index	Gamma index
Beale index	CCC index	Ptbiserial index	G plus index	DB index
Freg index	Hartigan index	Ratkowsky index	Tau index	Ball index
Trcovw index	Tracew index	Freidman index	Rubin index	KL index
SD index	Dunn index	McClain index	Gap index	Scott index
SD bw index	Hubert index	Silhouette index	Dindex	Marriot index

Date of Publication: 30 April 2018 | Pages: 3-8 | Volume 6, Issue 2

Clustering with index argument such as Gamma, Tau and Gplus, is computationally very expensive, especially when the number of clusters and objects in the data set grows very large. Thus the index value for the 27 indices are calculated using Nb Clust package.

From table 3, one can get an idea about how many clusters can be formed. The number of cluster suggested by each indices can be determined the index value. For each indices has its own cut-off value. According to that CH index, CCC index, DB index, Silhouette index, Beale index, Ptbiserial index, Freg index, McClain index, Dunn index and SD index suggest that there can be a possibility of two clusters.

From the NbClust package among all the 27 indices,

- 2 proposed 0 as the best number of clusters
- 11 proposed 2 as the best number of clusters
- 9 proposed 3 as the best number of clusters
- 1 proposed 5 as the best number of clusters
- 1 proposed 9 as the best number of clusters
- 3 proposed 10 as the best number of clusters

4. CONCLUSION

According to the majority rule, 2 would be the best number of clusters in the data set. In fact, 10 among 27 indices select 2 as the optimal number of cluster. Hence, the majority rule seems to be a more reliable solution for selecting the best number of clusters. Hence the different districts are grouped together to form two natural clusters. This implies that the health status of children in these district are interdependent. Not only the factors within one district are responsible for the health status of the children, there is also a great influence from the neighbouring districts. In order to mould a better future generation, the focuses should be made in the entire country.

Table 3: Index value of 27 indices

	2	3	4	5	6	7	8	9	10
KL	3.6326	0.6809	3.7383	0.6015	1.1393	1.2871	0.7209	0.177	12.395
CH	249.3232	198.4629	133.6608	110.8044	106.1318	91.0204	80.4337	70.8079	91.8066
Hartigan	102.3996	2.7871	24.9662	49.0723	8.3895	9.0082	2.2037	128.6065	6.289
CCC	-6.0067	-9.3754	-17.1871	-21.5015	-23.0604	-24.9375	-26.3212	-28.1073	-18.0479
Scott	621.0487	1035.085	1058.649	1207.545	1431.214	1559.202	1621.757	1638.309	2028.312
Marriot	5.45E+16	5.84E+16	9.95E+16	1.19E+17	1.15E+17	1.24E+17	1.45E+17	1.78E+17	1.09E+17
TrCovW	195437.9	107464.9	106295.6	100782.5	90558.97	88765.85	86323.02	85396.53	62196.88
TraceW	2691.893	2273.229	2261.87	2164.333	1987.928	1958.167	1926.668	1918.979	1554.766
Friedman	40.4513	42.2809	42.3711	47.8373	48.9981	50.335	51.8395	51.9516	61.2033
Rubin	1.4484	1.7152	1.7238	1.8015	1.9613	1.9911	2.0237	2.0318	2.5078
Cindex	0.3446	0.3408	0.34	0.3391	0.3634	0.3629	0.3618	0.3618	0.3701
DB	0.8615	1.574	1.3325	1.2505	1.1986	1.4364	1.3744	1.2845	1.2802
Silhouette	0.5006	0.179	0.1714	0.1589	0.1575	0.1488	0.1347	0.1263	0.1539
Duda	0.8262	1	0.6556	0.8844	0.8256	0.9305	1.0252	0.7168	0.4991
Pseudot2	107.7147	-0.0034	22.0655	45.7583	7.3936	11.8829	-0.4428	130.3696	8.029
Beale	0.9587	0	2.343	0.5952	0.9377	0.3391	-0.1064	1.7984	4.0733
Ratkowsky	0.3524	0.3617	0.3146	0.2889	0.2816	0.2627	0.2475	0.2339	0.2431
Ball	1345.947	757.7428	565.4674	432.8666	331.3213	279.7381	240.8335	213.2199	155.4766
Ptbiserial	0.6679	0.3865	0.3881	0.3907	0.4206	0.4206	0.421	0.421	0.397
Frey	10.3601	0.0642	-0.3158	0.0652	0.8819	0.6029	0.7202	0.7026	0.1017
McClain	0.0813	0.7065	0.7115	0.7141	0.8127	0.818	0.8301	0.8305	1.7036
Dunn	0.1284	0.0613	0.0613	0.0613	0.0687	0.0687	0.0687	0.0687	0.0811
Hubert	9.00E-04	8.00E-04	8.00E-04	9.00E-04	9.00E-04	9.00E-04	9.00E-04	9.00E-04	9.00E-04
SDindex	1.4467	1.9337	1.7099	1.6673	1.7027	1.662	1.6545	1.5665	1.5857
Dindex	2.08	1.8882	1.8822	1.8476	1.7888	1.7773	1.763	1.7579	1.5902
SDbw	1.0923	1.0717	0.6848	0.4737	0.4795	0.468	0.4323	0.3834	0.3776
Gap	0.2411	-0.4896	-0.9198	-1.2883	-1.568	-1.8824	-2.1885	-2.335	-2.4246

NB: The red colored value represent the critical index value used to determine the optimal number of cluster

Date of Publication: 30 April 2018 | Pages: 3-8 | Volume 6, Issue 2

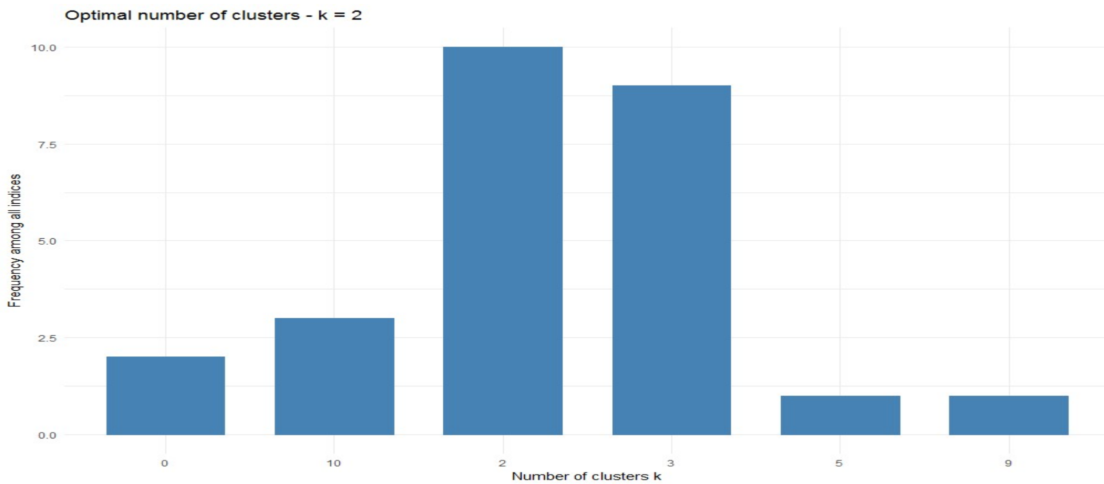


Fig -3 Histogram of 27 indices

REFERENCES

1 Abdi H, Williams LJ (2010) Principal component analysis. Wiley Interdiscip. Rev. Comput. Stat. 2:433–459.

2 Charrad M, Ghazzali N, Boiteau V, Niknafs A (2014) NbClust : An R Package for Determining the Relevant Number of Clusters in a Data Set. J Stat Softw 61:1–36. doi: 10.18637/jss.v061.i06

3 Hartigan JA, Wong MA (1979) A K-means clustering algorithm. Appl Stat 28:100–108. doi: 10.2307/2346830

4 Ross NA, Tremblay S, Graham K (2004) Neighbourhood influences on health in Montréal, Canada. Soc Sci Med 59:1485–1494. doi: 10.1016/j.socscimed.2004.01.016

5 Anderson TW (1958) An Introduction to Multivariate Statistical Analysis.

6 Boyer C, Jackson E, Bawah A, et al (2015) Estimating indices of health system readiness : an example from rural northern Ghana. Lancet Glob Heal 3:14. doi: 10.1016/S2214-109X(15)70133-6

7 Edmond KM, Wardlaw T, Brown DW, et al (2006) Delayed breastfeeding initiation increases risk of neonatal mortality. Pediatrics 117:380–386. doi: 10.1542/peds.2005-1496

8 Hair JF, Anderson RE, Tatham RL, Black WC (1998) Multivariate Data Analysis.

9 Hair JF, Black WC, Babin BJ, Anderson RE (2010) Multivariate Data Analysis. Vectors 816.

10 Hastie T, Tibshirani R, Friedman J (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction.

11 Imdad A, Mayo-Wilson E, Herzer K, Bhutta ZA (2017) Vitamin A supplementation for preventing morbidity and mortality in children from six months to five years of age. Cochrane Database Syst Rev. doi: 10.1002/14651858.CD008524.pub3

12 Jackson EF, Siddiqui A, Gutierrez H, et al (2015) Estimation of indices of health service readiness with a principal component analysis of the Tanzania Service Provision Assessment Survey. BMC Health Serv Res 15:536. doi: 10.1186/s12913-015-1203-7

13 Lai D (2003) Principal component analysis on human development indicators of China. Soc Indic Res 61:319–330. doi: 10.1023/A:1021951302937

14 Manly BFJ (1986) Multivariate statistical methods: a primer.

15 OECD (Organisation for Economic Co-operation and Development) (2008) Handbook on Constructing Composite Indicators: Methodology and User.

16 Office of the general & census commissioner I, ministry of home affairs goi, 2/a,mansingh road nd-110011 (2014) annual health survey report.

17 Rencher AC (2012) Methods of Multivariate Analysis, Second Edition.

18 Tripathy JP, Mishra S (2017) Causes and Predictors of Neonatal, Post-neonatal and Maternal Deaths in India: Analysis of a Nationwide District-Level Household Survey-4 (DLHS-4), 2012–13. J. Trop. Pediatr.