Original Research Article

Clustering Analysis of the Survey for Mobility Reasons in the US 1999-2017

ABSTRACT

This paper is to estimate the survey for 98000 addresses from 1999-2017 in United States bureau of Census by using cluster analysis. The analysis is mainly applied by Approximate Covariance Estimation for CLUSTING (ACECLUS), and procedure variables for CLUSTING (VARCLUS) to test some important parameters such as average linkage, two-stage density linkage, Cubic Clustering Criterions (CCC), R-Square, Ward's minimum variance techniques, as well as Tree procedure for deeper exploring the clusters or variables. After the overall analysis, the results show that there is existence of strong covariate correlation for variables X8 and X15 with respond variable Y (Mobility periods). Hence, Reason "Retired" from survey data is most important impact on mobility other than the reasons "Wanted better neighborhood or less crimes" and "Wanted cheaper housing" that are popular and highly frequent.

Key words: CLUSTERING analysis, Mobility, ACECLUS, CCC, R-Squared, Ward's minimum variance.

1. INTRODUCTION

1.1. Background

Mobility is involved an economic, health, neurological, thoughtful, and social activity. It needs move people and government have adaptability, versatility, adjustability, and flexibility[1,2].. In the United States, place of birth has long been an important measure of domestic mobility in the U.S. census [3] (Long, 1988). The American population experienced high mobility activity. Each year many people leave their original place to live, work, or study at another new locations [4] (Frey, 2009). However, although most of mobility are individual behaviors, they effect demographics, transportation, and economics, etc. Hence, population survey in United States is conducted each ten.years[5,6] This paper is a research based on redesigned questions for incomes and health insurance coverage, etc. from a population survey in the United States Census Bureau 2017. The improved income questions were implemented using a split panel design[7].

The Cluster analysis is one of strong practical statistical methods[8]. It has broadly application to all of fields in the world. The Clustering procedure is a hierarchical technique that finds the observations in a SAS data set with coordination and Euclidean.distances[9]. The clustering methods include average linkage, the centroid method, complete linkage, density linkage (i.e. Wong's hybrid, kth-nearest neighbor methods), maximum likelihood for mixtures of spherical multivariate normal distributions with equal variances but possibly unequal mixing proportions, the flexible-beta method, McQuitty's similarity analysis, the median method, single linkage, two-stage density linkage[10], and Ward's minimum-variance method, etc. These methods have corresponding to themselves characteristics based on general agglomerative hierarchical clustering procedures. Moreover, every observation or object gets start with one cluster, and then two mutual nearest clusters mix to create a new one that substitutes two old clusters. After that, new cluster would be reused until only one leaves[11]. However, since the CPU time varies over the numbers of observations, the cluster procedures are seldomly utilized for large data set. Hence, for these large data set, we should apply proc FASTCLUS to be preliminary cluster analysis hierarchically.

However, we might not look the statistical method application based on mobility reasons in the journals or magazines. Hence, the author in this paper would like to explore and use some cluster analysis to solve the survey data of mobility reason.

1.2. Related Works

Many Researchers and organizations make some statistical analysis for census survey. For example, Ansary, et al. thought that "new census towns in the 2011 might be the real driving force for this staggering increase" in India [12]. Flippo et al. used a continuum approach to compute the probability for observing a trip that arbitrary region and the fluxes between two regions. The finding was that the complex topological feature observed in large mobility and transportation networks might be the result of a single stochastic process[13]. Yolanda explored motivations, job satisfaction using qualitative semi-structural interviews for a sample of thirty African immigrant workers in Pittsburg metropolitan areas in the United States. The result reveals that more people must make direct care work attractive and rewarding for African immigrant workers[14]. Xie, L.M thought that, since time series models have strong application with flexibility, adaptability, versatility, adjustability, time series models do fit for all fields including mobility data set, such as spectral analysis and filtering, ARIMA[15], and SARIMA[16], etc. Nawrotzki, et al. used theoretical framework to analyze U.S- Mexico bound migration from rural locates. The result showed that a decrease in precipitation affect seriously migration life [17].

2. MATERIALS AND METHODS

The data comes from United States Census Government. This survey collected about 98000 addresses. Approximately 68000 addresses were selected to receive a set of income questions. It contains mobility periods from 1999-2017, the mobility reasons were "Total moves 1 year and over", "Changing in mental status", "To establish own household", "Other family reason", "New job or job transfer", "To look for work or lost job", "To the closer to work/easier commute", "Retired", "Other job related reason", "Want own home, not rent", "Wanted new or better home/apartment", "Wanted better neighborhood /less crimes", "Wanted cheaper housing", "Foreclosure/eviction" , "Other housing reason", To altered or leave college" , "Changing of climate" , "Health reasons" , "Natural disaster", and "Other reasons". In this paper, I used Y to represent mobility periods dependent, the above mobility reasons are replaced by X1-X20, respectively. the raw data set is shown in Table 1. Statistical Analysis is used SAS 9.4.

Table 1 Raw Data for this paper.

Table A-5. Reason for Move (All Categories): 1999-2017

(Numbers in thousands.)

Mobility Period	Total Movers 1 year and over	Change in marital status	To establish own household	Other family reason	New job or job transfer	To look for work or lost job	To be closer to work/easier commute	Retired	Other job related reason
NUMBER									
2016-2017/1	34,902	1,786	4,023	3,928	3,462	453	1,924	290	317
2015-2016/1	35,138	1,679	4,283	3,683	3,807	531	2,094	228	426
2014-2015	36,324	2,122	3,989	5,178	3,848	583	1,789	417	835
2013-2014 (98,000 address file)	35,681	1,742	3,966	4,797	3,471	745	2,211	249	694
2012-2013	35,918	1,817	3,753	5,301	3,242	750	1,941	237	809
2011-2012	36,488	2,300	3,906	4,487	3,470	659	1,997	182	750
2010-2011 (2010 controls)/3	35,038	1,939	3,325	4,494	2,801	909	2,068	110	534
2010-2011 (2000 controls)/4	35,075	1,949	3,334	4,501	2,829	924	2,081	108	539
2009-2010 (2010 controls)/3	37,445	2,731	4,188	4,407	2,935	951	1,583	199	481
2009-2010 (2000 controls)/4	37,540	2,756	4,211	4,409	2,928	971	1,590	198	488
2008-2009	37,105	1,993	3,514	4,266	3,244	1,005	1,872	148	385
2007-2008	35,167	1,987	3,682	5,069	2,940	794	2,183	140	1,295
2006-2007	38,681	2,296	3,808	5,555	3,789	648	1,858	214	1,551
2005-2006	39,837	2,395	3,389	5,241	3,481	638	1,440	170	1,599
2004-2005	39,888	2,841	3,108	4,882	4,157	758	1,366	216	540
2003-2004	38,995	2,401	2,716	4,357	3,569	922	1,444	128	561
2002-2003	40,093	2,679	2,814	5,055	3,546	749	1,275	101	576
2001-2002	41,111	2,517	3,140	4,930	4,331	952	1,239	230	648
2000-2001	39,006	2,330	2,943	5,337	4,023	805	1,215	224	434
1999-2000	43,388	2,651	3,144	5,755	4,517	726	1,465	181	578
1998-1999	42,636	2,761	3,273	4,947	4,042	676	1,324	245	842

Mobility Period	Wanted own home, not rent	Wanted new or better home/apartment	Wanted better neighborhood /less crime	Wanted cheaper housing	Foreclosure/eviction*	Other housing reason	To attend or leave college	Change of climate	Health reasons	Natural disaster**	Other reasons
NUMBER											
2016-2017/1	2,554	5,576	969	2,895	373	2,647	1,030	184	672	91	1,726
2015-2016/1	2,075	6,126	1,096	2,871	305	2,351	1,137	269	628	17	1,532
2014-2015	1,912	5,567	1,056	2,709	272	5,233	108	66	106	-	534
2013-2014 (98,000 address file)	2,004	5,629	1,054	3,356	477	4,578	177	31	129	10	359
2012-2013	2,099	5,332	1,135	2,989	654	5,016	215	20	136	11	462
2011-2012	1,711	5,810	1,229	3,260	792	5,239	198	16	101	55	328
2010-2011 (2010 controls)/3	1,544	5,690	1,379	3,696	412	3,062	896	149	564	32	1,435
2010-2011 (2000 controls)/4	1,530	5,665	1,360	3,684	412	3,085	890	149	564	31	1,439
2009-2010 (2010 controls)/3	1,725	5,772	1,553	4,056	-	3,269	1,024	238	565	108	1,658
2009-2010 (2000 controls)/4	1,734	5,800	1,530	4,067	-	3,275	1,031	237	565	105	1,646
2008-2009	2,031	5,387	1,871	4,125	-	3,610	955	196	592	135	1,774
2007-2008	2,033	4,866	1,778	2,872	-	2,549	872	212	460	61	1,373
2006-2007	2,292	6,096	2,130	3,089	-	2,628	743	146	531	177	1,130
2005-2006	3,415	7,090	1,754	2,451	-	3,679	1,064	175	504	669	683
2004-2005	3,704	7,091	1,609	2,648	-	3,759	1,289	234	633	-	1,052
2003-2004	3,627	8,235	1,840	2,865	-	4,009	1,118	224	394	-	583
2002-2003	4,078	7,942	1,530	2,622	-	4,406	1010	160	565	-	987
2001-2002	4,334	7,724	1,625	2,405	-	4,279	1,126	259	510	-	862
2000-2001	3,942	6,871	1,540	2,135	-	4,205	1,183	216	520	-	1,084
1999-2000	4,792	7,710	1,860	2,318	-	4,879	1,075	309	483	-	946
1998-1999	3,329	8,861	1,657	2,539	-	4,738	841	334	469	-	1,760

Statistical Analysis

Average Linkage

This linkage is the distance between two clusters that is defined by[18] $A = \pi r^2$

$$D_{ab} = \frac{1}{N_a N_b} \sum_{i \epsilon X_a} \sum_{j \epsilon X_b} d(z_i, z_j) \tag{1}$$

When $d(w, q) = ||w - q||^2$, we can obtain:

$$D_{ab} = ||\bar{w}_a - \bar{w}_b||^2 + \frac{S_a}{N_a} + \frac{S_b}{N_b} \tag{2}$$

Hence, we can obtain the following formula:

$$D_{lm} = \frac{N_a D_{la} + N_b D_{lb}}{N_m} \tag{3}$$

For average linkage, it is mainly measurement of distance between two clusters that the average distance for pairs of observations. It joins clusters with small variances, but it has the trend to be biased as creating clusters with the same variance (Sokal and Michener, 1958).

Uniform-kernel method

This method is using uniform-kernel density estimates. Suppose c is the value specified for the R=option. If a closed sphere centered at point *n* with radius c. Then, for the estimated density at *n*, *f (n)*, it has the proportion of observations within the sphere divided by the volume of the sphere:

$$d\,(n_i, n_j = \begin{cases} \frac{1}{2}\left(\frac{1}{f(n_i)} + \frac{1}{f(n_j)}\right) & \text{if } d(n_i, n_j) \leq c \\ \infty & \text{Otherwise} \end{cases} \qquad (4)$$

Two-Stage Density Linkage

Before all the points in the tails have clustered, this technique is used to merge the model clusters. The CLUSTER procedure in SAS is applied to display the same three varieties of two-stage density linkage as of ordinary density linkage, that is, kth-nearest neighbor, uniform kernel, and hybrid. (1) when disjoint model clusters are created, the algorithm is used as the single linkage algorithm. Finally, each point becomes one modal cluster. (2) single linkage joins the modal clusters hierarchically, as there are wide gaps between the clusters or small parameter, the final number of clusters often exceed one.

The TREE procedure plots a tree as every stage creates. For the second stage, In the proc Tree statement we can use the option HEIGHT=mode to obtain the tree. In addition, we could form a single tree diagram including two stages having the number of clusters at the height axis. Two-stage density was introduced by W.S. Sarle of SAS Institute Inc.

Ward's Minimum-Variance Method

This is a method that measure the distance between two clusters and it is defined as:

$$D_{ab} = G_{ab} = \frac{||\bar{x}_a - \bar{x}_b||^2}{\frac{1}{N_a} + \frac{1}{N_b}} \qquad (5)$$

When d (f, g) = $\frac{1}{2}||f - g||^2$, we can obtain the following equation:

$$D_{lm} = \frac{(N_L + N_a)D_{l_a} + (N_l + N_b)D_{lb} - N_l D_{lb}}{N_l + N_m} \qquad (6)$$

Ward's minimum-variance method measures the distance between two clusters. The distance represents that the analysis of variance sum of squares between the two clusters is over all the variables. The within-cluster sum of squares makes the minimum over all partitions received by merging two clusters from the former generation. When the two clusters are divided by the total sum of squares to give proportions of variance, the sum of squares displays the values of them.

There are some conditions for Ward's method that joins clusters to maximize the likelihood at every level of the hierarchy: (1) Multivariate normal mixture (2) equates spherical covariance matrices, (3) equate sampling probabilities.

3. RESULTS

Estimation of mobility is aligning of data analytics. The author thinks that using cluster analysis can understand a quick overview of data, master characteristics of groups in data, measure the similarity of variables in data, ordinate the deeper relationships, and enrich the ordination plots, etc. This paper is mainly applied Hierarchical clustering, fuzzy clustering, or division clustering by using SAS 9.4. I would to determine whether mobility figures for the variable "Want better neighborhood or less crimes", variable "Wanted cheaper housing" and other variables can be used to determine specific types or categories in reasons of mobility in the United States 1999 to 2017. So, the author uses a cluster analysis to confirm whether the observations might be incorporated to groups provided by the data.

The following plot is a kind of scatter figure that reflects variable 12 and its specific periods. The following is the plot that shows the relationship between the reason of mobility for “Wanted better neighborhood and less crimes” with their periods. It suggests the difficulty of dividing the points into clusters. Plots of the other variables (not shown) show similar properties. The clusters that comprise these data might poorly separate and elongated. Hence, it is necessary for the figure of the data with poorly separated or elongated clusters to be transformed.

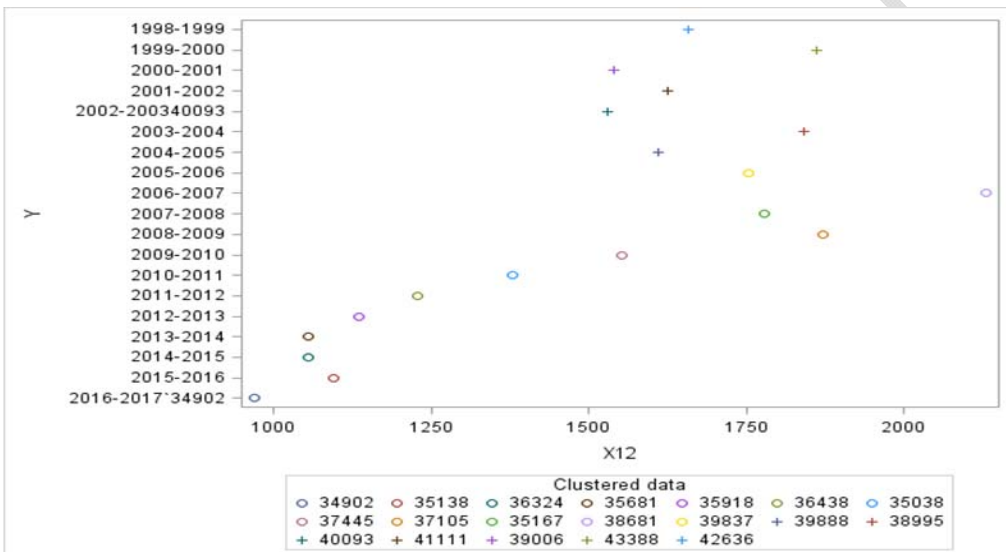


Figure 1. Plot for X12 (Wanted better neighborhood and less crimes) with their periods.

When we know the within-cluster covariance, I could transform the data to make the clusters spherical. However, we do not understand what the clusters are, so, the ACECLUS (Approximate Covariance Estimation for CLUSTERING) technique is used to estimate the within-cluster covariance matrix to transform the data. In Table 1. It is output of cluster analysis for the relationship between the reason of mobility that wanted better neighborhood and less crimes and total moves one year and over. The setting for the Proportion and Converge options. The Proportion option indicates set at 0.06, the CONVERGE parameter is set at default of 0.001. The result shows that 1508.7 of the mean and 336.6 of standard deviation, but the value of standard deviation for X8 is the lowest among the all of variables. So, X8 is significant one.

Table 2. Cluster Analysis using ACECLUS by X1*X12

Approximate Covariance Estimation for Cluster Analysis		
Means and Standard Deviations		
Variable	Mean	Standard Deviation
X1	38041.6	2631.1

X2	2261.4	376.1
X3	4824.7	475.0
X4	4824.7	546.0
X5	3614.5	473.6
X6	754.9	157.0
X7	1699.4	343.8
X8	205.7	72.2825
X9	729.2	370.3
X10	2800.1	1034.8
X11	6492.9	1159.6
X12	1508.7	336.6
X13	2942.2	552.7
X14	172.9	257.2
X15	3901.9	945.9
X16	842.7	375.1
X17	180.9	92.9591
X18	450.6	188.3
X19	71.3684	153.8
X20	1066.7	491.9
Variables 20	Proportion 0.0600	Converge 0.00100

The ACECLUS procedure is to get approximate estimates of the pooled within-cluster covariance matrix if the clusters area is assumed to be multivariate normal with equal covariance matrices. The type of matrix used for the initial within-cluster covariance estimate is the following table. In this case, that initial estimate is the diagonal covariance matrix. The threshold value that corresponds to 1.039077.

Table 3. Table of Iteration History from the ACECLUS Procedure

Initial Within-Cluster Covariance Estimate =Diagonal Covariance Matrix

Iteration	RMS Distance	Distance Cutoff	Pairs within Cutoff	Convergence Measure
1	6.325	6.572	90.000	0.243668
2	7.117	7.396	117.0	0.175394
3	29.313	30.458	136.0	0.117912
4	27.769	28.854	136.0	0.000000

Threshold=1.039077 Algorithm converged.

The above table indicates results that test root mean square (RMS) distance between all pairs of observations, distance cutoff for including pairs of observations in the estimate of within-cluster covariances (that is, RMS*Threshold), number of pairs within the cutoff, and convergence measure.

Table 4. Approximate covariance estimation for cluster analysis with eigenvalues of ACE* COV_ACE) (X1*X13).

Eigenvalues of Inv (ACE)*(COV-ACE)

	Eigenvalue	Difference	Proportion	Cumulative
1	294842	22.3915	0.9146	0.9146
2	7.0927	7.2038	0.9146	0.9146
3	-0.1111	8.5E-14	-0.00345	1.1312
4	-0.1111	1.08E-14	-0.00345	1.1278
5	-0.1111	7.22E-15	-0.00345	1.1174
6	-0.1111	3.11E-15	-0.00345	1.1209
7	-0.1111	2E-15	-0.00345	1.1174
8	-0.1111	2.44E-15	-0.00345	1.1140
9	-0.1111	2E-15	-0.00345	1.1105
10	-0.1111	7.77E-16	-0.00345	1.1071
11	-0.1111	1.89E-15	-0.00345	1.1037
12	-0.1111	6.66E-15	-0.00345	1.1002
13	-0.1111	9.77E-15	-0.00345	1.0968
14	-0.1111	6.22E-15	-0.00345	1.0933
15	-0.1111	4.32E-14	-0.00345	1.0899
16	-0.1111	0.1522	-0.00345	1.0864
17	-0.2633	0.2591	-0.00817	1.0782
18	-0.5224	0.4776	-0.0162	1.0620
19	-1.0000	3.23E-13	-0.0310	1.0310
20	-1.0000		-0.0310	1.0000
	ACE: Approximate Covariance Estimate Within Clusters			
	X1	X2	X12	X13
X1	7402599	913311.250	512071.000	-888621.688
X2	913311.250	157746.618	74764.004	-75147.629
X12	512071.000	74764.004	107786.154	-15241.654
X13	-888621.688	-75147.629	-15241.654	333951.154

The above table shows that the approximate within-cluster covariate matrix and the eigenvalues from the canonical analysis. In the first column of the table listed for the eigenvalues of Inv (ACE)*(COV-ACE), the next column of the table is the eigenvalues includes numbers for the eigenvectors. In other three columns for Eigenvalues of Inv (ACE)*(COV-ACE), it shows that the relative size and importance of the eigenvalues. “Difference” is defined as the region between each eigenvalue and its successor. The other two express the individual and cumulative proportions in which each eigenvalue is assigned to the total sum of eigenvalues.

The following table is a plot that is applied Ward’s minimum variance method to the data. The distance between two clusters is the analysis of variance sum of squares between the two clusters added up over all the variables. For each generation, the within-cluster sum of squares does have minimized values over all partition obtained by merging two clusters from the previous generation. The sums of squares are popular to explain as they are divided by the total sum of squares to give proportions of variance. This method involves clusters to maximize the likelihood at each level of the hierarchy under some conditions: Multivariate normal mixture; equal spherical covariance matrices; equal sampling probabilities, etc. However, Ward technique may enter clusters for small numbers of observations, and it is biased toward generating clusters for roughly the same number of observations. Also, it is sensitive to outlines (Milligan, 1980).

Table 5. Cluster Analysis

Ward's Minimum Variance Cluster Analysis

Number of Clusters	Clusters Joined	Freq	Semipartial R-Square	R-Square	Approximate Expected R-Square	Cubic Clustering Criterion	Pseudo F Statistic	Pseudo t-Squared
18	14-15 13-14	2	0.0000	1.00	.	.	6E4	.
17	02-03_40093 00-01	2	0.0000	1.00	.	.	4902	.
16	08-09 99-00	2	0.0000	1.00	.	.	3625	.
15	04-05 01-02	2	0.0000	1.00	.	.	2422	.
14	09-10 CL17	3	0.0001	1.00	.	.	1718	4.3
13	07-08 05-06	2	0.0001	1.00	.	.	1369	.
12	CL16 03-04	3	0.0002	.999	.	.	1101	7.2
11	15-16 12-13	2	0.0004	.999	.	.	824	.
10	CL15 89-99	3	0.0005	.999	.	.	669	8.3
9	CL11 CL18	4	0.0018	.997	.	.	376	9.3
8	CL12 CL13	5	0.0049	.992	.	.	190	38.1
7	16-17_34902 CL9	5	0.0053	.986	.	.	146	7.2
6	11-12 10-11	2	0.0055	.981	.	.	134	.
5	CL14 CL10	6	0.0059	.975	.	.	137	32.8
4	CL8 06-07	6	0.0391	.936	.	.	73.1	29.8
3	CL7 CL6	7	0.0410	.895	.918	-.83	68.2	15.7
2	CL5 CL4	12	0.1207	.774	.781	-.12	58.3	23..7
1	CL3 CL2	19	0.7743	.000	.000	.00	.	58.3

Eigenvalue 113297.871 Proportion 1.0000 Cumulative 1.0000
Root-Mean-Square Total-Sample Standard Deviation 336.5975
Root-Mean-Square Distance Between Observations 476.0207

To test if the parameters are coordinated with the data, I use proc CLUSTER. Because the output can check the estimators: (1) the eigenvalues of the correlation or covariance matrix; (2) the difference between successive eigenvalues; (3) the proportion of variance interpreted by each eigenvalue; (4) the cumulative proportion of variance interpreted; (5) the Root-Mean-Square Distance between observations; (6) the mean distance between observations. On the other hand, proc CLUSTER can also show that: (1) the number of Cluster; (2) the names of the Cluster jointed. The observations are identified by OBn (n, the observation numbers); (3) the number of observations in the new cluster, frequency of new cluster or Freq. But, if we want to see if the data joint into two clusters, semipartial R-Squared (SPRSQ) that this equals the between cluster sum of squares divided by the corrected total sum of squares, we should use Method=average or centroid. Therefore, we obtain the following results: a. the decrease in the proportion of variance accounted for resulting from jointing the two clusters, semipartial R-Squired (SPRSQ), which does equate the between cluster sum of square divided by the corrected total sum of squares; b. the squared multiple correlation, R-Squared (RSQ), R^2 is the proportion of variance explained by the clusters. Testing data coordination using CCC (Cubic Cluster Criterion) has the important output: (1) Approximate Expected R-Squared, which is the approximation expectation value of R^2 under the uniform null hypothesis; (2) the value outputs of CCC. Sometimes, the values of CCC and approximation expected R^2 might missing values if the number of clusters are larger than one-fifth the number of observations. For PSEUDO option, it does mainly exhibits: a. Pseudo F (PSE) that is the pseudo F statistic measuring the separation among all the clusters at the current level. b. Pseudo t^2(PST2) in which pseudo t^2statistic testing the separation between the two clusters most recently involved. Form the above table, CL6 and CL7 are better for comprehensive estimators, and then CL5 and CL4. But, the CCC and pseudo F statistics are not appropriate for use with single linkage because of the method’s tendency to shop off tails of distributions. The pseudo t^2statistic would be applied by looking for bigger values and getting the number of clusters to be one larger than the level at which the bigger pseudo t^2 value is listed. Therefore, level 8 is largest, which suggests13 cluster, although we cannot look for where local peak CCC is.

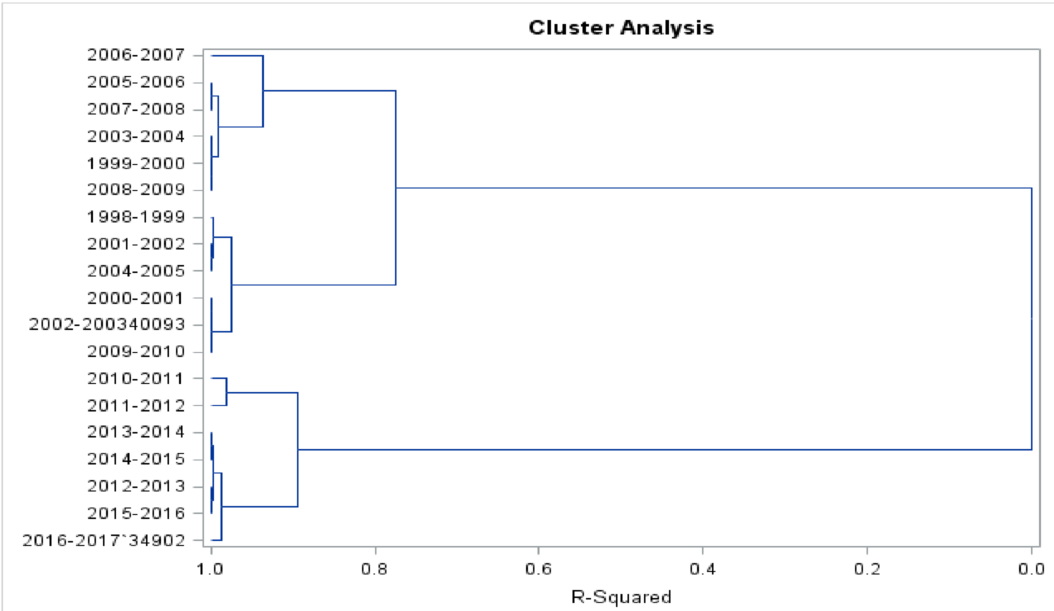


Figure 2. Tree diagram of clusters versus R-Squares.

For R^2, we know that it estimates if our models are appropriate for the data.it demines how much of the total variation in dependent variable. However, R^2 represents the differences between 2 clusters, but, at the beginning, the clustering process all entries are their own cluster, so, the $R^2$is 1. However, more clusters are combined with the decreasing value of R^2. Theoretically, the value of $R^2$is closed to be zero. The above plot, we see out that the group of clusters from 2002 to 2007 combines with the clusters from 2010 through 2016, R^2 at the final stage are closed to 0.

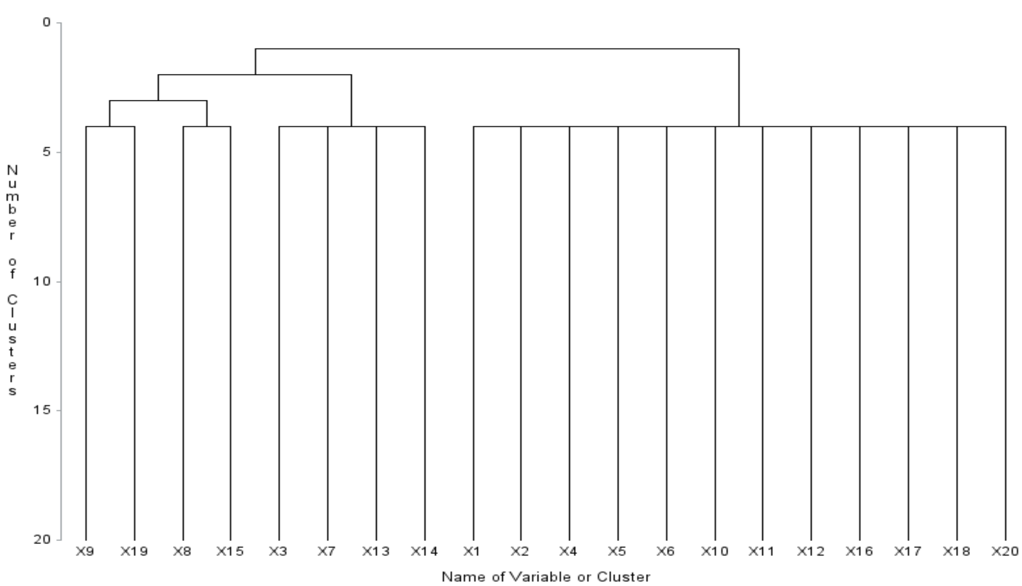


Figure 3. Dendrogram of single-linage Euclidean distance based on cluster solutions

Clearly, the above dendrogram displays that, at height about 4, a horizontal line segment joins two or two more variables X9 and X19, X8 and X15, and so on. In addition, we can find that the all of variables are contained. But, some variables do not change place without making shift of the clustering structures.

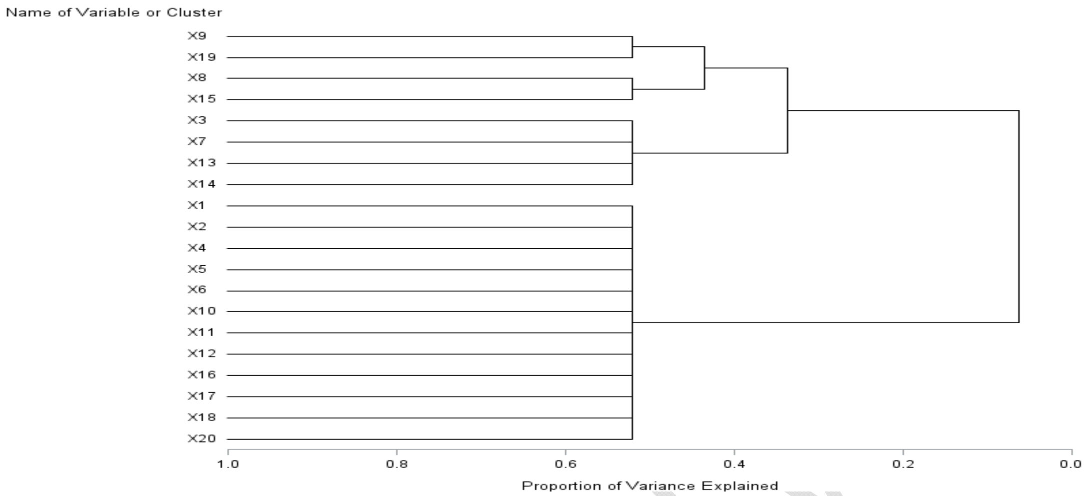


Figure 4. The plot of Oblique Centroid Component Clustering for Proportion of variance and numbers of variable cluster.

Looking from left to right in the above diagram, objects and clusters are progressively joins and becomes the single, all-encompassing clatter is created at the right for root. Clusters present in each level of the diagram. Every vertical line connects leaves and branches into progressively bigger clusters. The horizontal axis of the above dendrogram expresses the distance of the proportion of variance explained. For example, at approximately 0.5 of the proportion there are some variables or clusters (X1, X2, X4, X6, X10-X12, X16-X18, and X20) join (fusion) by splitting horizontal lines into two horizontal lines. The horizontal position of the split that are shown by the short vertical bar, which gives the distance between two or more clusters. Hence, combining X8, X15 with X9 and X19 suggests that there is significant for the data.

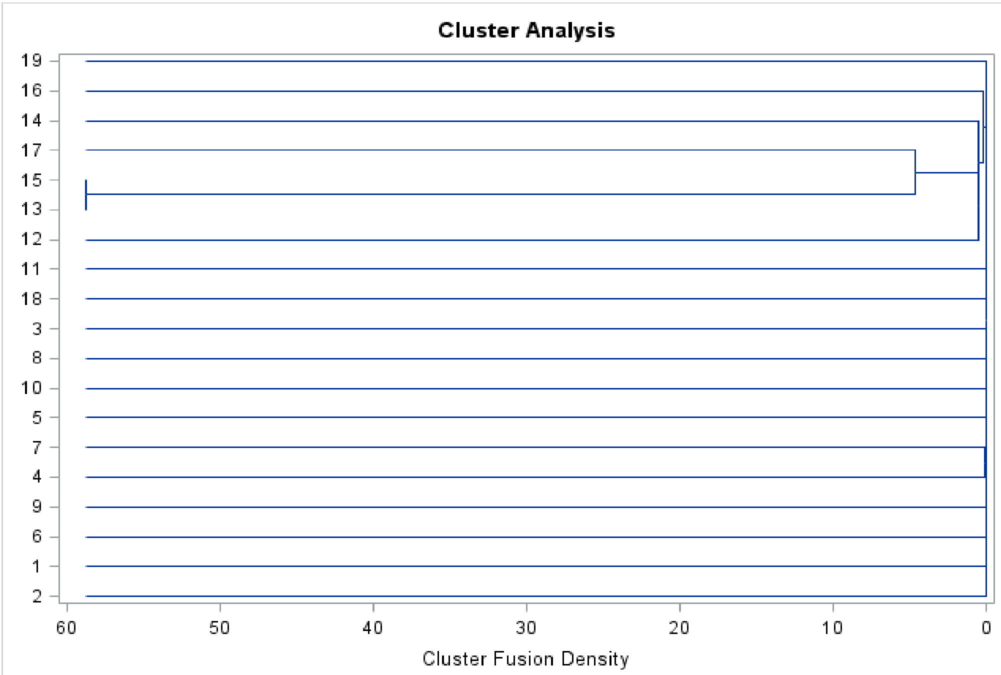


Figure 5. Cluster Analysis. It shows that two model clusters are formed.

For two-stage Density Linkage, the model clusters often need to merge before all the points in the tails have clustered. In the CLUSTER procedure the same the model clusters of two-stage density linkage as of ordinary density linkage: k-th nearest neighbor, uniform kernel and hybrid are displayed. For the first stage, generating disjoint model clusters, and then computing the single linkage algorithm ordinary with density linkage except for, two clusters are involved when one of two clusters has fewer members than the number specified by the option “MODE”. Finally, each point joints to one model cluster. For second stage, the model clusters are hierarchically formed by single linkage. When there are wide gaps between the clusters, the final number of cluster might be more one or if the smoothing parameter is small.

For the TREE procedure, the tree is plotted in the first stage, and then using HEIGHT in the proc Tree statement. In addition, we could make a single tree diagram for both stages, Hence, in the tree procedure, there are two model clusters set up in Figure 5: variables X13 and X15 have combined with X17.

4. CONCLUSION

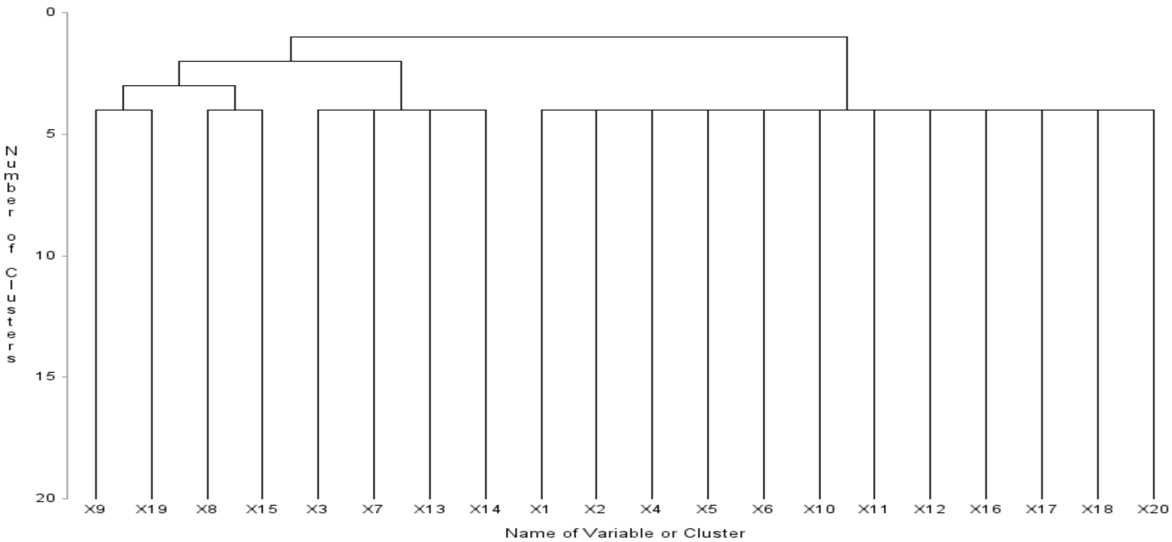
Cluster analysis reflects hierarchically clusters for the observations. There are many methods such as average linkage, the centroid, complete linkage, density, and maximum likelihood, and so on. These methods should be based on the usual agglomerative hierarchical clustering procedures. This paper is mainly to apply average, single linkages, two-stage density linkage, and Ward’s minimum-variance methods to analyze the mobility reasons. Since the variables in the data set do not have same as variance, I use some form of transformation to standardize the variables to mean zero and variance one. Moreover, I also apply proc ACECLUS procedure to transform the data in order to make within-cluster covariance matrix to be spherical. After using the above methods, the results show that, “Retired” reason in all the mobility reasons is significant, secondly, “Other housing reason” is more popular. But two reasons “Wanted better neighborhood or less crimes” and “Wanted cheaper housing”, that most people should make the decision of mobility, are not supported by this paper and the outputs of the data.

REFERENCES

1. Jason, S. Geographic mobility, 1990-1995. United States Bureau of the Census, 2000.
2. Jason, S., and Jeffrey J., K. Seasonality of moves and the duration and tenure of residence, 1996. United States Bureau of the Census, Population Division, 2002.
3. Long, L. E. Migration and Residential Mobility in The United States. Population of The United States in the 1980’s: Census Monograph Series. New York: Russell Sage Foundation 1988.
4. Frey, W. H. The Great American Migration Slowdown: Regional and Metropolitan Dimensions. Brooking Policy Brief. Washington, D.C.: Brookings Institute 2009.
5. United States Census Bureau. Reason for Move (Specific Categories): 2018.
6. Franklin, R. S. Migration of the young, single, and college education. U.S. Census Bureau 2003.
7. Jason, S. Migration by Race and Hispanic origin: 1995 to 2000. U.S. Census Bureau 2003.
8. Romesburg. H. Cluster analysis for researchers (2004). Lulu Press, Morrisville, NC 27560.
9. Brain, E. Cluster Analysis. Wiley Inc., Chichester, West Sussex, UK. 2011.
10. Daniel, S., and Hannah, S. Model-based on cluster analysis. Wiley Interdisciplinary Reviews, 2012: 4(4): 341-358.
11. Anderberg. M.R. Cluster analysis for application (1973). Academic Press, New York, NY 10019.
12. Ansary, and Rabiul. Emerging Patterns of Migration Streams in India: A State-level Analysis of the 2011 Census. Migration Letters; Luton 2018: 15(3): 347-360.
13. Flippo, S., Maritan, Amos, M., and Zoltan, N. Human Mobility in a Continuum Approach. PLOS One 2013: 8(3): e60069.
14. Yolanda, C.W. African Immigrants in Low-Wage Direct Health Care: Motivations, Job Satisfaction, and Occupational Mobility. Journal of immigrant and Minority Health 2017: 19 (3): 709-715.
15. Xie, L.M. Time Series Analysis and Prediction on Cancer Incidence Rates. J. Med. Discov. 2017: 2(3): jmd17030; DOI:10.24262/jmd.2.3.17030.
16. Xie, L.M. Analyzing and Forecasting HIV Data Using Hybrid Time Series Models. Asian Journal of Probability and Statistics 2018: 2(3): 1-12.
17. Nawrotzki, R. J., Riosmena, F., and Hunter, L.M. Do Rain fall Deficits U.S. -Border Migration from Rural Mexico? Evidence from the Mexican Census. Population Research and Policy Review 2013: 32(1): 129-158.
18. Duran, B.S. Cluster analysis: a survey (1974). Berlin, New York, NY 12022.

SUPPLEMENTARY FILES

S1 Oblique Centroid Component Clustering.



S2 The Matrix Table using proc distance procedure.

Obs	Y	_2016_2017_34902	_2015N2016	_2014N2015	_2013N2014	_2012N2013	_2011N2012	_2010N2011	_2009N2010	_2008N2009	_2007N2008	_2006N2007	_2005N2006	_2004N2005	_2003N2004	_2002N20034009 3	_2001N2002	_2000N2001	_1999N2000	_1998N1999
1	2016-2017'34902	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
2	2015-2016	0.08969	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
3	2014-2015	0.54045	0.45075	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	2013-2014	0.29607	0.20637	0.24438	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
5	2012-2013	0.38614	0.29645	0.15431	0.09007	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.
6	2011-2012	0.58378	0.49408	0.04333	0.28771	0.19763	0	.	.	.	.	.	.	.	.	.	.	.	.	.
7	2010-2011	0.05169	0.03801	0.48876	0.24438	0.33446	0.53209	0	.	.	.	.	.	.	.	.	.	.	.	.
8	2009-2010	0.9665	0.87681	0.42605	0.67043	0.58036	0.38272	0.91481	0	.	.	.	.	.	.	.	.	.	.	.
9	2008-2009	0.83728	0.74758	0.29683	0.54121	0.45113	0.2535	0.78559	0.12922	0	.	.	.	.	.	.	.	.	.	.
10	2007-2008	0.10072	0.01102	0.43973	0.19535	0.28543	0.48306	0.04903	0.86578	0.73656	0	.	.	.	.	.	.	.	.	.
11	2006-2007	1.43626	1.34656	0.89581	1.14019	1.05011	0.85248	1.38457	0.46976	0.59898	1.33554	0	.	.	.	.	.	.	.	.
12	2005-2006	1.87561	1.78592	1.33516	1.57954	1.48947	1.29183	1.82392	0.90911	1.03833	1.77489	0.43935	0	.	.	.	.	.	.	.
13	2004-2005	1.89499	1.8053	1.35454	1.59893	1.50885	1.31122	1.84331	0.92849	1.05772	1.79428	0.45874	0.01938	0	.	.	.	.	.	.
14	2003-2004	1.5556	1.4659	1.01515	1.25953	1.16945	0.97182	1.50391	0.5891	0.71832	1.45488	0.11934	0.32001	0.3394	0	.	.	.	.	.
15	2002-200340093	1.97291	1.88321	1.43246	1.67684	1.58676	1.38913	1.92122	1.00641	1.13563	1.87219	0.53665	0.0973	0.07791	0.41731	0	.	.	.	.
16	2001-2002	2.35981	2.27012	1.81936	2.06374	1.97367	1.77603	2.30812	1.39331	1.52253	2.2591	0.92355	0.4842	0.46482	0.80421	0.3869	0	.	.	.
17	2000-2001	1.55978	1.47008	1.01933	1.26371	1.17363	0.976	1.50809	0.59328	0.7225	1.45906	0.12352	0.31583	0.33522	0.00418	0.41313	0.80003	0	.	.
18	1999-2000	3.22522	3.13552	2.68477	2.92915	2.83907	2.64144	3.17353	2.25872	2.38794	3.1245	1.78896	1.3496	1.33022	1.66962	1.25231	0.8654	1.66544	0	.
19	1998-1999	2.93941	2.84971	2.39896	2.64334	2.55326	2.35563	2.88772	1.97291	2.10213	2.83869	1.50315	1.0638	1.04441	1.38381	0.9665	0.5796	1.37963	0.28581	0

S3. Plot of Eigenvectors for Raw Canonical Coefficients.

Eigenvectors (Raw Canonical Coefficients)																				
	Can1	Can2	Can3	Can4	Can5	Can6	Can7	Can8	Can9	Can10	Can11	Can12	Can13	Can14	Can15	Can16	Can17	Can18	Can19	Can20
X1	0.00955	0.00399	0.00052	0.00043	0.0002	-5E-05	0.00005	0.000037	0.00035	0.00028	0.00031	0.0011	3.9E-05	0.0011	0.0005	-3E-04	0.005	-0.005	-0.0085	-0.026
X2	-0.03175	-0.0077	-0.00022	-0.0004	0.00045	-0.0013	-0.00033	-0.00022	-0.00082	-0.0022	-0.0012	0.0001	0.0017	0.0038	0.0002	-0.004	-0.011	0.0085	0.0049	0.021
X3	-0.01506	-0.0071	0.00036	0.00088	-0.0002	-0.0018	6.8E-05	0.000321	-0.00317	-0.0009	0.00186	-0.0005	-0.00125	-0.0048	-0.002	0.0032	-0.005	0.0127	-0.0037	0.0233
X4	0.00085	-0.001	0.00065	-0.0003	-0.0002	6.7E-05	0.00014	-0.00047	0.00168	-0.0019	0.0003	0.0004	-0.00129	-0.0006	-0.003	0.0007	-0.006	0.0026	0.0083	0.0261
X5	-0.01701	0.00349	-0.00041	0.00048	0.0003	0.00184	-0.00028	0.001511	0.00236	-0.0005	-0.0002	-0.0007	0.00138	-0.0024	0.0015	-9E-04	-0.01	-0.004	0.0156	0.0285
X6	-0.03204	-0.0136	0.00044	-4E-06	-0.0005	-0.0003	0.00099	0.000691	-0.0011	-0.0012	0.00071	0.0047	-0.00182	0.0007	0.0086	0.0008	-0.018	0.0256	-0.0118	0.0093
X7	-0.01198	-0.0135	6.5E-05	-0.0001	2.1E-05	0.00158	9.7E-05	0.000676	-0.00126	0.00112	0.0002	0.0045	0.00073	0.0018	-0.003	-0.007	-0.006	0.0141	-0.0184	0.0239
X8	0.05847	-0.0395	0.00003	-0.0003	-7E-05	-0.0004	0.00078	0.000996	0.00075	0.00032	0.0002	0.0004	-4.2E-05	0.0215	0.0015	0.0013	0.0227	0.0242	-0.0405	0.0336
X9	0.01554	0.01116	0.0003	-0.0007	9.1E-05	-0.0002	-0.00096	-0.00056	-0.00087	0.00123	0.00039	-0.0022	8.7E-05	-0.0073	0.0059	-0.002	0.0099	-0.014	0.012	0.0436
X10	-0.00933	0.00261	-0.00038	0.00061	-0.0001	-0.0009	-0.00028	0.000487	-0.00203	0.00061	-0.0002	-0.003	-0.00151	0.0013	0.0014	-9E-04	-0.005	0.0028	0.0101	0.0274
X11	-0.00715	-0.0047	-0.00054	-0.0012	-1E-06	0.00023	0.00079	0.000082	3E-06	-0.0007	0.00103	-0.0004	-0.00046	-0.0011	-0.002	0.0006	-0.006	0.0069	0.0026	0.0247
X12	-0.03142	-0.0186	0.00012	-0.0002	0.00023	0.00184	-0.00102	-0.00023	-0.00304	0.00031	-0.0009	-0.0027	0.00067	0.0018	-0.004	0.006	-0.005	0.0236	-0.0044	0.0153
X13	0.0141	0.01225	0.00054	-5E-05	-0.0002	-0.0001	0.00115	0.001782	0.00144	-0.0002	-0.0002	-0.0031	-0.00079	-0.0013	8E-05	-0.001	0.0023	-0.013	0.0201	0.0384
X14	-0.02592	0.00765	0.00013	0.00105	-0.0002	0.00027	-0.00101	-0.00203	0.00228	-0.0016	0.00103	-0.0014	-0.00156	0.0058	0.0019	0.0037	0.0004	0.0049	0.0116	0.0233
X15	-0.00189	-0.0084	-0.00017	0.00018	0.00032	0.00015	0.00018	-0.00078	-0.00083	0.00138	-0.0007	0.0001	0.00029	-0.0057	-9E-04	0.0002	0.0026	0.0051	0.0011	0.0295
X16	0.02585	0.02061	-0.00026	0.00024	0.00011	-0.0004	-0.00049	-0.00033	0.00105	0.00021	-0.0008	0.0051	0.0013	-0.0067	-0.004	0.0015	0.013	-0.026	-0.018	0.0586
X17	0.02907	-0.0223	-5E-06	-1E-05	-2E-05	3.3E-05	0.00021	0.000113	-0.00108	0.00015	8.3E-05	0.0032	0.00026	0.0058	-5E-04	0.0004	0.0136	0.015	0.1088	0
X18	0.01218	-0.065	-9.2E-05	0.00052	7.7E-05	0.00011	-0.00079	-0.00117	0.00174	-0.0003	-0.0003	-0.0026	0.00021	-0.0192	0.0002	-0.001	0.0202	0.0608	-0.0053	-0.003
X19	-0.04782	-0.0139	3.5E-05	-0.0008	3.8E-05	-0.0015	0.00074	0.001556	0.00439	0.0033	-0.0005	0.0024	0.00127	0.01	-0.003	-0.001	-0.029	0.0201	0.0112	0.0042
X20	-0.01931	0.00867	0.00011	0.00054	-7E-05	0.00059	0.00055	-0.00215	-0.00051	0.00103	0.00003	-0.0019	0.00017	0.0015	0.0016	-3E-04	-0.012	-0.004	0.0042	0.0301