




Received: July 12, 2020
Accepted: December 31, 2021
Published Online: December 31, 2021

AJ ID: 2021.09.02.STAT.04
DOI: 10.17093/alphanumeric.970448
Research Article

A Proposal Method for Missing Value Analysis: Cluster Analysis Approach

Uğur Arcagök, Ph.D.* 

Res. Asst., Faculty of Economics and Administrative Sciences, Department of Business Administration, Mus Alparslan University,
u.arcagok@alparslan.edu.tr

Çiğdem Arıcıgil Çılan, Ph.D. 

Prof., School of Business, Department of Quantitative Methods, Istanbul University, Istanbul, Turkey, ccilan@istanbul.edu.tr

* Muş Alparslan Üniversitesi Külliyesi, Diyarbakır Yolu 7. km, 49250 Merkez/Muş, Türkiye

ABSTRACT

Imputing values to missing cases is a subject that is frequently met in the fields of Machine Learning and Data Mining, and that require the researchers to study it. It is known that many computer-based analysis algorithms operate under assumption that there is no missing case. The lack of sufficient search of missing case by the researchers is able to negatively affect the performance of analysis results. In this study, it was studied with a data set consisting of 52 variables in order to measure the performance of Corporate Sustainability of district municipalities in Istanbul. Little's MCAR was applied on 17 variables containing missing case, and it was determined that missing cases were MCAR, namely completely at random. And then Clustering Analysis was applied on 35 variables not containing missing case, and missing case imputations were made based on the clusters formed. It was observed that the cluster labels of municipalities, whose clustering analysis results obtained by data set with 35 variables that didn't contain missing case, and whose results obtained by the data set with 52 variables following imputation were the same, didn't change. The lack of change of cluster labels of municipalities indicates that the data set formed following imputation doesn't draw away from the main data, namely that the data structure doesn't get disrupted. Consequently, it can be said that clustering analysis is effective in terms of imputing more representative values in the imputation of missing case.

Keywords:

Missing Value Analysis, Little's MCAR Test, K-Nearest Neighbor Imputation Methods, Cluster Analysis



1. Introduction

The process with respect to missing data analysis initially began by Little and Rubin. The books of *Statistical Analysis with Missing Data* written by Little and Rubin, and *Multiple Imputation for Nonresponse in Surveys* written by Rubin in 1987 are deemed as the beginning of a revolutionary process about missing data. Because, by virtue of these books, the basis of computer programs to be used in missing data analysis was formed (Graham, 2009). Moreover, there are effective articles written about the problems arising from missing data before 1987.

The problem of missing data consists of two causes as being arising from the respondent (such as not answering the questions), or not arising from the respondent (such as data collection problems, or data entry mistakes). The researchers generally try to overcome the missing values in data by adding more new cases if they can be added to data set, or by various statistical approaches. If a variable or case in a data set is forming a great part of the missing cases in the data set, and if a great decrease in the number of missing cases in the data set is arising by the removal of that variable or case, then the removal of such variable or case from the data set is being a method used in the solution of the missing case problem. And sometimes under the control of researchers, and in definable circumstances, no method is used for the missing cases in the data set. And such missing data is called negligible missing data (Alpar, 2017).

Imputing values to missing cases is a subject that is frequently met in the fields of Machine Learning and Data Mining, and that require the researchers to study it. Because missing values are able to affect the quality of bias and controlled learning process, or the decomposition performance of classification algorithms. Moreover, many learning algorithms were designed under the assumption that there is no missing value in the data sets. For these reasons, the researchers should use a reliable method which may be able to preserve the distribution of data set while completing the missing values (Zhang et al., 2008).

Clustering analysis searches the cluster patterns in the data set by grouping the multi-variable cases in the form of clusters. The purpose is to find a group where the cases or objects in each cluster have similar characteristics, and where the clusters are different from each other. The purpose of the researchers is to find the significant natural homogenous groups in the data set. The difference of clustering analysis from classification analysis is lack of knowing beforehand the groups and number of groups in the clustering analysis. And in classification analyses, number of groups of the variable defined as dependent variable is definite, and it is coded before the analysis. In here, the purpose is to estimate to which known group the cases will be imputed (Rencher and Schimek 1997).

In clustering analysis, a kind of heuristic data analytic is formed by the use of uncontrolled learning algorithm. As it is known, while there are both class variable and features (other independent variables) in controlled learning, there is no class label, which will compare the performance of test and education set, in the uncontrolled learning. In other words, while clustering algorithm is maximizing the similarities of cases in the clusters, it also forms cluster label for the cases by minimizing the similarities among the clusters (Önder, 2020).

In this study, a data set, consisting of 52 variables collected for the sustainability of 39 district municipalities in Istanbul, was used. For some reasons, this data set contains missing cases. First the randomness of 17 variables containing missing case (approximately 4.93% of the data set) was searched. And then normalization method was used for clearing the variables from the effect of their units. In order to make imputation to variables containing missing case, imputation was made based on the most extensively used hierarchical and non-hierarchical (K-means) methods of clustering analysis, and their results were compared. As referred in many studies, based on the idea that it is required to use the averages as the imputation method in cases when the rate of missing case is below 5%, missing case imputation was made as per the municipalities' cluster label averages. As the result of that imputation, it was observed that the municipalities' cluster labels didn't change.

2. Literature Review

Acuna and Rodriguez (2004) compared the results of case deletion technique, the mean imputation, the median imputation, and the k-nearest neighbor imputation methods by the use of twelve data sets in order to search the effect of these methods on incorrect classification error rate for solving the missing case problem. The study was conducted considering the linear discriminant analysis (parametric), and k-nearest neighbor (non-parametric) classification methods. While case deletion showed the poorest performance, it was observed that k-nearest neighbor imputation showed the best performance when the number of missing case increased. Moreover, it was observed that these results were compatible with the results obtained by Dixon (1979).

Aljuaid and Sasi (2016) used the 5 imputation techniques such as Mean/Mode, K-Nearest Neighbor (KNN), Hot-Deck (HD), Expectation Maximization (EM), and C5.0 in order to impute to missing data, formed artificially from different data sets of different dimensions, and to compare their results. The comparison of the performance of these techniques is based on the data imputed, and on the correct classification of the original data. They advocated that the data type may be numeric, categorical or combined, and that the selection of suitable imputation method is based on the data types, missing data mechanisms, patterns, and methods. As the result of their studies, they concluded that the HD imputation for missing data may increase the accuracy of estimation to a statistically significant level in a large data set, and all features of the data set was not used for the C5.0 classification, but that it was still providing a good classification accuracy on different data types. In addition, they concluded that both the EM and KNN imputations may be effective for missing data imputation. They concluded that more time was being consumed when worked with KNN on large data sets, that the EM method was showing better performance on quantitative variables, and that the mean/mode imputation decreased the relationship with other variables as well as disrupting the normality assumptions, and that it may be used in case of presence of missing data less than 5%.

Chan and Dunn (1972) examined by the Monte Carlo methods the correct classification probabilities of a few methods extensively used in missing case analysis. They advocated that average imputation, and principal components

methods are generally superior compared to other methods. Dixon (1979) introduced the k-nearest neighbor (KNN) imputation technique for the solution of missing case in controlled classification. Tresp, Neuneier and Ahmad (1995) addressed the missing case problem in the context of controlled learning for artificial neural networks. And they showed the accuracy of their theory by the use of clustering and regressing methods. Bello (1995) used the regression method as the imputation technique of missing cases. In his study, he compared the results of imputation methods based on both dependent and independent variables (type 1), and based on only explanatory variables (type 2). The results of Monte Carlo indicated that the imputation values performed by the procedure type 1 may give the impression of high accuracy by creating spurious impression especially as the rate of missing data increases, but that in the estimations made by type 2, residual mean square error was being overestimated.

3. Research Methods

3.1. Missing Value Analysis

Before finding a solution for the missing data problem, it is required to search the randomness level in missing data. The level of randomness in missing data is searched in three manners as being MAR (Missing at Random), MCAR (Missing Completely at Random), and MNAR (Missing Not At Random). The term MCAR (Missing Completely at Random) expresses that being missing is not dependent on a variable, or on any variable in the data set. It indicates that the data set gathered randomly, and that the missing data is not dependent on another variable in the data set, and that it is random (Rubin, 1976). The term MAR (Missing at Random) arises when being missing is not random, but when being missing may be completely explained by the variables having full information. It is not possible to verify it statistically. By the term MNAR, the data not being MAR and MCAR are known as not responded, or not marked data (Scheffer, 2002).

Hair (2009) classified the data sets containing missing case under three groups as being ones below 10%, ones between 10% and 20%, and ones above 20%. He told that if the data set is containing missing cases below 10% then it is ignorable, and that any imputation method is applicable. If the data set is containing missing cases between 10% and 20%, then the missing case is visible. If the data set is MCAR, then Hot Deck imputation method should be applied, and if it is MAR, then model based missing case process should be applied. If the data set is containing missing cases above 20%, and if imputation is wanted to be made, then regression should be applied for MCAR cases, and model-based imputation methods should be applied for MAR cases (Hair, 2009).

3.2. Cluster Analysis

3.2.1. Hierarchical Clustering Method

It is the method used by the researchers when the number of clusters is not known beforehand, or when estimation of the number of clusters is not possible. This method is applied in two manners as being clusters formed by a series of successive combinations (Agglomerative), or clusters formed by a series of successive divisions (Divise).

In the Divise hierarchical method, all case values are deemed as a single cluster in the beginning. And then two new sub-clusters are formed consisting of cases that are away from each other. And then these sub-clusters are divided to dissimilar sub-clusters. The operation continues until each object forms a cluster, namely until having numerous sub-cluster objects. And in Agglomerative hierarchical method, each case value is deemed as a cluster against the Divise method. Then the two clusters combine as per their similar features, or the cases gather and form a new cluster. This operation continues until all the cases or objects form a cluster. The algorithms used in hierarchical clustering analysis are divided to five: Single Linkage Method, Complete Linkage Method, Average Linkage Method, Ward's Minimum Variance, and Centroid Method.

3.2.2. K-Means Method

The purpose of the analysis is to classify the cases as per the pre-estimated or pre-known number of clusters based on the prior knowledge and experience of the researcher. In the k-means analysis, the number of clusters is determined at least as 2, and at most as not to exceed the number of cases. In here, the symbol K represents the number of clusters. The analysis method also known as non-hierarchical clustering analysis, and k-means consists of three stages (Orhunbilge, 2010):

1. Division of cases to the required number of clusters.
2. Imputation of cases to the closest cluster as per the center (average) of cluster.
3. Stopping the operation when the imputation of all the cases ends. Otherwise, in case of having a new imputation rule, it is returned to stage 2, and the cases are imputed as per the newly determined rule, and the operation is completed.

Generally, three different clustering algorithms are referred in the sources: Sequential Threshold Method, Parallel Threshold Method, and Optimization Method. There are also other algorithms except these algorithms (Alpar, 2017).

Three different methods are used in the determination of most suitable number of clusters used in the hierarchical and non-hierarchical (K-mean) methods of clustering analysis. The Elbow Method, Silhouette Method, and Gap Statistic Method (Özdemir, 2020).

3.3. Normalization

Multi-variable statistical techniques, in which distances are used just like the clustering analysis, are sensitive against unit differences. Because the variable with high variability is able to effect the other variables. For this reason, before starting the clustering analysis, it is required to normalize the data set. As the result of the normalization operation, it is ensured to draw the data to a specific range (0, +1, or -1, +1). Thus, the comparison and interpretation of data being under different conditions are ensured (Vinh, Epps, and Bailey, 2010). Before starting the clustering analysis, the variables in the data set were normalized, and were drawn into the range of 0-1. The reason of the use of the below normalization method is to eliminate the problem of outliers among the unit values of the variables.

$$\text{Normalization} = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)}$$

3.4. Research Design

In the study, regarding the data set on which Missing Case Analysis would be made, the data of 52 variables in order to measure the corporate sustainability (economic, social, environment dimensions) of 39 district municipalities of Istanbul, which is one of the most significant metropolises of the world, was obtained from the activity reports published on the official internet pages of the municipalities. 17 of these variables contain missing case.

First, the randomness of missing cases was searched. In the imputation of value instead of missing cases which were determined to be random, Clustering Analysis was applied by normalizing 35 variables of 39 districts which didn't contain missing case. For the variable of the district containing missing case, the values of average of cluster (arithmetic mean for quantitative variables, and mode for qualitative variables) covering the district was imputed.

Following imputation, in the Clustering Analysis re-applied on all 52 variables, it was indicated that the clusters didn't change, namely that the Missing Case Imputation Method suggested in this study didn't disrupt the structure of data.

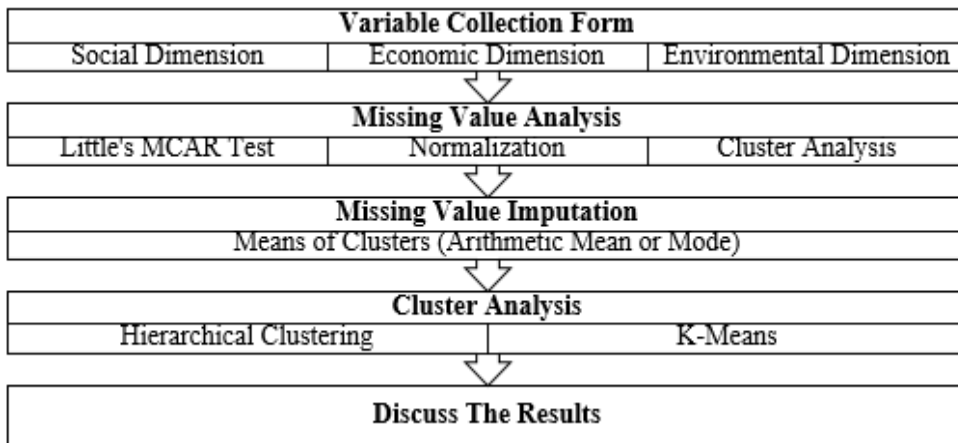


Figure 1. Research Design

4. Results

4.1. Results of Little's MCAR Test

When the data set was examined, it was observed that there were 100 missing cases in total in 17 variables. In other words, the missing cases formed the 4.93% of the data set consisting of 52 variables, 39 cases, and 2028 units. In order to search the randomness of missing cases, Little's MCAR test was conducted, and it was revealed that being missing was not dependent on a variable, or on any variable in the data set ($p \text{ sig} > 0.05$).

Univariate Statistics						
	N	Mean	Std. Deviation	Missing		P-value
				Count	Percent	
S31_1	27	1,44	0,85	12	30,80	
S31_2	36	2,81	2,49	3	7,70	
S31_3	35	1,63	1,35	4	10,30	
S31_4	37	3,11	1,91	2	5,10	
S31_5	36	2,33	1,27	3	7,70	
S31_6	38	4,05	3,39	1	2,60	
E4	34	0,02	0,04	5	12,80	
E6	34	62,11	36,61	5	12,80	0,582
E7	37	23,31	41,22	2	5,10	
E8	32	14,05	22,92	7	17,90	
E9	28	32,81	18,43	11	28,20	
E10	36	31,51	59,45	3	7,70	
E11	33	76,90	101,05	6	15,40	
E15	29	284,11	248,64	10	25,60	
E19	34	26,34	28,41	5	12,80	
ENV5	24	66,29	17,64	15	38,50	
ENV6	33	54,85	44,80	6	15,40	
Total	N	Percent				
Missing	100	4.93				
Completed	1928	95.07				
Case	2028	100				

Table 1. Little's MCAR Test Result

4.2. Clustering Analysis

4.2.1. Hierarchical Clustering Analysis

First Hierarchical Clustering Analysis was applied on the data set consisting of 35 variables not containing missing cases, and clusters that were similar in terms of sustainability were formed. And then central tendency measures (arithmetic mean value for quantitative variables, and mode value for qualitative variables) of municipalities included in the relevant cluster were imputed to the missing cases. The normality assumption, having a significant place in multi-variable statistical analyses, is ensured by the distance values in clustering analysis. In the study, Euclidean distance matrices were used in the measurement of distances.

According to the Euclidean distance, being the most extensive distance measurement used in clustering analysis, the 39 municipalities of Istanbul were divided to clusters as per 4 different algorithms of Agglomerative hierarchical method analysis. As shown below, when the algorithm with the highest agglomeration coefficient was compared with the other three algorithms, it was obtained by the use of Ward's minimum variance method (0.969).

Average	Single	Complete	Ward
0.9056290	0.7955809	0.9525100	0.9695144

Table 2. Integration coefficients of Hierarchical Clustering Analysis.

Number of clusters was determined by the use of Elbow and Silhouette methods. As seen below, it was decided that the most suitable number of clusters was required to be 4.

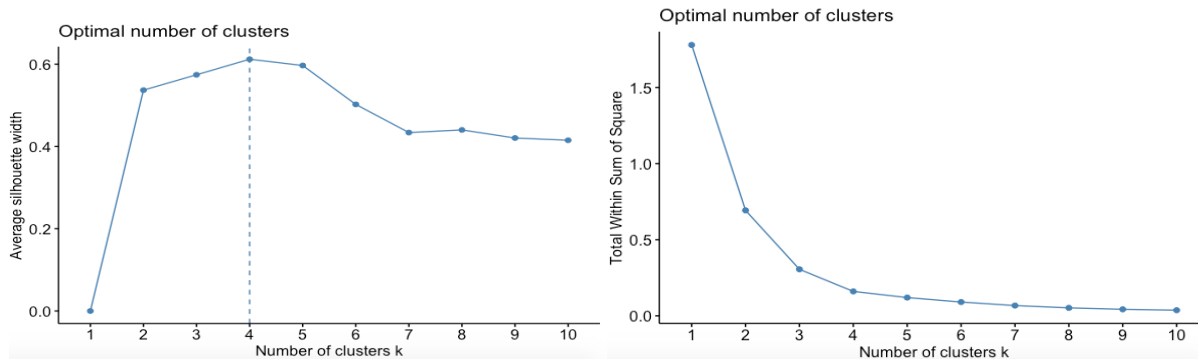


Figure 2. Results of Elbow and Silhouette Methods

Dendrogram Of Agnes

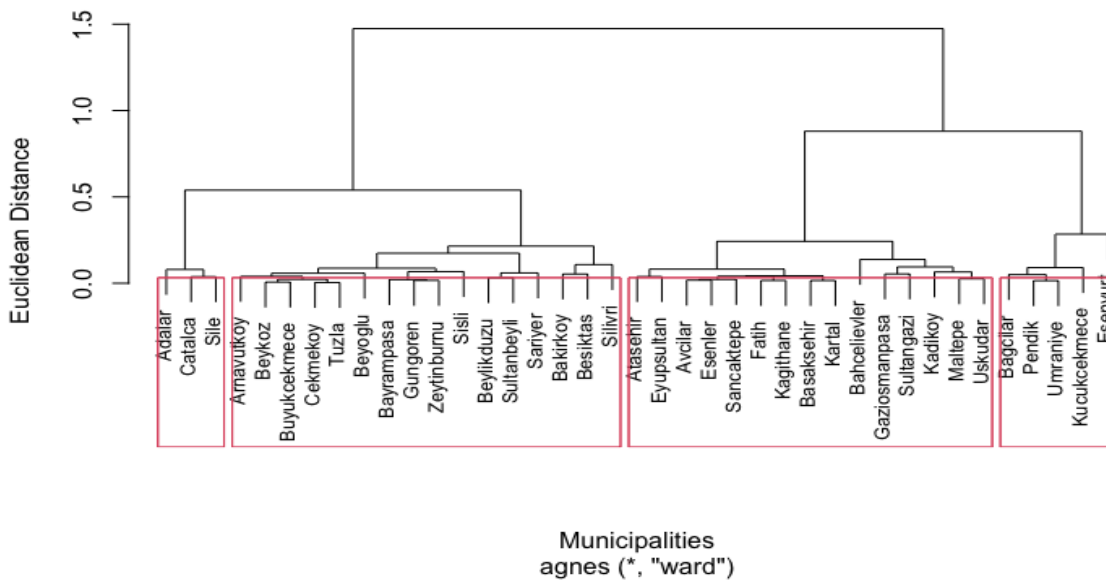


Figure 3. Result of Hierarchical Cluster Analysis

4.2.2. K-Means

The purpose of the analysis is to classify the cases as per the pre-estimated or pre-known number of clusters based on the prior knowledge and experience of the researcher. In the k-means analysis, the number of clusters is determined at least as 2, and at most as not to exceed the number of cases. In here, the symbol K represents the number of clusters. When the number of clusters is determined as four (K=4), it is observed that the municipalities are being divided to exactly the same clusters considering the results of hierarchical clustering analysis, and non-hierarchical (K-means) clustering analysis.

Cluster 1 (n=3)	Cluster 2 (n=16)	Cluster 3 (n=15)	Cluster 4 (n=5)
Adalar	Arnavutköy	Ataşehir	Bağcılar
Çatalca	Bakırköy	Avcılar	Esenyurt
Şile	Bayrampaşa	Bahçelievler	Küçükçekmece
	Beşiktaş	Başakşehir	Pendik
	Beykoz	Esenler	Ümraniye
	Beylikdüzü	Eyüpsultan	
	Beyoğlu	Fatih	
	Büyükkçekmece	Gaziosmanpaşa	
	Çekmeköy	Kadıköy	
	Güngören	Kağıthane	
	Sarıyer	Kartal	
	Silivri	Maltepe	
	Şişli	Sancaktepe	
	Sultanbeyli	Sultangazi	
	Tuzla	Üsküdar	
	Zeytinburnu		

Table 3. Cluster membership of the municipalities

4.3. Missing Value Imputation

In order to search the randomness of missing cases, Little's MCAR test was conducted, and it was determined that being missing was not dependent on a variable, or on any variable in the data set. By the 35 variables not containing missing cases, it was concluded that the 39 municipalities of Istanbul were being divided to 4 clusters. And then central tendency measures (arithmetic mean value for quantitative variables, and mode value for qualitative variables) of municipalities included in the relevant cluster were imputed to the missing cases. Then, in order to check whether the missing case imputation operation was successful or not, in other words in order to check whether the cluster patterns covering the municipalities change or not, the clustering analysis was repeated with 52 variables not containing missing case. As seen in the following figures, cluster labels and number of clusters, in which the municipalities were represented, didn't change as per the results of hierarchical clustering, and k-means.

Average	Single	Complete	Ward
0.9056193	0.7955541	0.9525057	0.9695115

Table 4. Integration coefficients of Hierarchical Clustering Analysis.

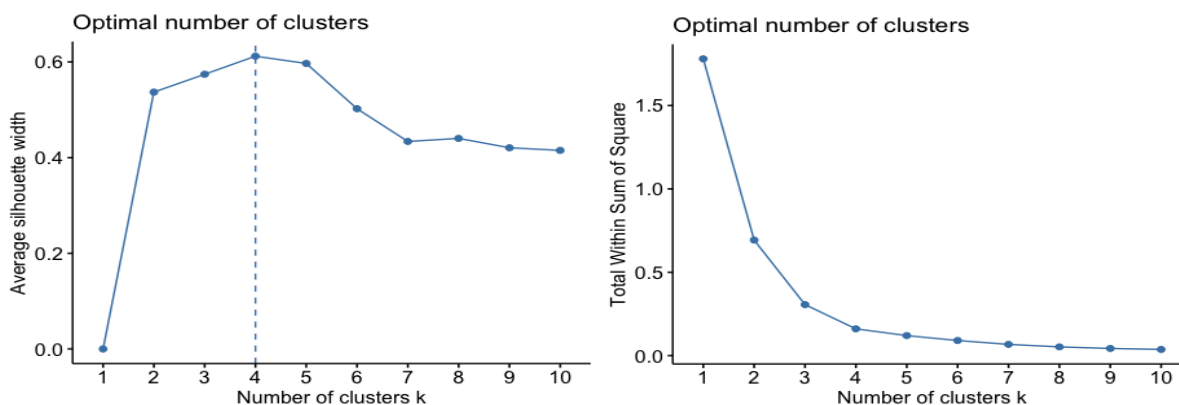


Figure 4. Results of Elbow and Silhouette Methods

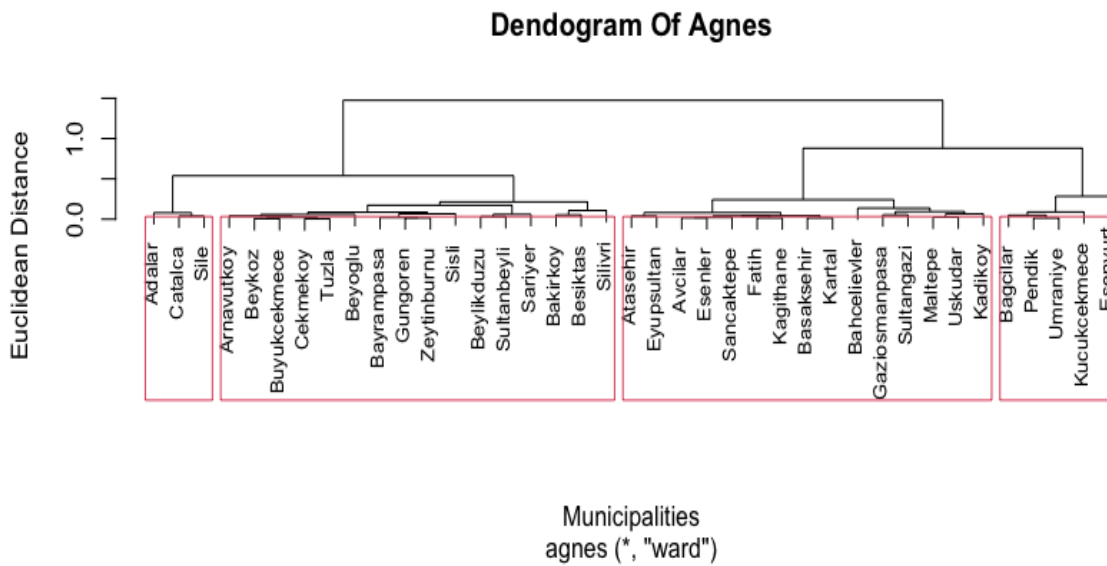


Figure 5. Result of Hierarchical Cluster Analysis

Cluster 1 (n=3)	Cluster 2 (n=16)	Cluster 3 (n=15)	Cluster 4 (n=5)
Adalar	Arnautköy	Ataşehir	Bağcılar
Çatalca	Bakırköy	Avcılar	Esenyurt
Şile	Bayrampaşa	Bahçelievler	Küçükçekmece
	Beşiktaş	Başakşehir	Pendik
	Beykoz	Esenler	Ümraniye
	Beylikdüzü	Eyüpsultan	
	Beyoğlu	Fatih	
	Büyükçekmece	Gaziosmanpaşa	
	Çekmeköy	Kadıköy	
	Güngören	Kağıthane	
	Sarıyer	Kartal	
	Silivri	Maltepe	
	Şişli	Sancaktepe	
	Sultanbeyli	Sultangazi	
	Tuzla	Üsküdar	
	Zeytinburnu		

Table 5. Cluster membership of the municipalities

5. Discussion and Conclusion

It was studied with a data set consisting of 52 variables in order to measure the performance of Corporate Sustainability of district municipalities in Istanbul. Little's MCAR was applied on 17 variables containing missing case, and it was determined that missing cases were MCAR, namely missing completely at random. Clustering Analysis was applied on 35 variables not containing missing case, and missing case imputations were made based on the clusters formed. As the rate of missing cases in data set was below 5%, and as it was MCAR, mean and mode values were imputed for missing cases as in the studies of Hair (2009), Acuna & Rodriguez (2004), Aljuaid and Sasi (2016), and Dixon (1979). In other words, as per the district municipality and indicator that the missing cases in the data set were subject to, they were imputed by taking the average of the district municipalities in their relevant cluster.

It was observed that the cluster labels of municipalities, whose clustering analysis results obtained by data set with 35 variables that didn't contain missing case, and

whose results obtained by the data set with 52 variables following imputation were the same, didn't change. The lack of change of cluster labels of municipalities indicates that the data set formed following imputation doesn't draw away from the main data, namely that the data structure doesn't get disrupted. Consequently, it can be said that clustering analysis is effective in terms of imputing more representative values in the imputation of missing case.

References

- Acuna, E., & Rodriguez, C. (2004). The treatment of missing values and its effect on classifier accuracy. In *Classification, clustering, and data mining applications* (pp. 639-647). Springer, Berlin, Heidelberg.
- Aljuaid, T., & Sasi, S. (2016, August). Proper imputation techniques for missing values in data sets. In *2016 International Conference on Data Science and Engineering (ICDSE)* (pp. 1-5). IEEE.
- Alpar, C. (2017). Uygulamalı çok değişkenli istatistiksel yöntemler.
- Bello, A. L. (1995). Imputation techniques in regression analysis: looking closely at their implementation. *Computational statistics & data analysis*, 20(1), 45-57.
- Chan, L. S., & Dunn, O. J. (1972). The treatment of missing values in discriminant analysis—I. The sampling experiment. *Journal of the American Statistical Association*, 67(338), 473-477.
- Dixon, J. K. (1979). Pattern recognition with partly missing data. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(10), 617-621.
- Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual review of psychology*, 60, 549-576. <https://doi.org/10.1146/annurev.psych.58.110405.085530>
- Hair, J. F. (2009). *Multivariate data analysis*.
- Önder, E. (2020). *Sağlıkta Gelişmekte Olan Teknolojiler, Yapay Zekâ & R İle Makine Öğrenimi Uygulamaları*. Bursa: Dora Yayıncılık.
- Orhunbilge, N. (1999). *Zaman Serisi Analizi Tahmin ve Fiyat İndeksleri*. İstanbul: Tunç Matbaacılık.
- Özdemir, M. (2020). *R ile Programlama ve Makine Öğrenmesi*.
- Rencher, A. C., & Schimek, M. G. (1997). Methods of multivariate analysis. *Computational Statistics*, 12(4), 422-422.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581-592.
- Scheffer, J. (2002). *Dealing with missing data*.
- Tresp, V., Neuneier, R., & Ahmad, S. (1995). Efficient methods for dealing with missing data in supervised learning. In *Advances in neural information processing systems* (pp. 689-696). Morgan Kaufmann Publishers.
- Vinh, N. X., Epps, J., & Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research*, 11, 2837-2854.
- Zhang S., Zhang J., Zhu X., Qin Y., Zhang C. (2008) Missing Value Imputation Based on Data Clustering. In: Gavrilova M.L., Tan C.J.K. (eds) *Transactions on Computational Science I. Lecture Notes in Computer Science*, vol 4750. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-79299-4_7

