EKONOMETRIA ECONOMETRICS 4(54) • 2016

ISSN 1507-3866
e-ISSN 2449-9994

Marcin Gąsior, Łukasz Skowron
Lublin University of Technology
e-mails: m.gasior@pollub.pl; lukasz.m.skowron@gmail.com

METHODS FOR IMPUTATION OF MISSING VALUES AND THEIR INFLUENCE ON THE RESULTS OF SEGMENTATION RESEARCH*

METODY UZUPEŁNIANIA BRAKÓW DANYCH I ICH WPŁYW NA WYNIKI BADAŃ SEGMENTACYJNYCH

DOI: 10.15611/ekt.2016.4.04
JEL Classification: C38

Summary: The lack of answers is a common problem in all types of research, especially in the field of social sciences. Hence a number of solutions were developed, including the analysis of complete cases or imputations that supplement the missing value with a value calculated according to different algorithms. This paper evaluates the influence of the adopted method for the supplementation of missing answers regarding the result of segmentation conducted with the use of cluster analysis. In order to achieve this we used a set of data from an actual consumer research in which the cases with missing values were deleted or supplemented with the use of various methods. Cluster analyses were then performed on those sets of data, both with the assumption of ordinal and ratio level of measurement, and then the grouping quality, as expressed by different indicators, was evaluated. This research proved the advantage of imputation over the analysis of complete cases, it also proved the validity of using more complex approaches than the simple supplementation with an average or median value.

Keywords: missing values, cluster analysis, k-means algorithm, k-medoids algorithm.

Streszczenie: Braki odpowiedzi są częstym problemem we wszelkiego rodzaju badaniach, zwłaszcza z obszaru nauk społecznych. W konsekwencji opracowane zostało wiele sposobów rozwiązania tego problemu, uwzględniających między innymi analizę przypadków kompletnych czy imputacje – polegające na przypisaniu w miejsce braku wartości wyznaczonej przy wykorzystaniu różnych algorytmów. W niniejszym artykule dokonano oceny wpływu przyjętej metody zastępowania braków odpowiedzi na wyniki badań segmentacyjnych, prowadzonych przy wykorzystaniu analizy skupień. W tym celu wykorzystano zbiór danych z rzeczywistego badania konsumenckiego, w którym braki odpowiedzi zostały usunięte bądź zastąpione przy wykorzystaniu różnych, możliwych podejść. Na tak przygotowanych zestawach przypadków przeprowadzono analizy skupień, zarówno przy założeniu porządkowego, jak i przedziałowego poziomu pomiaru, następnie zaś porównano jakość grupowania, wyra-

* The presented research was a part of a research project financed by the Polish National Science Centre, decision DEC-2011/03/D/HS4/04311.

żoną wybranymi wskaźnikami. Tak przeprowadzone badanie wskazało na przewagę imputacji nad analizą przypadków kompletnych, dowiodło także zasadności stosowania podejść bardziej złożonych niż zastępowanie braków średnią lub medianą.

Słowa kluczowe: *missing values*, analiza skupień, metoda *k*-średnich, metoda *k*-medoidów.

1. Introduction

Missing data, understood as missing values for some or all variables for measurement, are a typical and common problem in the majority of social research, especially when they are connected with acquiring respondent's opinion with the use of intermediate research tools, such as survey questionnaires. This shortage can assume two different characters – manifested either by the refusal to take part in the research by the individual qualified for participation in the panel or the unavailability of such a person (*unit non-response*), or the lack of response of the individual to one or several questions, with correct, valid replies to the remaining questions (*item non--response*).

When discussing the lack of data within one of the variables, we will first describe the completely random gaps – MCAR (*missing completely at random*) – which occur when the absence of answer has no relation to the value of the variable in which it occurs, nor to any other observed or non-observed variables [Rubin 1976], and the subset of complete cases is a random sample of the original set of cases.

The opposite are the MNAR (*missing not at random*) gaps, in which the fact of the occurrence of absence may depend on variables that remain outside the measurement model, and thus fall outside observation, or on the observed value itself. The lack of non-random data may lead to loss of information, and – as their predictors are not observed – the diagnosing of their lack of randomness may be inhibited.

The last possible case are the MAR (*missing at random*) gaps, in which the probability of occurrence of such a lack depends on other variables observed during the respective measurements. This makes those variables suitable for the estimation of missing values.

There are several procedures for missing data. First, it is possible to reject all cases that contain incomplete data, that is the analysis of complete cases only. As far as this is simple and convenient, such approach leads to the partial loss of information and also the possible biasing of the remaining part of sample [Schafer and Graham 2002].

The alternative to the deletion of cases is data imputation, that is the supplementation of missing values with values determined with the use of a certain algorithm. The simplest approaches include the methods of supplementation with an arithmetic average (or median) calculated from the available values of the variable in question, the supplementation of a value with one determined with regression, or for example the assignment of a value from a complete case that is similar when the remaining variables are considered, to the case in question.

A more complex approach is the use of methods based on maximum probability, where the missing values are estimated with use of an EM algorithm or multiple imputation, that is the repeated process of the estimation of values of missing inputs several times, that generates not a single, but a set of datasets, each of which is then subjected to further analysis. Linking the subsequent results allows the production of a general estimation, the evaluation of potential sample bias resulting from the conducted imputation, and the estimation of the values of standard errors [Cole 2008].

It is worth stressing that the majority of algorithms require the missing data to be at least missing at random (MAR), an assumption that is hard to prove in the case of the majority of empirical datasets, without reaching the individuals who failed to answer in order to repeat the questionnaire [Schafer and Graham 2002]. At the same time it is worth stressing, after Schafer [Schafer 1997], that this type of methods is characterized by good practical results, even if the assumption of the random character of gaps cannot be verified or is doubtful.

The question may thus arise if, and if so to what degree, the procedure adopted for dealing with missing values influence the results of the subsequent analyses conducted on the basis of the results prepared in such a way. The aim of the present research was to evaluate the influence of applying different methods of data imputation (as described in subject literature) on the quality of the segmentation research conducted using cluster analysis. This evaluation enables us to formulate conclusions pertaining to the optimal procedure for data gaps in cases of such research.

2. Research assumptions

The research procedure applied for the evaluation of the influence of imputation methods involved; in first place, the generation of datasets with missing values estimated using several common methods. Then, for each of those sets, cluster analyses were conducted with use of k-mean and k-medoids methods, in both cases with 2 to 10 clusters. Finally, a range of indicators that were displaying the achieved quality of clustering were calculated for all analyses conducted in this way.

The present analysis is based on empirical material gathered during completion of a research project focused around the search for links between employer motivation and satisfaction, and the satisfaction and loyalty of customers. It utilizes a set of 8 variables expressing the different possible reasons for visiting a shopping mall that included the replies of 1375 respondents.

All the discussed variables were measured in a ten-grade semantic scale, within which the low values expressed the low importance of the respective reason for visiting, and the high ones – its large significance. It should be also pointed out that the aforementioned variables were not strongly [Sambandam 2003] correlated.

At this stage we also had to make an assumption about the level of the measurement conducted. Adopting a conservative approach, the scale of that type

should be regarded as ordinal [Marcus-Roberts, Roberts 1987; Stevens 1946], together with all the implications related to the possible methods of its analysis. Still, both literature and the research praxis commonly accept the mitigation of that criterion, indicating that even though the level of measurement was in fact ordinal, the numerical character of scale allows for its treatment as an interval or ratio level, and thus the application of a broader set of analytical methods [King et al. 2001; Labovitz 1967]. Taking the above into consideration, the analysis will be conducted in two paths, first assuming that the level of measurement was ordinal, and then that it was ratio.

The characteristic of the used data is presented in Table 1.

Table 1. Parameters of variables used for analysis

Variable	Complete	Missing data		Median	Mean	Standard deviation
		n	%			
Variable 1	1356	19	1.38	8	7.01	2.36
Variable 2	1317	58	4.22	6	5.63	2.73
Variable 3	1228	147	10.69	2	3.16	2.61
Variable 4	1291	84	6.11	4	4.49	2.79
Variable 5	1298	77	5.60	5	4.89	2.87
Variable 6	1259	116	8.44	3	3.68	2.69
Variable 7	1261	114	8.29	5	5.21	2.99
Variable 8	1230	145	10.55	2	2.96	2.56

Source: author's research.

For the purpose of the conducted analysis, seven datasets were prepared:

1. Set of complete cases.

2. Set with missing values filled with random values from the measurement scale, a solution deemed acceptable by many sources, as at the cost of worsening measurement precision of one variable (the one filled with random values) the quality of the remaining, complete ones, which in case of no imputation would be excluded from analysis together with the whole case, increases.

3. Set with missing values filled with mean values of the respective variable.

4. Set with missing values filled with median of the respective variable.

5. Set with missing values filled with PMM (*predictive mean matching*) [Little 1988], a method that requires a ratio scale, while still allowing for its discretion and lack of normality.

6. Set with values filled with use of classification and regression tree (C&RT/CART).

7. Dataset with values inputted with use of the Expectation-Maximization (EM) algorithm – in the latter two cases the predictors for the missing values were all the variables used for the analysis.

Taking the reservation in respect to the measurement level in consideration, two methods for cluster analysis were used. Assuming the ratio level, the cluster analysis was performed with the k-means method with the Euclidean measure of distance. Assuming the ordinal level, that excluded the calculation of mean values and using the Euclidean distance – the k-medoids method (PAM) with GDM (*General Distance Measure*) distance measure were used [Jajuga, Walesiak, Bąk 2003; Walesiak 2006].

The last aspect of the proposed test procedure that ought to be described is the choice of indicators reflecting the achieved quality of the clustering. The literature suggests a very wide set of possible criteria and evaluation methods – a comprehensive review is presented by [Migdał-Najman 2011] and [Charrad et al. 2014]. The present research used:

1. The silhouette index [Rousseeuw 1987], in a range of –1 to 1, where values closer to one represent a small distance between elements of cluster in relation to the closest elements of separated clusters, thus the higher quality of classification.

2. Hubert-Levin's C index [Hubert, Levin 1976], in a range of 0 to 1, where values closer to nil represent better classification.

3. (Assuming the ratio level of measurement) the Caliski-Harabasz index [Caliński, Harabasz 1974], expressing the relation of variances between clusters to the variances within clusters, in this case the higher the ratio the better quality of classification.

It is worth stressing that the first two indexes can be applied also in the case of an ordinal type of measurement [Walesiak, Dudek 2006] which enables the preservation of comparability of quality of clustering with the use of k-means and k-medoids methods.

All the calculations were performed using the *clusterSim* software suite, an add-on within the R software environment.

3. Imputation and the k-means method

The first method for cluster analysis that was subjected to evaluation was the k-means method, used with the Euclidean measure of distance. The indexes used to represent the quality of the resulting clustering were the silhouette index, the C Hubert-Levin index, and – because of the ratio character of the measurement – the Celiński-Harabasz index.

The analysis of values of the first of these indicators (Figure 1) shows first that the imputation of random values visibly lowers the quality of clustering. The substitution of gaps with median or mean average of the test sample gives much the same results, which is probably a consequence of the fact that both values were similar. What is important is the fact that even if such an imputation is better than inputting random values, the quality of clustering is still visibly lower than when analyzing complete cases. The remaining algorithms seem to improve the quality of clustering to some extent, with a visible advantage of EM algorithm in cases of a lower number of clusters, and of CART and PMM algorithms with a growing number of clusters.

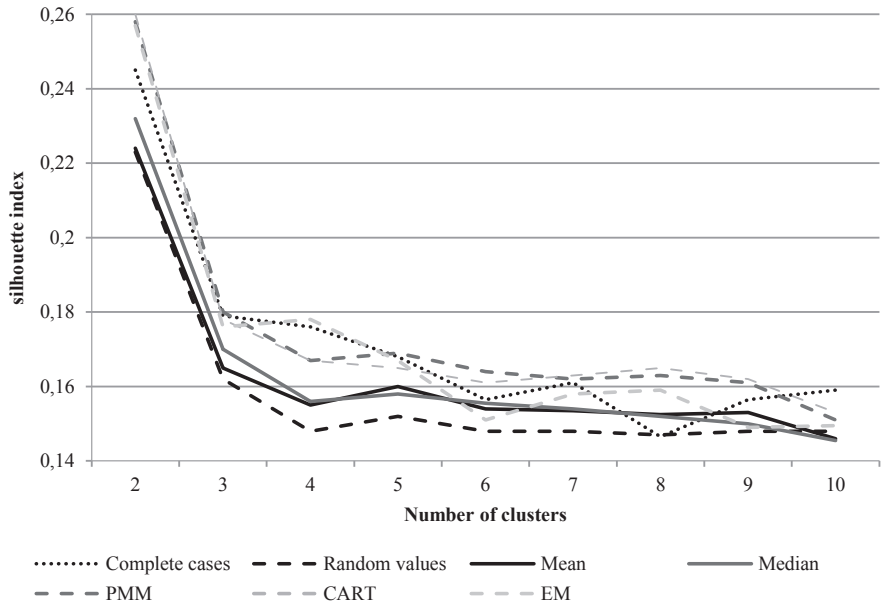

Fig. 1. Silhouette index values for different imputation methods – k-means method

Source: own research.

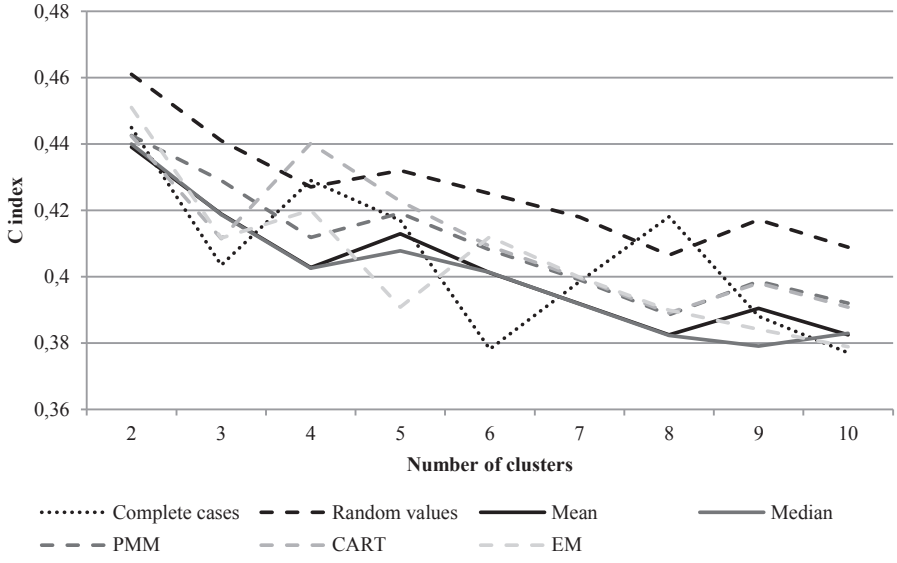

Fig. 2. C index values for different imputation methods – k-means method

Source: own research.

The evaluation of the subsequent methods of estimating missing values performed on the basis of the value of the C index (Figure 2), gives somewhat different results. In this case we can see again that the lowest quality of clustering (expressed by highest index values) is achieved in with imputation of random values. Nevertheless, the results are still far from unambiguous as in the case of the silhouette index – in the case of 4 and 8 clusters the set of complete cases and supplemented with use of the CART algorithm were characterized by a quality lower than the random values imputation.

Contrary to the previous index, the C index values suggest that the supplementation of data with mean and median seems to bring better results than those using the PMM and CART methods, and in cases of a larger number of clusters – better than the EM algorithm.

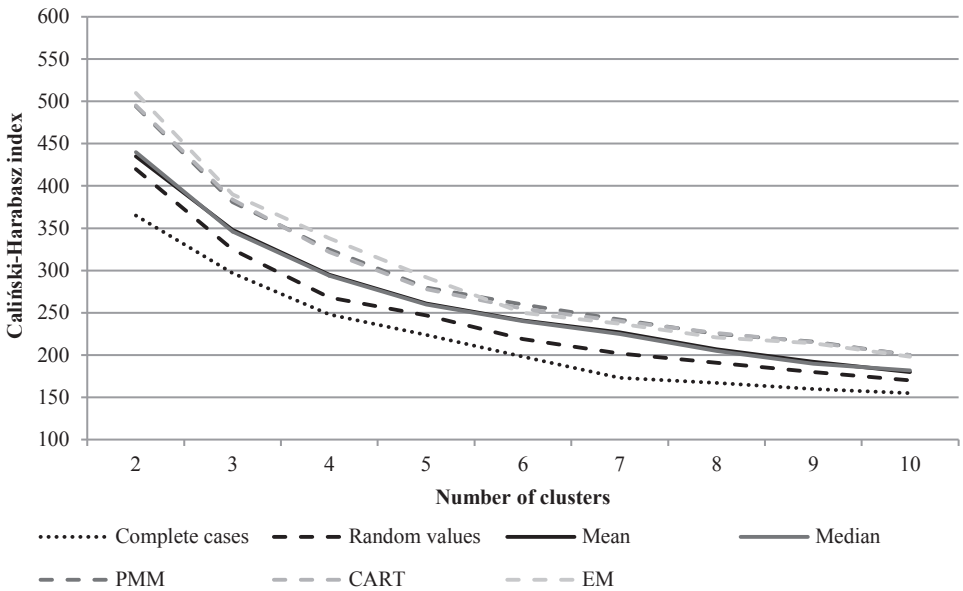

Fig. 3. Caliński-Harabasz index values for different imputation methods – k-means method

Source: own research.

A very clear picture was achieved with use of the Caliński-Harabasz index (Figure 3). We can see the gradual increase of the index value, and thus of the quality of clustering after using subsequent possible methods. Its lowest values were recorded for the set of complete cases, with clustering quality improving with imputation of random values, then the median or mean, and finally, in a very similar range, by other methods applied.

4. Imputation and the k-medoids method

The other analyzed method which assumes the ordinal level of measurement was the k-medoids method, with the GDM measure of distance applied. Unfortunately the values of the silhouette index calculated for the subsequent numbers of clustersselected with this method (Figure 4) does not lead us to such unequivocal conclusions as was the case with the k-means method.

What is worth observing here is that the imputation of random values seems to be an option that is, in most numbers of clusters (excluding 5 and 6), not worse than the others. Opposite to the situation observed in the case of the k-means method, we cannot say that the imputation of missing values with median (or possibly mean, which still - taking the ordinal level of measurement – should not be used) results in the worse quality of clustering than the analysis of complete cases, we also cannot confirm the higher quality achieved with the use of the remaining methods.

The comparison of the applied methods produces more legible, yet still ambiguous results in the case of the C index. Similarly to the k-means method, also in this case the imputation of random values is reflected, in the majority of situations, by lower clustering quality (high index values) than that achieved with other methods. The remaining approaches seem comparable in the case of the lower number of cluster (2-6), in cases of a larger number of clusters the best quality is observed in the analysis of complete cases and the CART algorithm.

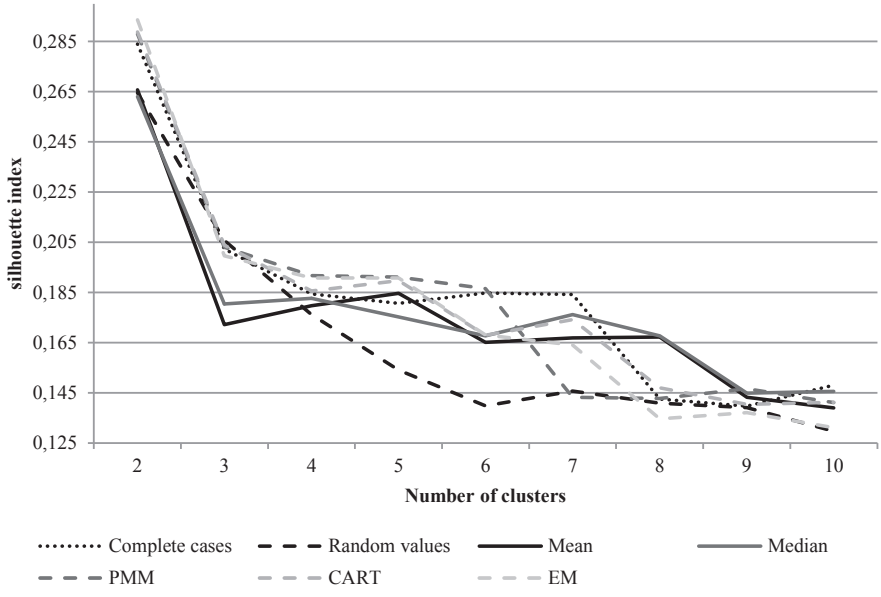

Fig. 4. Silhouette index values for different imputation methods – k-medoids method

Source: own research.

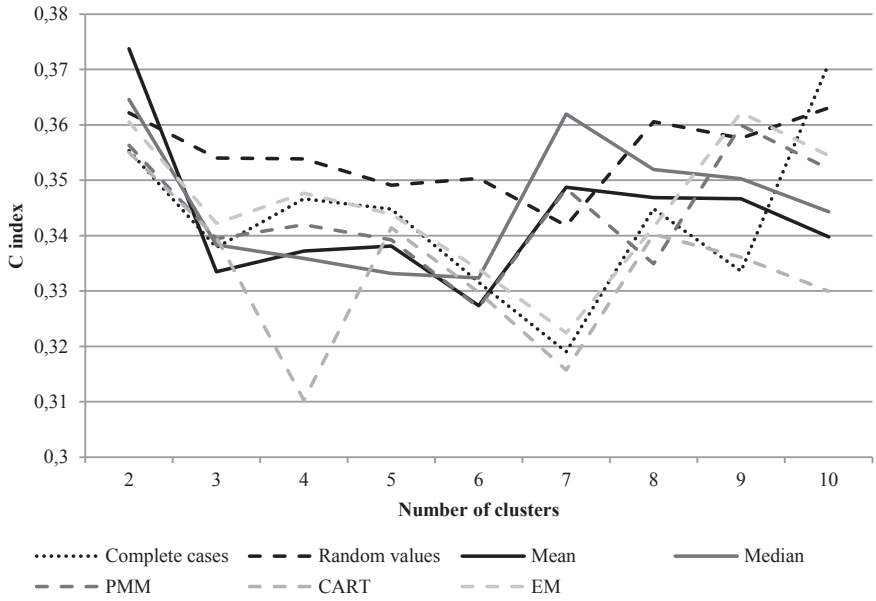

Fig. 5. C index values for different imputation methods – k-medoids method

Source: own research.

5. Conclusions

The presented research material leads to a number of significant conclusions pertaining to the imputation of missing data and the way the adopted method influences the cluster analysis.

First of all, we ascertained that the imputation of data is justified; the analysis performed with sets of complete cases and those performed on datasets in which the missing values were supplemented with random values are usually – as it does not include all possible numbers of clusters that were investigated – characterized by a better clustering quality. It should be noted that the approach of supplementing the gaps with one, constant value (median, mean) is usually (the silhouette index and the Caliński-Harabasz index for the k-means method) less effective. Imputation of the same value flattens the differences between the cases, thus worsening the clustering quality.

The situation is more complex in the case of the k-medoids method. It is hard to define clearly the optimal method for gap imputation in classifications led with the use of this method. For small number clusters the methods seem comparable, for larger numbers – the analysis of indexes results in divergent conclusions.

The evaluation of other methods for data imputation that measure the missing values on the basis of the remaining observations is also not unambiguous. Still we

do observe that from the viewpoint of the resulting cluster qualities the best effects, in the majority of cases, are achieved with the use of the classification and regression tree (CART).

Finally, it is worth indicating several problems that appeared during the analyses conducted. First of all the approach that analyses the complete cases and their comparability with datasets where gaps were filled requires further evaluation. This comparability may not be full, as the number of complete cases is lower than the total number of cases, and we may assume that the clustering of complete cases will be characterized by lower variability than the clustering of all cases. It is still to be recognized that the values of the Caliński-Harabasz index, based on variance, to some extent contradicts that hypothesis.

Secondly, we must ask the question about the connection between the growth of clustering quality and the fact that the data gaps were filled exclusively on the basis of values of internal variables, that is those used for cluster analysis. The increase of that quality may be connected with the fact that the EM and CART algorithms chose the filled values so that they conformed with the regularities already existing in the data, further enhancing them. That is why it would be extremely interesting to further develop the analyses conducted with datasets in which the gaps were filled on the basis of other data that did not form the basis for the conducted clustering.

Bibliography

Caliński T., Harabasz J., 1974, *A dendrite method for cluster analysis*, Communications in Statistics, 3 (1), pp. 1-27.

Charrad M., Ghazzali N., Boiteau V., Niknafs A., Charrad M.M., 2014, *Package 'NbClust'*, Journal of Statistical Software 61, pp. 1-36.

Cole J.C., 2008, *How to deal with missing data*, Best Practices in Quantitative Methods, pp. 214-238.

Hubert L.J., Levin J.R., 1976, *A general statistical framework for assessing categorical clustering in free recall,* Psychological Bulletin, 83(6), pp. 1072-1080.

Jajuga K., Walesiak M., Bak A., 2003, *On the General Distance Measure*, [in:] *Exploratory Data Analysis in Empirical Research*, Springer Berlin Heidelberg, pp. 104-109.

King G., James H., Anne J., Kenneth S., 2001, *Analyzing incomplete political science data: an alternative algorithm for multiple imputation*, American Political Science Review 95 (1, March), pp. 49-69.

Labovitz S., 1967, *Some observations on measurement and statistics*, Social Forces, 46(2), pp. 151-160.

Little R.J.A., 1988, *Missing data adjustments in large surveys*, Journal of Business Economics and Statistics, 6, pp. 287-301.

Marcus-Roberts H.M., Roberts F.S., 1987, *Meaningless statistics*, Journal of Educational Statistics, 12, pp. 383-394.

Migdał-Najman K., 2011, *Ocena jakości wyników grupowania-przegląd bibliografii*, Przegląd Statystyczny, 58(3-4), pp. 281-299.

Rousseeuw P., 1987, *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis*, Journal of Computational and Applied Mathematics, 20, pp. 53-65.

Rubin D.B., 1976, *Inference and missing data*, Biometrika, 63, pp. 581-592.
Sambandam R., 2003, *Cluster analysis gets complicated*, Marketing Research, vol. 15, no. 1.
Stevens S., 1946, *On the theory of scales of measurement,* Science, 103(2684), pp. 677-680.
Schafer, J.L., 1997, *Analysis of Incomplete Multivariate Data*, Chapman & Hall, New York.
Schafer J.L., Graham J.W., 2002, *Missing data: our view of the state of the art*, Psychological Methods, 7(2).
Walesiak M., 2006, *Uogólniona miara odległości w statystycznej analizie wielowymiarowej*, Wydawnictwo Akademii Ekonomicznej im. Oskara Langego, Wrocław.
Walesiak M., Dudek A., 2006, *Symulacyjna optymalizacja wyboru procedury klasyfikacyjnej dla danego typu danych – oprogramowanie komputerowe i wyniki badań*, Taksonomia 13, Prace Naukowe Akademii Ekonomicznej we Wrocławiu 1126, pp. 120-129.