DOI: 10.53122/ISSN.0452-5027/2026.1.08 kontrola i audyt

Data Analytics in Auditing

Use of Data Clustering and its Simple Implementation

In his article, the author presents one of the methods used in audit data analytics (ADA) and in unsupervised machine-learning algorithms, namely data clustering. It has found practical application in various decision-making areas (e.g. marketing, banking, insurance, social research, medicine, biology), while its application in auditing is still relatively limited. This is mostly because the audit community perceive it as overly complex, in terms of both mathematics and tools. The author discusses the subsequent stages of data clustering, presents selected measures and methods with a view to their application in auditing, and proposes a simple implementation of selected methods that allows for analysis automation. He also pays attention to a broad potential that data clustering may have at different stages of an audit, beyond typical application of this method for detecting anomalies in financial audits. Such analysis may be particularly useful for Supreme Audit Institutions due to the wide thematic scope and coverage of audits they conduct. This also applies to audits carried out by NIK, especially performance audits. The issue is presented from a perspective of an auditor and engineer simultaneously, rather than from a purely academic one, and the proposed implementation is based on tools that are widely known to the community of auditors.

WIESŁAW KARLIŃSKI

General Assumptions and Basic Methods of Data Clustering

Data clustering is the process of grouping a set of data objects into multiple groups (clusters) so that objects within a cluster are very similar to one another, but differ from objects in other clusters. Differences and similarities are identified based on the values of the variables describing the objects, and distance measures are

frequently used for this purpose[1]. The method belongs to the category of unsupervised machine-learning methods and, in the context of audit, to audit data analytics (ADA).

J. Korzeniowski[2] points out that after many years of experience in developing and applying data clustering, a complete analysis can be divided into the following stages:

- selection of objects and variables,
- visualization of objects (or variables),
- normalization of variables,
- selection of a distance measure between objects,
- selection of a clustering method,
- setting the number of clusters,
- grouping of objects – the proper stage of data clustering,
- assessment of grouping results,
- description and profiling of clusters.

The selection of objects depends on the objectives and scope of the examination and – while in some areas (e.g. in marketing) examination is based on a random sample – in auditing is most often dealt with an analysis of a complete population of transactions, cases or entities. Objects with missing or erroneous data must be removed from the set, which means that selection of objects is closely linked to selection of variables. Selection of variables that describe objects is one of the most important and most difficult stages of the examination. M. Walesiak[3] proposes to divide it into two phases:

- Phase I – establishing an initial list based on knowledge of the subject, availability of relevant data and potential cooperation with experts in the field;
- Phase II – reducing the initial list by removing variables with low variability and variables that are mutually related (with a high Pearson linear correlation coefficient r).

New variables can also be created to eliminate relationships among original variables (e.g. using BMI instead of weight and height), or to obtain information relevant to the audit objectives (e.g. deriving timeliness from dates, or calculating key indicators). When defining the final set of variables, the type of variables: quantitative (metric) or qualitative (ordinal or nominal), and their number should also be considered (with more than four variables and hundreds of thousands of observations, computations may be time-consuming).

Data visualization prior to the actual analysis allows for an optical estimate of the number of clusters and their structure, which may have an impact on the choice of analysis methods. For a set of two variables, a clear picture is obtained on a scatter plot. For three variables (X, Y and Z), a three-dimension scatter plot can be constructed (if appropriate tools are available), or three two-dimensional plots (X–Y, X–Z and Y–Z). Obtaining a third dimension via a bubble chart is not effective with a large number of observations.

---

[1] J. Han, J. Pei, H. Tong: *Data Mining. Concepts and Techniques* (4th edition), Morgan Kaufmann Publishers, Cambridge, MA, USA, 2023.

[2] J. Korzeniowski: *Methods of Selection of Variables in Data Clustering. New Methods*, University of Łódź, 2012.

[3] M. Walesiak: *Issue of Variables Selection and Measurement in Classification*, *Taxonomy 12*, "AE Scientific Works in Wrocław" No 1076/2005.

If the number of variables exceeds three, visualization becomes even more complicated and may be impossible at this stage.

An important element of most data clustering methods applied to metric variables is measuring the distance between individual observations. Among numerous distance measures, a commonly used and simple one is the Euclidean distance, which for a pair of observations (A and B) described by n variables is expressed by the formula:

$$d = \sqrt{(A_1 - B_1)^2 + (A_2 - B_2)^2 + \cdots + (A_n - B_n)^2} \quad (1)$$

Because for each variable (1 to n), squared differences in observation values are summed, if variables are presented on scales with different ranges (often differing by one or even several orders of magnitude), variables with large ranges will affect the result, and the result can even depend on measurement units (e.g. whether the area is provided in m^2, km^2, or hectares). To avoid this, data normalization is typically applied, i.e. transforming variables to a comparable measurement scale. Some researchers (e.g. G. Milligan[4]) argue that although normalization is not mandatory, and it even may distort the cluster structure, in auditing – where variable scales can vary a lot – it appears indispensable. Only when variables are expressed in the same units and have similar value ranges, can normalization be omitted and clustering performed on raw data.

In the literature (see e.g. M. Walesiak[5]) various normalization methods can be found. In general, normalization can be expressed by the following formula:

$$z_i = \frac{x_i - a}{b} \quad (2)$$

where x_i is the original (raw) variable value, z_i is the normalized value and a, b are parameters depending on the normalization type.

The simplest normalization methods commonly used in data clustering include:

- Min-max normalization, known as zero unitarization, where the parameter a is assigned to the minimum of variable x, namely a=min(X), the parameter b is assigned to its range, which means that b=max(X) – min(X), and the result lies in the interval [0,1];
- Classical standardization (z-score), where the parameter a is the mean a=m(X), the parameter b is the standard deviation b=sd(X) and the standardized variable has the mean m=0 and standard deviation sd=1.

The min-max normalization is the most intuitive, but when strong outliers (extreme observations) occur, which is frequent in full-population audit datasets examinations in auditing, it becomes less useful for clustering. As a result, an extremely large observation becomes 1 after normalization, while the vast majority of the remaining observations cluster near 0 (or vice versa

[4] G. Milligan: *Clustering Validation: Results and Implications for Applied Analyses*, Clustering and Classification, P. Arabie (ed.), L.Hubert, G. de Soete, World Scientific, Singapore, 1996.

[5] M. Walesiak: *Review of Formulas for Normalization of Variables Values and Their Properties in Statistical Multi-dimensional Analysis*, "Przegląd Statystyczny", R. LXI, volume 4, 2014.

with an extremely small observation), so they will vary only a little. For example, if one variable were the number of citizens, and the population were 314 municipalities of the Mazowiecki Region, about 87% of municipalities would obtain a value below 0.01 after zero unitarization.

Classical standardization handles outliers much better, but even here the parameters (mean, standard deviation), and thus the normalization result, are sensitive to extreme values.

A normalization method robust to outliers is the so called positional standardization. In this method, the parameter a is the median a=Me(X), and the parameter b is the median absolute deviation MAD(X), which is often additionally adjusted by the factor 1.4826, i.e. b=1.4826·MAD(X).

MAD is calculated as:

$$MAD_{med} = Me(|x_i - Me|) \quad (3)$$

The "med" subscript in the above formula has been introduced on purpose for those readers who may have encountered a different concept of MAD (mean absolute deviation) used in forensic analytics, especially in Benford analysis[6].

Positional normalization based on MAD_{med} is not as popular as the other two, but some authors emphasize its particular usefulness in the presence of outliers in clustering in such areas as image analysis, biostatistics[7], and financial and economic analyses[8]. The method is also used when integrating metric (quantitative) data with ordinal scale data[9].

The selection of the normalization method depends on the data distribution (no outliers, isolated outliers, or skewed distribution) and the objective of the examination (segmentation of objects, detection of single anomalies, or visualization or ranking only). This issue has been discussed later in the article, with identification of the so-called global anomalies.

According to the classification proposed by J. Han *et al.*[10], basic data clustering methods can be divided into three groups:
- partitioning methods (often called non-hierarchical) that group objects around cluster centres (centroids),
- hierarchical methods, which consist in a cluster tree (dendrogram) that shows relationships among objects at different levels of detail,
- density-based methods that identify clusters as regions of high point density.

While advanced methods include, among others, model-based and probabilistic methods, graph-based methods and methods for high-dimensional data, partitioning methods are considered the simplest to implement and the most efficient for large datasets.

---

6 M. Nigrini: *Forensic Analytics. Methods and Techniques for Forensic Accounting Investigation*. John Willey & Sons, New Jersey, 2011.
7 D. M. Rocke & B. Durbin: *A Model for Measurement Error for Gene Expression Arrays*, "Journal of Computational Biology" No 6/2001.
8 C. C. Aggarwal: *Outlier Analysis,* Springer 2013.
9 P. Filzmoser, R. Maronna, M. Werner: *Outlier Detection in High Dimensions*, "Computational Statistics & Data Analysis" No 3/2008.
10 J. Han, J. Pei, H. Tong: *Data Mining*..., op.cit.

Simple Data Clustering Methods Useful in Auditing

Among partitioning methods, the most popular for metric (quantitative) data is k-means. In this method, assigning objects to a given number k of clusters is achieved by moving objects between clusters until within-cluster variability and between-cluster variability are optimized. The process is iterative and carried out in several steps[11]:

1. determining the number of clusters,
2. defining the stopping criterion (no object moves between clusters and a maximum number of iterations),
3. selection of a metric, i.e. how distance between objects is measured,
4. determining cluster centres (centroids): in the first iteration, e.g. randomly, and in subsequent iterations, e.g. as the arithmetic mean of coordinates of points belonging to a given cluster,
5. measuring distances of objects to cluster centres,
6. assigning objects to clusters – for each observation, we compare its distances to all clusters and assign it to the cluster whose centre is the closest,
7. checking the stopping criterion; if it has not been met, return to step 4.

The method is relatively simple and fast, but it requires specifying the number of clusters *a priori*, it is less effective for eccentric clusters, it is sensitive to outliers and additionally it produces results that are little stable, due to the random selection of centroids in the first iteration, which creates the need to run it many times and to choose the most advantageous variant.

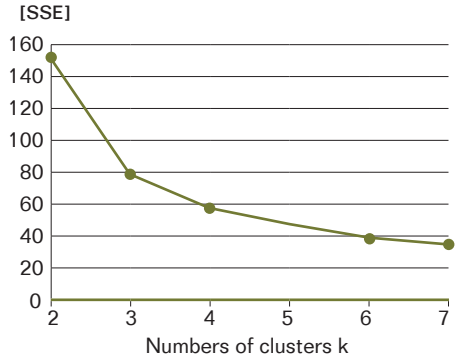
Most publications suggest that the k number of clusters be selected experimentally, with the use of the so called elbow method. It is a heuristic method that consists in computing the sum of squared errors (SSE) for each data point from the cluster centroid for various values of k and presenting the results on a simple chart. SSE decreases as the number of clusters increases, the optimal k is taken as the value below which the rate of SSE decrease diminishes. The figure below presents the elbow method for the iris[12] dataset, a typical dataset used in statistics and machine learning to test algorithms, containing descriptions of three iris species (setosa, versicolor and virginica) with four variables. Although the dataset is not optimal for using the k-means method, we can see that the elbow point corresponds to the number of clusters k=3.

The impact of extreme outliers (global anomalies) on the k-means result can be avoided by excluding such observations after normalization, but before the proper data clustering. In audit practice, this does not mean removing observations from the dataset, but rather flagging them, excluding them from data clustering, and subjecting them to a detailed analysis in another mode. To identify outliers under classical and positional standardization, one can use a simple Euclidean distance

[11] A. Królak-Nowak, K. Kotarba: *Basics of Machine Learning*, AGH, Kraków 2022.

[12] <https://ocw.mit.edu/courses/15-097-prediction-machine-learning-and-statistics-spring-2012/resources/iris/> (access 22.1.2026).

Figure 1. **Selection of numbers of clusters with the *k-means* method for the iris dataset (raw data)**



Source: Author's own calculations, made with the use the algorithm described later.

in the standardized space. For each standardized observation, a sum of squared deviations from the central point is calculated, which after standardization has a zero value. As a rough selection criterion, threshold values can be taken from χ^2 (chi-square) distribution quantiles. Thus, the selection criterion for outliers (global anomalies) for n variables can be presented as follows:

$$D_i^2 = z_{1i}^2 + z_{2i}^2 + \cdots + z_{ni}^2 > \chi^2(n, \alpha) \quad (4)$$

The materiality threshold α can be set at the conservative level of 0.01.

The proposed algorithm is based on the measuring distance from the typical value in the whole dataset, which is the mean (in classical standardization), or the median in positional standardization. The algorithm cannot be used for clustering raw data or data normalized by min-max due to the lack of a typical value in a probabilistic sense. In both cases, the Euclidean distance should be replaced with the Mahalanobis distance from the central point, and then the above chi-square selection criterion can be applied. The Mahalanobis distance is a popular, although somewhat more complex, measure used in data clustering that accounts for correlations between variables, and hence the author abandoned presenting its mathematical apparatus.

The choice of normalization methods in the context of global anomalies can be summarized as follows. Positional standardization will effectively eliminate the impact of isolated outliers, while preserving clear relationships among the remaining observations. However, when the dataset is strongly skewed and the upper tail includes 10% or more of observations, it may classify too many observations as outliers. Under classical standardization, the number of global anomalies will be lower because in such distributions the mean is higher than the median. In auditing, it is worth using an adaptive normalization scheme depending on the distributions of variables. For near-symmetric distributions, classical z-score standardization should be used. While for data with large outliers or strong skewness, which are typical of financial and operational variables, positional standardization can be applied. If the share of observations flagged as outliers after positional standardization is small, they can be treated as global anomalies. Otherwise, this identification should be interpreted as preliminary separation of units of a different scale of activity, and further segmentation should be performed on the remaining observations, which will improve the quality of grouping. If no preliminary separation is intended, and the share of observations classified as global

anomalies exceeds an operationally useful threshold (e.g. 20%), it is better to replace positional standardization with classical one.

Data unitarization (min-max normalization) can be then used for visualization and in cases when position rather than scale matters, and when there are no outliers. Using raw data for clustering applies to special cases – when the variable scale is comparable and units are the same.

As for concerns about stability of results, a modification of k-means that optimizes the selection of initial points can be a solution. The relevant algorithm, called k--means++, was suggested in 2007[13] and it gives more repeatable results. A satisfactory effect can also be achieved by increasing the number of random drawings in classical k-means, as applied in the procedures described below.

The use of k-means in audit is proposed, among others, by S. Thiprungsri and M. Vasarhelyi[14], who applied it in practice to study claims in group life insurance, based on quantitative data. Data type is crucial here, because k-means is based on measuring distance (differences) and therefore it can only be applied to metric (quantitative) data. For ordinal data, one cannot compute differences – only the relation smaller/equal/greater can be applied, while for nominal data only equality can be used. In certain situations (if a single variable is important and it is ordinal, and the response portfolio is sufficiently large (e.g. five or more)), it may be acceptable to include it in k-means alongside quantitative data. However, positional normalization (with the use of MAD_{med}) should be applied, and in the distance measure (formula 1) an appropriate weight lower than 1 can be assigned to that variable. The best solution, however, is to use a method dedicated to qualitative (ordinal or nominal) variables, namely the k-modes method.

The k-modes method works similarly to k-means (see e.g. J. Han *et al.*[15]), with the following differences:

- original data do not have to be numeric and are not normalized,
- similarity of objects A and B is not the Euclidean distance (A-B), but the number of variables for which Ai=Bi,
- centroids are updated using the mode (dominant) that is the most frequent category in the cluster, instead of the mean.

The algorithm can be described as follows:

1. Select k of initial centroids, one for each cluster.
2. Assign each object to the cluster whose centroid is closest, using the similarity measure.
3. For each cluster, the centroid is updated by computing the mode of each attribute within the given cluster.

[13] D. Arthur, S. Vassilvitskii: *k-means++: The Advantages of Careful Seeding. Proceedings of the Eighteenth annual ACM-SIAM Symposium on Discrete Algorithms*, Society for Industrial and Applied Mathematics Philadelphia, PA, USA, 2007.

[14] S. Thiprungsri, M. Vasarhelyi: *Cluster Analysis for Anomaly Detection in Accounting Data: An Audit Approach*, "The International Journal of Digital Accounting Research", Vol. 11, 2011.

[15] J. Han, J. Pei, H. Tong: *Data Mining*..., op.cit.

4. Repeat steps 2 and 3 until there are no changes in cluster memberships, or a specified number of iterations is reached.

The quality measure in k-modes is not the value of SSE, but the minimizing of the so-called cost function (J), defined as the total number of attribute mismatches between observations and the cluster dominant. The cost function is also used to select the optimal number of clusters k with the use of the elbow principle analogous to k-means.

In audit practice, it is common that datasets describing a process contain both metric (quantitative) and qualitative variables. There is a hybrid approach called k-prototypes that combines both methods (k-means and k-modes), but using it creates numerous complications (e.g. selecting the optimal number of clusters) and the results obtained are less clear to the reader. A better solution is to split variables into two groups (quantitative and qualitative) and apply clear methods to each group. A division into two categories was used by D. Wei *et al*.[16] in a study of a set of accounting entries. For quantitative variables they propose the LOF (Local Outlier Factor) method, while for qualitative variables – k-modes. The LOF method is based on density analysis, and it is not a classic data clustering method; it rather identifies outliers, which is in line with the objective of the examination intended to detect outliers at the transaction level.

After applying the two algorithms (in our case: k-means and k-modes), for each observation we obtain a quantitative cluster assignment and an independent qualitative cluster assignment. The two results can be combined by creating a cluster matrix, a mechanism commonly used in auditing e.g. in risk analysis (a risk matrix). It should be noted, however, that cluster numbers have no substantive meaning, so before building the matrix it is worth assigning them with labels (ordinal numbers or names), corresponding to the cluster's meaning, i.e. the post-hoc cluster labelling. For numeric data, cluster meaning may be indicated by high mean values of certain variables within the cluster, and for qualitative data – e.g. by individual types of cases. To increase interpretability and identify rare feature combinations, for k-means × k-modes matrix it is worth indicating cluster counts, optionally supplemented with percentages. Such a table is clear and it does not require additional tools to interpret.

If it turns out that a dataset contains a single quantitative variable and the rest are qualitative, a solution is to convert the quantitative variable into an ordinal one through stratification and then apply only k-modes.

Data clustering is often used in audit to identify anomalies. S. Thiprungsri and M. Vasarhelyi[17] note that anomalies may include:

• observations that belong to none of the clusters,

[16] D. Wei, S. Cho, M. Vasarhelyi, L. Te-Wierik: *Outlier Detection in Auditing: Integrating Unsupervised Learning within a Multilevel Framework for General Ledger Analysis*, "Journal of Information Systems" No 2/2024, American Accounting Association.

[17] S. Thiprungsri, M. Vasarhelyi: *Cluster Analysis for Anomaly*..., op.cit.

• observations farthest from the cluster centroid,
• observations forming small or rare clusters.

In the method described above, globally outlying observations (belonging to none of the clusters) are identified using the mechanism in formula (4). Identification of small clusters is simple, as it is performed by counting observations in each cluster. Classical k-means method, however, does not handle identifying observations far from centroids (local anomalies) or rare clusters. This can be addressed by extending standard k-means functionality to measure and interpret distances. For each observation, in addition to the cluster number, we also obtain the distance (d_i) from the centroid. It should be emphasized that in the k-means algorithm centroids are theoretical points and do not have to correspond to any actual observation.

If we want a distance measure comparable across clusters, we should normalize d_i, e.g. by dividing it by the median[18] computed for each k cluster:

$$d'_i = \frac{d_i}{Me(d_k)} \quad (5)$$

The process of identifying local anomalies can then be automated with the use of the following rule:
• $d'_i < 2$ – typical observation,
• $d'_i = 2 - 3$ – atypical observation,
• $d'_i = 3$ and more – outlying observation.

In a situation when the number of variables exceeds 5, it is worth increasing the threshold values (2 and 3) by introducing a correction factor such as $(n/3)^{1/2}$, where n is the number of variables.

Implementation of Simple Clustering Methods in Auditing

P. Byrnes[19] notes that performing clustering step by step is a major challenge for typical auditors. In his view, efforts should be made to automate the process so that attention can be focused on interpreting the results. Knowledge of appropriate analytical tools is also important. Commercial statistical packages and popular open-source IT tools provide procedures and libraries dedicated to data clustering (in Python: the libraries Scikit-learn, NumPy and Pandas; while in R – the packages cluster, stats and flexclust). However, these tools are not yet popular among auditors. The objective of this article is to present relatively simple data clustering solutions, operating semi-automatically and implemented in an environment that auditors are familiar with, such as MS Excel. Therefore, the author attempted to develop and test procedures in VBA, the working names of which are *Audit_kMeans*, *Audit_kMeans_LocalRisk* and *Audit_kModes*, that implement the methods discussed earlier: k-means, k-means with local anomaly identification and k-modes.

[18] We can also consider normalization with the interquartile range.
[19] P. Byrnes: *Automated Clustering for Data Analytics*, "Journal of Emerging Technologies in Accounting" No 2/2019, American Accounting Association.

The procedures assume that input data are recorded in an Excel worksheet in columns (according to the number of variables) starting from column A, and the first row contains variable names.

The *Audit_kMeans* procedure requires input parameters: number of variables, normalization method (classical, positional, min-max, or raw data) and the maximum number of clusters (k_{max}). Data are automatically normalized and checked for outliers (global anomalies) and the verification result is recorded in the worksheet. Under min-max normalization and raw data, global anomalies are determined with the use of the Mahalanobis distance. Clustering is performed with the use of the k-means method for successive values of k, starting from two up to the maximum indicated value (k_{max}). Observations identified as global anomalies are automatically excluded. For each k, a specified number of drawings (default 10) is performed and the best result is recorded, together with the respective SSE value. The maximum number of iterations per run is 100 by default. The entire operation is automatic and after the procedure it is enough to create a simple line chart of SSE versus k (see Figure 1) and to decide which k in the range 2-k_{max} and the corresponding assignment of observations to clusters are optimal.

After selecting the optimal number of clusters k, it is worth applying the *Audit_kMeans_LocalRisk* procedure. The procedure is run for a specified number of variables, normalization type (method same as above) and number of clusters. As a result, for each observation we achieve: the cluster assignment, the distance to the centroid, the median-normalized distance and the note on typicality/atypicality (typical/atypical/outlying). In addition, for presentation purposes, the procedure foresees an option to save normalized data to the worksheet.

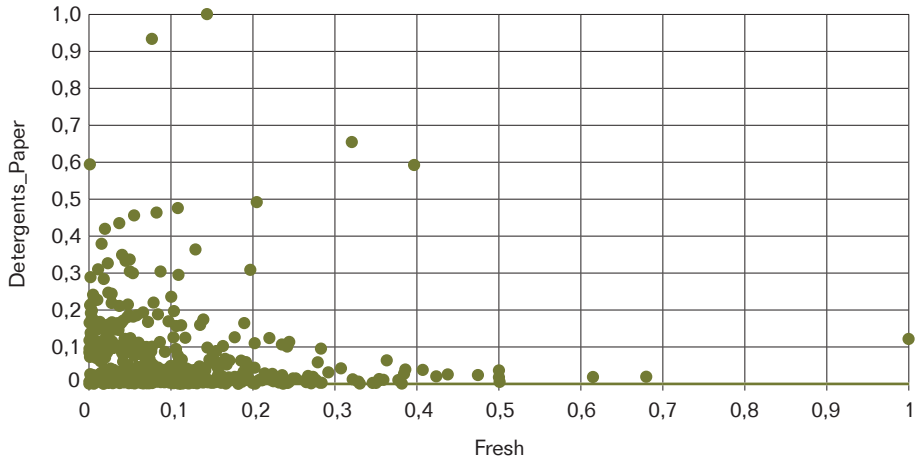
The *Audit_kModes* procedure is handled analogously to *Audit_kMeans*, i.e. the number of qualitative variables and the maximum number of clusters should be provided, but without specifying the normalization method. After computations, the best results from drawings (default 10) are presented and the cost function values (J) for each number of clusters. Also for k-modes, automatic selection of the number of clusters k is not provided, as the decision is left to the user.

The procedures were tested on several reference datasets. For the iris dataset mentioned above, k-means with classical normalization achieved the following agreement level: setosa 50/50, versicolor 39/50 and virginica 36/50, while with raw data: setosa 50/50, versicolor 48/50 and virginica 36/50. Using of raw data was justified due to similar variable scales and it gave a slightly better accuracy, although the dataset is specific because versicolor and virginica clusters are not spherical and overlap.

To estimate the efficiency of the *Audit_kMeans_LocalRisk* procedure, an analysis was conducted on a general ledger dataset containing three quantitative variables and about 230,000 observations, which obtained a runtime below two minutes, which should be satisfactory for financial auditors.

To demonstrate the effectiveness of different analytical solutions in data

Figure 2. **Scatter plot of *Fresh* and *Detergents_Paper* variables after unitarization**



Source: Author's own study.

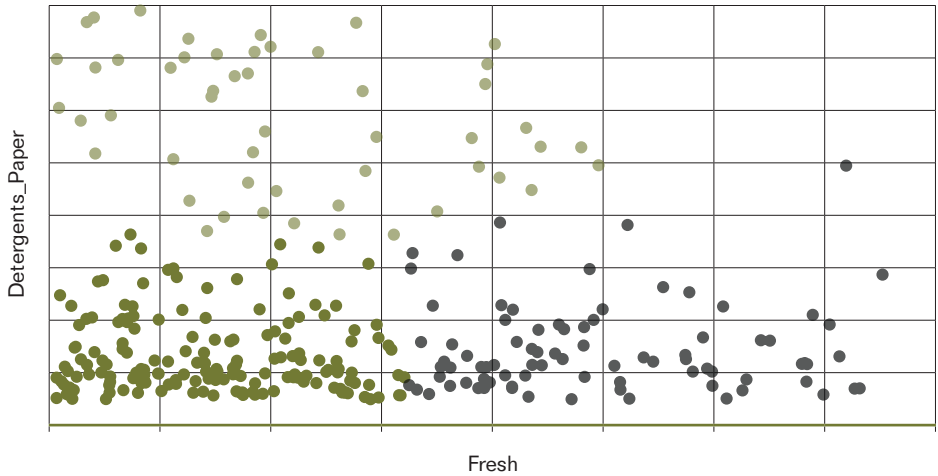
clustering, the *Wholesales_customer* dataset, available in the UCI Machine Learning Repository[20], was used. The dataset contains data on 440 customers across six product categories (Fresh, Milk, Grocery, Frozen, Detergents_Paper, Delicassen). Additionally, the set included two qualitative variables (Channel and Region). It was assumed that the objective of the examination was customer segmentation at the stage of preparing an audit or final report.

Quantitative variables were analysed for mutual correlations (Correlation option in Excel's Data Analysis Tools Add-In), identifying strong correlations between some variables. For demonstration purposes, two variables were selected, reflecting the customer business type and operational intensity (*Fresh* and *Detergents_Paper*). A scatter plot for *Fresh* and *Detergents_Paper* after zero unitarization is shown in the figure below.

With the application of the two normalization methods mentioned earlier to these variables, the following number of global anomalies was obtained: 18 – for classical standardization and 135 – for positional standardization. The very high share (about 34%) of outliers under positional standardization results from skewed distributions and indicates units operations at a different scale. For segmentation of customers, they were treated as a separate category rather than as global anomalies. Such an operation improved the clustering quality for the remaining customers (a more clearly articulated

[20] https://archive.ics.uci.edu/dataset/292/wholesale+customers (access 19 Feb 2026)

Figure 3. **Customer clustering (excluding units with very high turnover)**



Source: Author's own study.

elbow point indicating three clusters). After applying the *Audit_kMeans_LocalRisk* procedure, the remaining 305 customers were grouped into three clusters, named according to their characteristics: 1 – mixed purchasing (moderate level), 2 – chemicals dominant, 3 – food dominant. At the same time, the procedure revealed six atypical objects in cluster 1 and five atypical plus one outlying object in cluster 3. The result of the customer grouping is shown in Figure 3.

As for the two qualitative variables in *Wholesales_customer* (*Channel* and *Region*), an attempt was made to apply the k-modes data clustering. The *Audit_kModes* procedure indicated an optimal number of five clusters, while the two variables together contained 2×3=6 different values. Thus data clustering did not provide a meaningful reduction compared to a simple contingency table, due to the small number of categories and their unambiguous combinations. Therefore, only one qualitative variable was used to stratify k-means results. The variable was selected that described the mode of process operation, and in that dataset it was the variable describing the distribution channel (*Channel*). The results of the analysis can be presented in the table below with category counts.

The table shows that preliminary separation of very high-scale units allowed for obtaining a stable and interpretable segmentation of the remaining observations, i.e. separating purchasing profile from returns to scale. It can be observed that distribution channels also differ: channel 2 is focused on high turnover, while channel 1 is structurally diversified. Such a breakdown can be used in risk assessment, audit strategy planning, auditee selection, or reporting on audit results.

Table 1. **Summary of clustering results by quantitative and qualitative variables**

k-means/k-modes	Channel 1	Channel 2	Total
1-mixed purchase	166	5	171
2-chemical dominant	23	28	51
3-food dominant	79	4	83
Very high turnover	30	105	135
Total	298	142	440

Source: Author's own study.

In the example above, for ease of graphical presentation, the number of quantitative variables was deliberately limited to two, which may raise doubts about the method's usefulness and whether simple filtering could replace it. A fundamental qualitative difference becomes apparent only with a larger number of variables, where filtering is clearly insufficient. An additional advantage is full process automation, allowing the auditor to focus on substantive issues.

Areas of Application of Data Clustering in Auditing

Among numerous applications of data clustering in business, marketing, biology, medicine, social research, or image analysis, segmentation and detection of unusual behaviour prevail. While current applications in auditing focus mainly on anomaly detection and concern financial audits. The author's experience in the methodology of audit planning and conducting indicates that such a narrow application of data clustering does not exploit the method's potential. Data clustering in auditing can have not only a detection function, but also exploring, planning and benchmarking functions, and it can support risk assessment, audit procedures design and interpretation of complex data populations. These functionalities may be particularly useful for supreme audit institutions, given the broad scope of their audit activity. In the case of the Supreme Audit Office (NIK), data clustering can be applied in planned audits, and especially in coordinated ones, as well as in certain *ad hoc* audits focused on specific issues.

Potential areas of application of the method in auditing include:

1. Segmentation of processes and auditees: data clustering allows for grouping auditees on the basis of similarity of their financial and operational indicators, thanks to which high-risk groups of entities can be identified (important in assurance audits) and for properly selecting entities for performance audits. While in reporting, ex post segmentation based on data gathered during the audit allows for presenting more balanced assessments;
2. Benchmarking and efficiency assessment: grouping entities by unit costs, productivity, and operational indicators to avoid false conclusions and comparing similar entities only;
3. Audit planning: grouping areas by materiality, process complexity and results of previous audits;
4. Identification of anomalies and irregularities: isolating observations significantly different from others – not only in financial audits, but also in compliance (identification of deficiencies patterns) and performance audits (e.g. suspicious transactions in cost analysis);
5. Identification of atypical processes (not transactions) and detection of alternative process paths;

Table 2. **Potential application of k-means and k-modes procedures in various audit areas**

Audit area	Data clustering purpose	Data type	Method	Result
Planning	Auditee selection	Quantitative	k-means	Prioritization of risk areas
Reporting	Ex post segmentation	Mixed	k-means k-modes	Balanced assessments
Audit of financial statements	Segmentation of transactions	Quantitative	k-means	Grouping of similar transactions, support in sampling
Audit of financial statements	Detection of anomalies	Quantitative	k-means with anomalies analysis	Identification of atypical observations
Process examination	Typology of process paths	Quantitative	k-modes	Detection of untypical processes
IT audit	User profiling	Mixed	k-means k-modes	Identification of segregation of duty conflicts
Performance audit	Benchmarking jednostek	Quantitative	k-means	Assessment of relative efficiency
Performance audit	Efficiency assessment	Quantitative	k-means	Identification of weaknesses
Compliance audit	Analysis of violations	Nominal	k-modes	Identification of violations patterns
Compliance audit	Analysis of procurement	Mixed	k-means k-modes	Detecting threshold circumvention
Data audit	Data quality assessment	Quantitative	k-means	Identification of systemic errors
Sampling	Stratification	Quantitative	k-means	Better sample representativeness

Source: Author's own study.

6. Supplier and public procurement analysis: grouping suppliers by number of contracts, procurement mode, repetition of amounts and time relationships;
7. Analysis of ERP users' behaviour: grouping by operation frequency, authorization scope and types of activities performed.

A summary of potential application areas of data clustering with a focus on k-means and k-modes is presented in the table below.

This summary shows that data clustering can be used in various types of audit, particularly in performance audits. It can help:

- identify entities, processes, or programmes that perform significantly worse than others, supporting more objective selection of areas that call for a deeper examination;
- select entities and cases for examination while accounting for cluster heterogeneity (optimal Neyman stratified sampling);
- detect hidden inefficiency by grouping entities based on input-output relations.

It is also interesting to combine data clustering for benchmarking of entities with Data Envelopment Analysis (DEA)[21]. Data clustering does not create an efficiency

[21] W. Karliński: *Data Envelopment Analysis (DEA) in Performance Auditing*, "Kontrola Państwowa" No 4/2022.

frontier nor it identifies "ideal benchmarks", it rather groups similar entities, is more robust to atypical observations and its conclusions are better received by auditees. In public sector auditing, where the objective is not to maximize efficiency, but to identify areas where improvements are needed, data clustering is a practical and communicatively safer alternative to DEA. On the other hand, data clustering does not directly measure technical efficiency, although it may address the input – output relation. Therefore, a hybrid approach is worth considering, in which DEA is applied not to the whole population, but to entities that operate under similar conditions, identified through data clustering.

Summary

Data clustering can support decision-making of audit organisations and individual auditors through segmentation of entities, identification of risks and anomalies and optimization of audit activities. The method may be particularly useful for audit institutions in the public sector, including supreme audit institutions, due to the large number and diversity of audited entities.

The use of data clustering in auditing should contribute to:

- increased efficiency of resource use by focusing on high-risk groups;
- easier decision-making thanks to synthetic characterization of object groups;
- ability to detect structures and relationships in data that are difficult to observe with traditional methods;
- reduced subjectivity in selecting objects for audit;
- better structure of assessments in audit reports.

The fact that data clustering has not yet gained sufficient popularity in auditing results, to a large extent, from being perceived as too complex methodologically and tool-wise, and dedicated to narrow applications (anomaly detection in financial audits). In large audit organizations with analytical departments, methodological and tooling barriers are secondary, while for individual auditors – they appear to be key.

The article highlights potentially broad areas of application of data clustering in auditing, it presents key methodological issues and practical solutions and it demonstrates that relatively simple implementation of selected data clustering methods is possible with the use of popular IT tools.

Eng. **WIESŁAW KARLIŃSKI**, PhD, specialist in the field of audit methodology and the application of analytical methods

Key words: data clustering, use of data clustering in auditing, data segmentation methods in auditing, anomaly identification, identification of risk areas, audit data analytics

Bibliography:

1. Aggarwal C. C.: O*utlier Analysis*, Springer, 2013.
2. Arthur D., Vassilvitskii S.: *k-means++: The Advantages of Careful Seeding*, "Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete algorithms", Society for Industrial and Applied Mathematics Philadelphia, PA, USA, 2007.
3. Byrnes P.: *Automated Clustering for Data Analytics*, "Journal of Emerging Technologies in Accounting" No 2/2019, American Accounting Association.
4. Filzmoser P., Maronna R.,Werner M.: O*utlier Detection in High Dimensions*, "Computational Statistics & Data Analysis" No 3/2008.
5. Han J., Pei J., Tong H.: *Data Mining. Concepts and Techniques* (4th edition), Morgan Kaufmann Publishers, Cambridge, MA, USA, 2023.
6. Karliński W. : *Data Envelopment Analysis (DEA) in Performance Auditing*, "Kontrola Państwowa" No 4/2022.
7. Korzeniowski J.: *Methods of Selection of Variables in Data Clustering. New Methods*, University of Łódź, 2012.
8. Królak-Nowak A., Kotarba K.: *Basics of Machine Learning*, wyd. AGH, Kraków, 2022
9. Milligan G: C*lustering Validation: Results and Implications for Applied Analyses, Clustering and Classification*, P. Arabie (ed.), L. Hubert, G. de Soete, World Scientific, Singapore, 1996.
10. Nigrini M.: For*ensic Analytics. Methods and Techniques for Forensic Accounting Investigation*. John Willey & Sons 2011.
11. Rocke D. M. & Durbin B.: *A Model for Measurement Error for Gene Expression Arrays*, "Journal of Computational Biology" No 6/2001.
12. Thiprungsri S., Vasarhelyi M.: C*luster Analysis for Anomaly Detection in Accounting Data: An Audit Approach*, "The International Journal of Digital Accounting Research" Vol. 11/2011.
13. Walesiak M.: *Issue of Variables Selection and Measurement in Classification, Taxonomy 12*, "AE Scientific Works in Wrocław" No 1076/2005.
14. Walesiak M.: Revi*ew of Formulas for Normalization of Variables Values and Their Properties in Statistical Multi-dimensional Analysis*, "Przegląd Statystyczny", R. LXI, volume 4/2014.
15. Wei D., Cho S., Vasarhelyi M., Te-Wierik L.: O*utlier Detection in Auditing: Integrating Unsupervised Learning within a Multilevel Framework for General Ledger Analysis*, "Journal of Information Systems" No 2/2024, American Accounting Association.

ABSTRACT

Use of Data Clustering and its Simple Implementation – Data Analytics in Auditing

Data clustering is the process of grouping a set of data objects into multiple groups (clusters) so that objects within a cluster are very similar to one another, but differ from objects in other clusters. Differences and similarities are identified based on the values of the variables describing the objects, and distance measures are frequently used for

this purpose. The method belongs to the category of unsupervised machine-learning methods and, in the context of audit, to audit data analytics (ADA). Data clustering has found practical application in various decision-making areas (e.g. marketing, banking, insurance, social research, medicine, biology), while its application in auditing is still relatively limited. This is mostly because the audit community perceive it as overly complex, in terms of both mathematics and tools. The author discusses the subsequent stages of data clustering, presents selected measures and methods with a view to their application in auditing, and proposes a simple implementation of selected methods that allows for analysis automation. He also pays attention to a broad potential that data clustering may have at different stages of an audit, beyond typical application of this method for detecting anomalies in financial audits. Such analysis may be particularly useful for Supreme Audit Institutions due to the wide thematic scope and coverage of audits they conduct. This also applies to audits carried out by NIK, especially performance audits. The issue is presented from a perspective of an auditor and engineer simultaneously, rather than from a purely academic one, and the proposed implementation is based on tools that are widely known to the community of auditors.

Eng. WIESŁAW KARLIŃSKI, PhD, specialist in the field of audit methodology and the application of analytical methods

Key words: data clustering, use of data clustering in auditing, data segmentation methods in auditing, anomaly identification, identification of risk areas, audit data analytics