OmniaScience

Journal of Industrial Engineering and Management
JIEM, 2024 – 17(3): 641-649 – Online ISSN: 2013-0953 – Print ISSN: 2013-8423
https://doi.org/10.3926/jiem.6791

A Review of Unsupervised K-Value Selection Techniques in Clustering Algorithms

Ana Pegado-Bardayo*, Antonio Lorenzo-Espejo, Jesús Muñuzuri, Alejandro Escudero-Santana

Universidad de Sevilla (Spain)

*Corresponding author: apegado@us.es
alorenzo@us.es, munuzuri@us.es, alejandroescudero@us.es

Received: October 2023
Accepted: May 2024

Abstract:

Purpose: Automatic grouping of data according to certain characteristics is made possible by clustering algorithms, which makes them an essential tool when working with large datasets. However, although they are unsupervised tools, they generally require the specification of the number of clusters to be formed, *k*, a task that may be simple for a human, but quite complex to automate. Despite the most commonly used k-value selection techniques offer acceptable results, they are not without shortcomings, suggesting that there is ample room for improvement. This paper briefly introduces clustering techniques, discusses the main shortcomings of conventional k-value selection techniques and examines the advantages and limitations of nine promising alternatives presented in recent years.

Design/methodology/approach: An evaluation of the main shortcomings of classic *k*-value estimation techniques has been carried out, and the newest proposals have been explained and compared.

Findings: New *k*-value estimation indices and methodologies proposed by authors guarantee better results, extending the use of these techniques to large volumes of data, and complex shapes and structures. However, no generical methodology able to overcome all the described shortcomings has still been developed.

Research limitations/implications: This research is limited to the newest developed techniques for *k*-value estimation, including proposals published since 2019. Older proposals have not been considered, as the newest ones overcome the former's shortcomings. A *k*-value estimation techniques review published in 2019 is cited in the test as a base reference.

Practical implications: Although the examples listed in the text apply to industry, the techniques described and discussed in this review are applicable to any area of science that can benefit from the use of clustering techniques.

Originality/value: To date, there has been no paper comparing the new *k*-value estimation techniques. Although there are literature reviews comparing the classical methods, these methods are nowadays nearly obsolete due to the complexity of the data usually faced.

Keywords: clustering, k-means, unsupervised learning, k-value

To cite this article:

Pegado-Bardayo, A., Lorenzo-Espejo, A., Muñuzuri, J., & Escudero-Santana, A. (2024). A review of unsupervised k-value selection techniques in clustering algorithms. *Journal of Industrial Engineering and Management*, 17(3), 641-649. https://doi.org/10.3926/jiem.6791

1. Introduction

Clustering is a key machine learning process that aims to group sets of unlabeled objects according to their characteristics, in order to build subsets of data known as clusters. Each cluster is formed by a collection of data that, in terms of the considered features, are similar to each other and differ from the rest of the data belonging to the dataset. Specifically, the features of the observations are represented numerically. Therefore, the similarity between two points can be measured as the distance between them (e.g. Euclidean distance). Thus, clustering algorithms will attempt to group observations in such a way as to maximize the similarity between group members as well as the difference with members of other groups.

Data availability and quality keep increasing due to technological advancements, automation, and the pervasive use of interconnected devices. However, large volumes of data are not useful if they cannot be easily managed and if conclusions cannot be drawn from them. Hence, clustering techniques are essential in data mining, enabling the handling of substantial amounts of data based on common characteristics. As an example of this, numerous developments can be found in recent literature in which data clustering is a necessary tool within the research, covering industry-related areas such as risk and quality assessment (Er-Kara, Oktay-First & Ghadge, 2020; Orak, Akkoyunlu & Can, 2020), logistics (Pegado-Bardayo, Lorenzo-Espejo, Muñuzuri & Aparicio-Ruiz, 2023), or production optimization (Hong, Lee, Cho, Jang & Kim, 2023), among others.

As this tool finds application across various fields, the characteristics of different datasets and the needs of data scientists can vary significantly, giving rise to the development of numerous clustering algorithms. These algorithms have been conventionally classified according to whether they employ a partitional approach, in which observations are segregated into previously specified number of groups, or a hierarchical strategy, in which clusters are created iteratively, either by building them from individual observations and merging them into larger clusters, or by dividing a cluster containing the entire set into smaller clusters until individual clusters are obtained (Saxena, Prasad, Gupta, Bharill, Patel, Tiwari et al., 2017). This means that, while in the latter approach the final output consists of dendrograms expressing relationships between all observations in the dataset, in the former one observations are assigned exclusively to one cluster.

One of the most widely used clustering algorithms in machine learning and data analysis is the K-means clustering algorithm, due to its simplicity, ease of implementation, and computational efficiency. K-means is a partitioning technique aimed at dividing a dataset into k distinct, non-overlapping subsets. Grouping is done by minimizing the sum of distances between each object and the centroid of its cluster. The naïve version of this algorithm, proposed by Lloyd (1982), follows the described steps:

1. Initialization: k points are initially placed in the data domain (centers), either randomly or following an initialization method.
2. The Voronoi diagram of the k sites is computed and all points inside each cell are assigned to its corresponding center.
3. The center of each Voronoi cell is substituted by the mean value of points corresponding to that cell.

Steps 2 and 3 are repeated until a stopping criterion is met, usually when a number of iterations is reached. However, the algorithm has converged when the assignments no longer change.

This algorithm has given rise to extensions that seek to adapt the technique to more complex or extensive datasets, such as BFR (Bradley, Fayyad & Reina, 1998) or Fuzzy C-means algorithms. Partitional algorithms also encompass another widely used group of medoid-based algorithms, such as PAM (Kaufman & Rousseeuw, 1990), CLARA (Kaufman & Rousseeuw, 1990) or CLARANS (Ngand & Han, 2002). Unlike K-means, where cluster centers are represented by the mean value of data points in each cluster, in K-medoids cluster centers are actual data points, specifically the most centrally located or "medoid" point within a cluster. Among them, PAM algorithm (Partitioning Around Medoids) represents the simplest approach, which proceeds with the following steps:

1. Initialization: k points of the dataset are initially placed as medoids.
2. All remaining points are assigned to their closest medoid.

3. For each medoid *m*, for each non-medoid *o*:
    i) Swap *m* and *o*, and recalculate the cost function (sum of the distance of the points to their medoids).
    ii) If the total cost of the configuration increased in the previous step, undo the swap.

Step 3 is repeated until no improvement is achieved in the objective function.

Clustering tasks with popular algorithms such as k-means and k-medoids are essential in data analytics and are considered unsupervised tasks. However, they require the specification of the number of clusters to be formed a priori, and said value directly affects the result. This can become a problem when dealing with large volumes of data, which is often the case due to the very purpose of these techniques.

There are several approaches in the literature for identifying the optimal number of clusters, but there is still much room for improvement. The classical approach to estimating this *k* value involves performing a brute-force search. The first step is to establish a range of variation of *k*, i.e., all the values considered for *k*. Then, the data is clustered for each value within the range. Finally, the accuracy of the result is evaluated by using a clustering validation index, inferring the final value of *k* according to the score evaluation.

The main disadvantage of this technique lies in its computational cost as the algorithm has to be run as many times as numbers are contemplated in the range of *k*. This range should not be too small, as many options would be left unexplored, nor too large, as this may imply high execution times.

Also, the commonly used indices to evaluate each iteration show weaknesses and are sometimes not sophisticated enough to achieve good results on complex datasets. Yuan and Yang (2019) lists some of the most widespread classical validation indices used for estimating this parameter. The main ones currently in use are the Silhouette Score (S), Elbow Method, Gap Statistic, and Calinski-Harabasz (CH), according to the authors.

All of them offer acceptable results, but they fail to solve the problem satisfactorily with certain point distributions. Observations in a dataset will show both similarities and differences among them, generating multiple clusters. In ideal cases, the clusters will have clearly differentiated, however there may be instances in which two or more clusters have fuzzy boundaries (overlapping), or in which a cluster encompasses different "subgroups" (hierarchy), as illustrated in Figure 1. Cluster overlapping and hierarchy. When these situations occur, the classical indices tend to underestimate the number of clusters, resulting in a loss of information that may be relevant in future analyses.

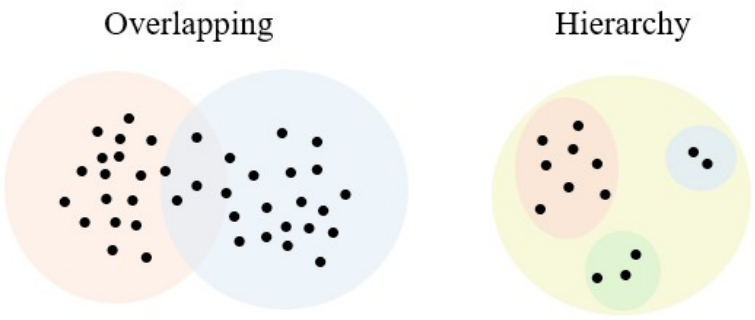

Figure 1. Cluster overlapping and hierarchy

Moreover, traditional indices show high sensitivity to outliers. Outliers are a common occurrence in real-world datasets, referring to observations significantly different from the rest of the data which can result from measurement errors, experimental variability, or genuine anomalies in the data distribution. Guerra, Robles, Bielza and Larrañaga (2012) compare in their study the performance of six classic indices in this situation, concluding that in the presence of average levels of outliers (5% of the total observations), none of them reaches an acceptable performance.

These, together with the computational complexity of this technique, have encouraged researchers to design alternatives for this process, which is essential for an appropriate handling of data.

This paper gathers more sophisticated methodologies proposed in recent years that attempt to address the weaknesses of the classical techniques. The following section compiles novel indices and strategies to estimate this parameter, and discussion and conclusions are given in Section 3.

2. Trends in *K*-value Estimation Methodologies

New developments in cluster validation indices as well as new methodologies designed to estimate the number of clusters are presented hereafter.

2.1. New Cluster Validation Indices

This section includes the most relevant studies focused exclusively on the improvement of classic validation indices or the design of new ones.

Some commonly used scores do not perform as desired when clusters "shapes" are not perfectly defined. The Silhouette score (Rousseeuw, 1987), for example, can be affected by irregular distances between clusters, while the Calinski-Harabasz index (Calinski & Harabasz, 1974) is highly sensitive to outliers. The proposal by Wang and Xu (2019) tries to solve these fluctuations by identifying peak points in Silhouette and Calinski-Harabasz indices and combining them into a single metric called Peak Weight Index (PWI), which seeks to balance both indexes by applying weights. The research shows promising results; however, it cannot be considered fully unsupervised as peak boundaries need to be specified according to the distribution of the dataset.

Yang, Lee, Choi and Joo (2020) focuses his research on the Gap statistic (Tibshirani, Walther & Hastie, 2001) as one of the most reliable indices in the estimation of k and describes the main weaknesses in its performance in order to improve it. The identified shortcomings are cluster overlapping, hierarchy within a cluster (i.e., clusters that may be formed by two or more smaller clusters, which classical methodologies are unable to identify), and high computational time. To overcome the first two aspects, the research proposes a new metric that evaluates the evolution of the Gap value with k, based on the premise that the Gap(k) function increases with constant or accelerated speed as k value is incremented, up to the point where k reaches its optimal value. At that point, the Gap(k) value suddenly decreases its increasing speed or starts slowing down. This deceleration of the Gap statistic (Dacc statistic) is calculated as follows:

$$Dacc(k) = [Gap(k) - Gap(k-1)] - [Gap(k+1) - Gap(k)] \quad (1)$$

However, statistical measurements sometimes fail to obtain proper results when data is not clearly separated or presents asymmetries. Aiming to overcome this, Rojas-Thomas, Santos and Mora (2017) propose an index adapted to real data patterns, based on the inner cohesion of clusters and the distance to others. The methodology divides clusters into sub-clusters based on the Principal Component Analysis (PCA), and the minimum spanning tree is obtained from the resulting centroids. Here, the concept of cohesion is introduced. To measure the cohesion between two adjacent sub-clusters in the spanning tree, the arc that connects them is evaluated: the greater the dispersion of data in the center of this arc, the lower the cohesion between these sub-clusters is (and vice versa). Finally, the cluster validation index is constructed by combining the distance between all the clusters being evaluated. To assess the distance between two clusters, the closest pair of sub-centroids (one from each cluster) is searched according to Euclidean distance, and the distance between them is calculated by adding the costs of the spanning tree branches joining them. In terms of scalability, the experimental results show that, as the number of clusters increases, the index's performance level decreases.

Also based on distance concepts, Abdalameer, Alswaitti, Alsudani and Isa (2022) present a novel index, according in this case to Euclidean distances. Two features are considered when evaluating clustering accuracy in this index: the distance between each point within a cluster to its centroid, namely Distance Within Cluster (DWC), and the Distance Between Centroids (DBC). Thus, good clustering will minimize DWC while maximizing DCB. Applying

this concept, the authors design a new metric, namely Validity Clustering Index based on Mean of clustered Data (VCIM), that aims to achieve more accurate and computationally cheaper estimations of *k*, obtained as:

$$VCIM = 1 - \left(\frac{1}{\exp\left(\exp(DBC_{total} - DWC_{total}\right)}\right) \quad (2)$$

Where and DWC_{total} represent the overall DBC and DWC for the dataset respectively at each iteration. The use of this metric is therefore limited to clustering algorithms that use Euclidean distances, but in these cases the results obtained are satisfactory, and better than those obtained with classical metrics in terms of accuracy.

Xie, Lawniczak and Gan (2022) try to solve the *k*-value underestimation of classic algorithms with an effective modification in the standard iterative method using Gap statistic. As previously discussed, the Gap statistic stands out among the classic indices for being the most sophisticated, however, there are two reasons that lead this technique to estimate low *k* values. These are the standard deviations of the data to be clustered, and the local fluctuations of this statistic, which have a direct impact on the evolution graphic and, therefore, on the estimation of k. To address this, the study proposes to smooth the curves of this graph. In order to do so, the authors propose benefiting from the power-law relationship between the Gap value and k, so that the derivative of the smooth function can be used to approximate the differential of gap statistics. This smooth curve allows overcoming the aforementioned fluctuations, thus avoiding the underestimation of clusters of this statistic.

Finally, a validity index based on the point pairs with fewer shared nearest neighbors (ANCV) is proposed by Duan, Ma, Zhou, Huang and Wang (2023), following the mentioned approaches that evaluate compactness within clusters and separation between clusters.

To calculate this index, an initial search is carried out to identify small, loose clusters within actual clusters, and their compactness is obtained as an indicator for the entire cluster. Consecutively, the average distance between pairs of data points at the intersection of two clusters is used as the between-cluster separation, making the index performance less influenced by the cluster shape. These measurements are obtained using equations (3) and (4):

$$COM = \frac{1}{K}\sum_{i=1}^{K} com(c_i) \quad (3)$$

$$SEP = \frac{1}{K-1}\sum sep(c_m, c_n) \quad (4)$$

Where *K* is the number of clusters formed, $com(c_i)$ is the within-cluster compactness for cluster *i*, and $sep(c_m,c_n)$ is the average distance between all pairs of between cluster augmented non-shared nearest neighbors. Both compactness and separation are combined in the final index ANCV.

$$ANCV = SEP - COM \quad (5)$$

Experiments show the best performance against the classic indices; however, this index quality may fail when the clustering results are incorrect for the actual number of clusters, and thus, authors consider improving this index to achieve optimal results in all different clustering situations.

2.2. New *K*-value Selection Methodologies

Classical methods are computationally expensive, mainly because the clustering algorithms are run iteratively for the whole range of possible *k* values. Therefore, authors have shown interest in creating new alternatives to these methodologies but still offering competitive results.

Computational complexity problem is compounded when talking about big data. Alibuhtto and Mahat (2020) present a local search algorithm to find a local optimum based on distances between centroids in big data. The main point of this technique is to establish a stop criterion. The research proposes an estimation of a threshold value based on Euclidean Distance so that the clustering algorithm is run until an acceptable value is found. This technique avoids evaluating the clustering over the whole range of possible k values, thus streamlining of the handling of large volumes of data.

Also trying to solve this problem, Ri, Kang, Kim, Choe and Han (2022) propose the Ratio of Variance to Range (RVR) and Dispersion-Width Ratio (DWR) separation measurement metrics as key to identifying different populations in a dataset. The authors performed Montecarlo simulations to study the behavior of DWR, initially on samples from a single population (cluster) and subsequently adding different populations. The analysis of the evolution of the DWR value revealed that for each new cluster, it is possible to observe a boundary in the graph, which allows the estimation of k. The results are promising, as this technique is able to reduce the runtime considerably, but still requires improvements to achieve more reliable results on some types of data such as heterogeneous distribution, sparse, and abnormal data.

Lastly, trying to address the inefficiency in execution times in conjunction with the accuracy of the results, Patil and Baidari (2019) propose to estimate k based on "depth difference" (DeD), following a similar approach to the exposed by (Abdalameer et al., 2022) in the previous section. Data depth measures a median in a multi-variate dataset, which is considered the deepest point in the given dataset. This metric assigns values from 0 to 1 to each point in the dataset according to their centrality. Then, the aforementioned Distance Between Clusters and Distance within Cluster are obtained and averaged, and finally, DeD is calculated as the difference among them.

Unlike classical methods, DeD does not employ any clustering algorithm for partitioning data, but rather iterates on the function itself, achieving significantly lower run times while achieving more accurate results than those obtained with indices such as Calinski and Harabasz (1974), Krzanowski and Lai (1988) Silhouette, and Gap.

3. Discussion and Conclusions

The increasing availability of data due to the growing presence of technologies in daily tasks enables its use and exploitation for several multidisciplinary purposes. However, the handling of large data volumes is challenging and clustering techniques are usually required in order to group data according to designed characteristics. This article reviews trends and new developments in unsupervised methodologies for estimating the optimal number of clusters in a dataset. There are widespread simple methods that present acceptable results in some cases, but they show numerous shortcomings.

Table 1 summarizes all approaches reviewed in this article, including the identified problems and each technique's limitations. Note that (I) and (M) correspond to new Index and Methodology, respectively.

Six novel indices and three methodologies are discussed in this paper, which evidence those mentioned shortcomings. These novel improvements get closer to an optimal solution to this problem, offering new approaches that mainly speed up the execution time and/or offer more accurate results, overcoming obstacles such as the underestimation of clusters, hierarchy within clusters, detection of small clusters or fuzzy shapes or even classical issues in real datasets such as the presence of outliers.

The findings of this review highlight the absence of a universally applicable approach to the aforementioned challenges: after a global comparison, it is observed that in order to solve the identified shortcomings it is necessary to trade-off either their accuracy, running time, or applicability. The latest advances in this field significantly facilitate and enhance these estimations considerably, and thus, although there is no global solution, data scientists can refer to the table to identify the approach that best aligns with their needs based on their dataset's characteristics.

Moreover, the table shows that there are still some aspects that can be improved and limitations that suggest that further advancements and refinements are still possible, underscoring the value and potential of this field of study.

Reference	Identified problem	Solving methodology	Limitations
Abdalameer et al. (2022)	Incorrect centroid position, computation inefficiency, and low accuracy in complex datasets	(I) Index based on the mean values of Distance Within Cluster and Distance Between Clusters (VCIM)	Specific for Euclidean-based clustering algorithms
Alibuhtto & Mahat (2020)	Inefficiency of classic methods in big data	(M) Proposes a local search and stopping criterion	Only valid in big bata. In low-medium datasets, the technique may underestimate k value. Data must be numerical
Duan et al. (2023)	Inaccuracy when identifying low-density clusters and abnormal shapes	(I) Index based on the shared neighbors between pairs of data	Sensible to low-quality clusterizations
Patil & Baidari (2019)	Computational cost and accuracy of classic methods	(M) Infers k-value by observing the evolution of Depth Difference function	Requires the specification of a range of k-values a priori
Rojas-Thomas et al. (2017)	Limitations of statistical-based indices when working with real data.	(I) Evaluates the distance between clusters and inside of them by assessing the cohesion (and discohesion) among them.	Performance degradation when the number of clusters increases
Ri et al. (2022)	Computational cost	(M) Obtains k-value by observing the evolution of Dispersion Width Ratio (DWR)	Limited performance in some datasets such as heterogeneous distributions, sparse, and abnormal data.
Wang & Xu (2019)	S and CH fluctuations due to data distribution	(I) Combines weighted peak values of S and CH into a single index (PWI)	Index boundaries to identify peak points need to be selected according to the distribution of the data set.
Xie et al. (2022)	K-value underestimation	(I) Proposes a function to overcome fluctuations of classic Gap statistic (Smoothing Gap)	Computationally inefficient
Yang et al. (2020)	Cluster overlapping and hierarchy within a cluster	(I) Creates an index that assesses the evolution of Gap statistic value (Dacc statistic)	Tested only in synthetic datasets

Table 1. Comparative table of new methodologies and indices to estimate *k* value

Declaration of Conflicting Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

The publication of this paper has been financed by project CAROLUM (PID2021-125125OB-I00), funded by MICIU/AEI/10.13039/501100011033 and the ERDF, UE. The research involved was also financially supported by project TRACSINT (P20_01183), funded by the Consejería de Economía, Conocimiento, Empresas y Universidad of Andalusia, and by the Ministry of Universities of Spain through the grant for the Training of University Researchers (Ayuda para la Formación del Profesorado Universitario, reference FPU20/05584).

References

Abdalameer, A., Alswaitti, M., Alsudani, A., & Isa, N. (2022) A new validity clustering index-based on finding new centroid positions using the mean of clustered data to determine the optimum number of clusters. *Expert Systems with Applications,* 191, 116329. https://doi.org/10.1016/j.eswa.2021.116329

Alibuhtto, M., & Mahat, N. (2020) Distance based k-means clustering algorithm for determining number of clusters for high dimensional data. *Decision Science Letters,* 9, 51-58. https://doi.org/10.5267/j.dsl.2019.8.002

Bradley, P.S., Fayyad, U., & Reina, C. (1998) *Scaling EM (Expectation- Maximization) Clustering to Large Databases.* Technical Report MSR-TR-98-35. Microsoft Research.

Calinski, T., & Harabasz, J. (1974) A dendrite method for cluster analysis. *Communications in Statistics,* 3(1), 1-27. https://doi.org/10.1080/03610927408827101

Duan, X., Ma, Y., Zhou, Y., Huang, H., & Wang, B. (2023) A novel cluster validity index based on augmented non-shared nearest neighbors. *Expert Systems With Applications*, 223, 119784. https://doi.org/10.1016/j.eswa.2023.119784

Er-Kara, M., Oktay-Firat, S.Ü., & Ghadge, A. (2020) A datamining-based framework for supply chain risk management. *Computers and Industrial Engineering,* 139, 105570. https://doi.org/10.1016/j.cie.2018.12.017

Guerra, L., Robles, V., Bielza, C., & Larrañaga, P. (2012). A comparison of clustering quality indices using outliers and noise. *Intelligent Data Analysis,* 16(4), 703-715. https://doi.org/10.3233/IDA-2012-0545

Hong, S., Lee, J., Cho, H., Jang, K., & Kim, J. (2023) Cluster-Based Multiobjective Particle Swarm Optimization and Application for Chemical Plants. *International Journal of Intelligent Systems,* 2023, 5275262. https://doi.org/10.1155/2023/5275262

Kaufman, L., & Rousseeuw, P.J. (1990) Partitioning Around Medoids (Program PAM), *Wiley Series in Probability and Statistics,* 68-125. https://doi.org/10.1002/9780470316801.ch2

Krzanowski, W.J., & Lai, T. (1988) A Criterion for Determining the Number of Groups in a Data Set Using Sum-of-Squares Clustering. *Biometrics,* 44(1), 23-34. https://doi.org/10.2307/2531893

Lloyd, S.P. (1982) Least square quantization in PCM. *IEEE Transactions on Information Theory,* 28(2), 129-137. https://doi.org/10.1109/TIT.1982.1056489

Ngand, R., & Han, J. (2002) CLARANS: A method for clustering objects for spatial data mining. *IEEE Transactions on Knowledge and Data Engineering*, 14(5), 1003-1016. https://doi.org/10.1109/TKDE.2002.1033770

Orak, E., Akkoyunlu, A., & Can, Z.S. (2020) Assessment of water quality classes using self-organizing map and fuzzy C-means clustering methods in Ergene River, Turkey. *Environmental Monitoring and Assessment,* 192(10), 638. https://doi.org/10.1007/s10661-020-08560-3

Patil, C., & Baidari, I. (2019) Estimating the Optimal Number of Clusters k in a Dataset Using Data Depth. *Data Science and Engineering,* 4, 132-140. https://doi.org/10.1007/s41019-019-0091-y

Pegado-Bardayo, A., Lorenzo-Espejo, A., Muñuzuri, J., & Aparicio-Ruiz, P. (2023) A data-driven decision support system for service completion prediction in last mile logistics. *Transportation Research Part A: Policy and Practice,* 176, 103817. https://doi.org/10.1016/j.tra.2023.103817

Ri, Y., Kang, C., Kim, K., Choe, Y., & Han, U. (2022) A New Method to Determine Cluster Number Without Clustering for Every K Based on Ratio of Variance to Range in K-Means. *Mathematical Problems in Engineering,* 2022, 6866747. https://doi.org/10.1155/2022/6866747

Rojas-Thomas, J.C., Santos, M., & Mora, M. (2017) New internal index for clustering validation based on graphs. *Expert Systems with Applications,* 86, 334-349. https://doi.org/10.1016/j.eswa.2017.06.003

Rousseeuw, P.J. (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics,* 20, 53-65. https://doi.org/10.1016/0377-0427(87)90125-7

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A. et al. (2017). A review of clustering techniques and developments. *Neurocomputing,* 267,664-681. https://doi.org/10.1016/j.neucom.2017.06.053

Tibshirani, R., Walther, G., & Hastie, T. (2001) Estimating the Number of Clusters in a Data Set via the Gap Statistic, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 63(2), 411-423. https://doi.org/10.1111/1467-9868.00293

Wang, X., & Xu, Y. (2019) An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index. *IOP Conference Series: Materials Science and Engineering,* 569(5), 052024. https://doi.org/10.1088/1757-899X/569/5/052024

Xie, S., Lawniczak, & A., Gan, C. (2022) Optimal number of clusters in explainable data analysis of agent-based simulation experiments. *Journal of Computational Science*, 62, 101685. https://doi.org/10.1016/j.jocs.2022.101685

Yang, J., Lee, JY., Choi, M., & Joo, Y. (2020). A New Approach to Determine the Optimal Number of Clusters Based on the Gap Statistic. In: Boumerdassi, S., Renault, É., & Mühlethaler, P. (Eds.), *Machine Learning for Networking (MLN)* (227-239). Springer, Cham. https://doi.org/10.1007/978-3-030-45778-5_15

Yuan, C., & Yang, H. (2019) Research on K-Value Selection Method of K-Means Clustering Algorithm. *J – Multidisciplinary Scientific Journal,* 2(2), 226-235. https://doi.org/10.3390/j2020016