Discover Artificial Intelligence

BRIEF REPORT Open Access

The Gap Statistic can be misleading when used to evaluate near box shaped clusters in the Euclidean space

Elisa Maria Merigo[1], Alessandro Anfossi[1] and Davide Chicco[1,2*]

*Correspondence:
Davide Chicco
davide.chicco@unimib.it
[1]Università di Milano-Bicocca, Milan, Italy
[2]University of Toronto, Toronto, ON, Canada

Abstract

The Gap Statistic is a metric designed and employed for the internal assessment of results from clustering analyses. Despite its popularity, we noticed a series of unexpected behaviors of this coefficient in some specific contexts. We therefore designed this study to understand why the Gap Statistic can take on negative values and under what circumstances this occurs. To this end, we introduce the concept of cages (box-shaped rectangular clusters in the Euclidean space), and calculate the Gap Statistic on the results obtained by *k*-Means applied to them, using the R open source programming language. We provide a mathematical explanation of how rectangular clusters were used and the reasoning behind their choice, starting with the original formula for the Gap Statistic. The results we obtained are inconsistent with the interpretation of the Gap Statistic, which suggests that a negative value indicates overlapping clusters or the presence of outliers around them. In contrast, we implemented well-separated groupings with no data points in between and still obtained a negative value for this metric. Considering these results, the Gap Statistic cannot be considered a reliable and standard assessment score for clustering experiments, as its resultant value alone does not provide a clear and universal understanding of the data distribution, in some specific cases. In fact, negative values of the statistic may arise in well-separated rectangular clusters closed to each other, reflecting sensitivity to reference distribution geometry. We therefore advise readers to avoid placing trust in the Gap Statistic when it yields negative values, to avoid employing the Gap Statistic alone but rather using it alongside more reliable metrics, such as Silhouette coefficient, Davies-Bouldin index, and DBCV index.

Keywords Gap Statistic, Unsupervised machine learning, Clustering, Internal clustering metrics, Clustering internal results evaluation

1 Introduction

Machine learning is a branch of artificial intelligence that aims to create systems that can function by learning from past examples. These algorithms can be classified as supervised and unsupervised. In unsupervised learning, the machine receives input data without access to past results (as in supervised algorithms). The main goal of this learning

method is to enable the machine to perform tasks independently, without our supervision. The most common application of the unsupervised algorithm is clustering, the process of creating clusters from data given some specific correlations among them. A cluster is a natural grouping of similar data points, and its centre is called centroid. To measure the effectiveness of clustering, metrics that can be divided into internal and external metrics are used: the former are used to assess the quality of the clusters without reference to external data, while the latter compare the results obtained with a reference dataset.

External metrics include the popular Adjusted Rand Index [1], which can be considered a generalization of the Matthews correlation coefficient [2, 3], and other scores such as Normalized Mutual Information, Mutual Information, Homogeneity, Completeness, V-measure, Jaccard index [4]. Examples of traditional internal metrics for convex-shaped clusters include Calinski-Harabasz index [5], Dunn index [6], Davies-Bouldin index [7], Silhouette score [8], plus the Gap Statistic [9] treated here. More recent internal coefficients for concave-shaped clusters' evaluation include CVNN [10], CDbw [11, 12], DBCV [13, 14], LCCV [15], CVDD [16], VIASCKDE [17], DCSI [18], and DISCO [19]. In this study we focus on the Gap Statistic, by exploring its properties and describing its unexpected behavior in some specific geometrical scenarios.

Scientific literature review. Several studies on the Gap Statistic were published in the past. For instance, Jaekyung Yang et al. [20] address and resolve an open problem raised in the original Gap Statistic study [9]. At the conclusion of that study, Robert Tibshirani and coauthors noted that the Gap Statistic cannot be applied when clusters are not well separated. Jaekyung Yang et al. [20] propose a novel methodology, based on the deceleration of the Gap metric, which distinguishes between two scenarios: when there is minimal overlap between the clusters, and when the clusters overlap heavily.

The Gap Statistic is also utilized in studies that seek to compare the performance of various internal evaluation metrics. This metric has been shown to outperform others in identifying the correct number of clusters in a two-dimensional dataset [21]. However, when a significant number of outliers are present between two clusters, resulting in the formation of a third or fourth artificial cluster, the Gap Statistic is unable to accurately distinguish between them. In the study by Chunhui Yuan et al. [22], the Gap Statistic demonstrates comparable accuracy to that of the Elbow and Silhouette methods, but with significantly higher computational complexity and time. Consequently, its utilization is not recommended for large-scale datasets.

Suneel Kumar Kingrani et al. [23] expound on the notion that the Gap Statistic is widely regarded as a meticulously delineated metric for estimating the number of clusters, and is esteemed as one of the most accurate methodologies in scenarios devoid of any outliers. However, this regard is due to the fact that in the presence of small clusters or outliers, the metric tends to overestimate the number of clusters and lacks robustness and performance.

In numerous studies, modifications have been made to the Gap Statistic with the objective of enhancing performance. One such example is the study by Rosana Ribeiro and coauthors [24], which introduces the Temporal Gap Statistic (TGS), a method based on the original metric, but redesigned specifically for time series data. The objective of this study is to propose a clustering validation index that is specifically designed for time-dependent data. This approach is distinct from the conventional Gap Statistic in

that it utilises a temporal distance metric and medoid-based clustering. The primary benefit of this approach is that it facilitates the identification of the optimal number of clusters (k) for temporally dependent datasets.

Conversely, the study of Chinatsu Arima et al. [25] provides a detailed exposition on the Modified Fuzzy Gap Statistic. The standard Gap Statistic is not well suited to fuzzy k-Means, a variation of the classic k-Means algorithm in which each data point can belong to multiple clusters simultaneously. Consequently, the metric is modified to ensure compatibility with this fuzzy clustering approach. The resulting method has been demonstrated to be more accurate than the original Gap Statistic when applied to data containing noise and outliers. In Yan and Ye [26], however, two other types of Gap Statistic are introduced: the Weighted Gap and the DD-Weighted Gap. The former is more robust to variation in density between clusters than the original Gap, while the latter is important for identifying the point beyond which additional clusters no longer need to be added.

As has been stated on multiple occasions, the Gap Statistic is susceptible to the presence of outliers and overlapping clusters. To address this challenge, an optimized version of k-Means is employed [27], incorporating a refined Gap Statistic. This enhancement facilitates a more reliable selection of the number of clusters, even in the presence of noise or overlaps. In addressing the well-documented issues surrounding the Gap Statistic, modifications have been made to the formula of the metric itself. For instance, Mojgan Mohajer et al. [28] propose the removal of the logarithm from the Gap Statistic formula. This advice is due to the finding in other studies that the logarithmic version may overestimate k, especially in cases where clusters are imbalanced in terms of density. Following the application of both versions of the Gap Statistic (with and without the logarithm), we observed that the version proposed by the authors more accurately identifies the correct number of clusters in cases where the original version often fails.

The article by Iliyas Karim Khan et al. [29] proposes an improved version of the Gap Statistic, along with a novel approach that uses the k-Means algorithm to handle clusters with complex, elongated or curvilinear shapes.

Another study by Iliyas Karim Khan et al. [30] introduces the Enhanced Gap Statistic (EGS) method for determining the optimal number of clusters k in k-Means clustering. The conventional Gap Statistic compares the within-cluster dispersion of the observed data, w_k, with the expected within-cluster dispersion, W_k*, computed from a set of randomly generated reference datasets. However, when the reference datasets are not properly generated or when the variables are expressed in different measurement units, this comparison may be biased. To address these limitations, the proposed EGS method standardizes the reference datasets prior to computing the Gap Statistic. This standardization ensures a fair and consistent comparison between the within-cluster dispersions of the original and reference data, thereby reducing the influence of scale differences and improving the reliability of the estimated optimal number of clusters. That study, although interesting, is beyond the scope of our research.

To the best of our knowledge, no previous studies have addressed the issue of geometry-driven negative values of the Gap Statistic in well-separated clusters as presented in our work. Moreover, the focus of our study is not to propose a new rule for selecting the "best" number of clusters, but rather to investigate the meaning of negative values in the analyzed metric.

This study. Here we analyze one of the specific cases in which the Gap Statistic becomes problematic, namely, when the clusters are not circular in shape but rectangular. The objective of this study is to determine and explain the circumstances under which this metric assumes negative values, and the present article offers a mathematical exposition of the problem. Even though we present a special case with only $k = 2$ clusters, uniform square clusters, and clusters generated solely by the *k*-means method, we believe the problematic behavior of the Gap Statistic highlighted in this study should be kept in mind by anyone conducting clustering analysis.

2 Methods

The Gap Statistic is a metric introduced in the early 2000 s specifically to assess the results of clustering analyses and to estimate the best number of clusters in a dataset [9]. The Gap Statistic was designed to be applicable to any clustering algorithm, but for simplicity we consider here its use on the *k*-Means algorithm's results. To summarise how the statistic works, it identifies a domain of the dataset (that is, a region of space where the points are placed), then uses the Monte Carlo method for sampling reference datasets, and eventually applies the clustering algorithm to these new reference datasets. In fact, to see what dispersion would look like if there were no real clusters, the Gap Statistic generates many artificial reference datasets by Monte Carlo sampling from a null distribution that preserves the broad scale of the data but has no cluster structure [9].

Our analysis will be limited to the use of the *k*-Means algorithm with $k = 2$ clusters.

We report here the final computable formula of the Gap Statistic:

$$\mathrm{Gap}_n(k) = \frac{1}{B}\sum_{b=1}^{B} \log\left(W^*_{k,b}\right) - \log\left(W_k\right), \tag{1}$$

where:

- n denotes the total number of observations in the dataset.
- k denotes the number of clusters considered.
- W_k is the within–cluster dispersion obtained by clustering the observed data into k clusters, defined as

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} \sum_{i,i' \in C_r} \|x_i - x_{i'}\|,$$

where C_r is the r-th cluster and $n_r = |C_r|$.

- $W^*_{k,b}$ is the within–cluster dispersion computed on the b-th Monte Carlo reference dataset, obtained by applying k-means clustering with k clusters to data sampled from the reference distribution.
- B is the number of Monte Carlo reference datasets, where $\{Cl^*_{k,b}\}_{0 \le b \le B}$ if we look at Fig. 2.
- $\frac{1}{B}\sum_{b=1}^{B} \log\left(W^*_{k,b}\right)$ is a Monte Carlo estimate of the expected value $\mathbb{E}^*_n[\log(W_k)]$ under the reference distribution.
- The reference distribution is usually uniform over the minimal bounding rectangle (even if the original authors of the Gap Statistic do not mandate this choice [9]).

The parameter B denotes the number of reference data sets generated by Monte Carlo simulation, used to estimate the expected dispersion within the cluster under the null hypothesis of no underlying cluster structure. For each of the B datasets, a clustering algorithm (for example, *k*-means) is applied to compute the corresponding within-cluster dispersion $W^*_{k,b}$. The logarithmic mean of these values serves as a benchmark against which the observed dispersion W_k from the original dataset is compared, allowing for the assessment of the effectiveness and significance of the identified clustering.

After a detailed study of the behavior of the metric, we concluded that its limits are determined by the dataset to which it is applied. Using the typical cluster shape, the circle, and applying several tests, we observed that when the Gap value falls between negative infinity and zero, the data are not well separated, leading to the interpretation of negative values as indicating poor clustering. However, when the same tests were performed using a different cluster shape (the rectangle) we observed that, despite negative values of the metric, the clusters were well separated and free of intermediate outliers. This observation highlights a critical issue with the metric: the Gap Statistic can be considered misleading and uncertain in some cases.

To explain why the use of rectangles gives different results than circles, we have relied on mathematical analysis. Indeed, a critical point in the construction of the Gap Statistic is the choice of the domain over which the simulated data are generated. This aspect can lead to significant distortions in the value of the statistic, especially in cases that are structured but simple.

Consider a dataset uniformly generated in two disjoint squares of side ℓ, separated along the horizontal axis by a distance $d - 2l$. We call this the *two cages scenario* (Fig. 1). Applying *k*-Means with $k = 2$, the clustering is correct and detects the two cages as distinct clusters.

The Gap Statistic for $k = 2$ is given by:

$$Gap(2) = \frac{1}{B}\sum_{b=1}^{B}[log(W^*_{2,b})] - log(W_2) \tag{2}$$

Where $1/B\sum_{b=1}^{B}[log(W^*_{2,b})]$ is the mean of the dispersions in the two clusters found by *k*-Means ($k = 2$) in the artificial datasets generated using the uniform distribution within the rectangle $l \times d$. For sufficiently large n we can assume that our rectangle will be separated almost perfectly in half by *k*-Means (Fig. 2). From Eq. 2 we can conclude that what we are interested in, once we make sure that we have *n* very large and sufficient sample datasets, is the expected value of the distance between two points generated within one of the halves of the rectangle and the expected value of the distance

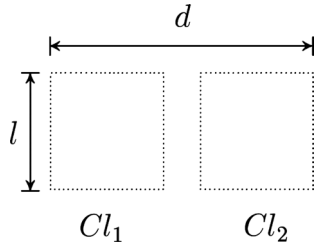

Fig. 1 Two-cage scenario. "Squares" of points laid with parallel sides along a line, where *d* and *l* represent the length of the sides of the smallest rectangle containing the dataset. Cl_1, Cl_2 are the clusters selected by a *k*-Means algorithm, $k = 2$

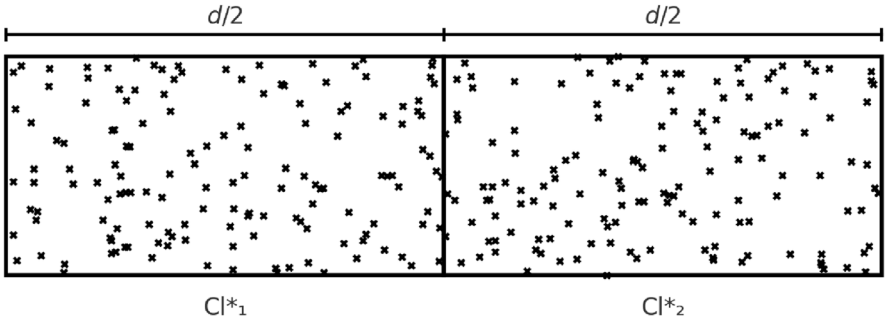

Fig. 2 Reference dataset. Example of a reference dataset generated using the Monte Carlo method, where Cl^*_1, Cl^*_2 are the "likely" clusters that *k*-Means, $k = 2$ would identify. Dividing the rectangle region into two smaller rectangles with a smaller side *d/2*

between two points generated on the perimeter of the corresponding square. This follows from the fact that the logarithm is an increasing monotone function and the W_k are essentially averages of the distances between pairs of points (we will call *dist*) in the individual clusters.

These two quantities give the following approximation, for a big enough dataset:

$$W_2 \approx 2 \cdot \mathbb{E}[dist] \tag{3}$$

From [31] we derive that the average distance between two points on the perimeter of a square is approximately:

$$\mathbb{E}[dist] \approx 0.7351 \cdot \ell \tag{4}$$

which applied to the latter formula (Eq. 3):

$$W_2 \approx 2 \cdot \mathbb{E}[dist] \approx 2 \cdot 0.7351 \cdot \ell \tag{5}$$

Furthermore, again from [31], we consider a rectangle with side lengths *a* and *b*, assuming without loss of generality that $b > a$. Let two points be chosen independently and uniformly at random within this rectangle. We are interested in computing the expected Euclidean distance between these two points, which we denote as $\mathbb{E}[dist^*]$. The result is expressed in closed form as:

$$\begin{aligned}\mathbb{E}[dist^*] = \Bigg\{ &\frac{a^3}{b^2} + \frac{b^3}{a^2} + \sqrt{a^2+b^2}\left(3 - \frac{a^2}{b^2} - \frac{b^2}{a^2}\right) + \\ &+ \frac{5}{2}\left(\frac{b^2}{a}\ln\left(\frac{a+\sqrt{a^2+b^2}}{b}\right) + \frac{a^2}{b}\ln\left(\frac{b+\sqrt{a^2+b^2}}{a}\right)\right)\Bigg\}/15\end{aligned} \tag{6}$$

For $a = \ell$ (Fig. 1) and $b = \frac{3}{2}\ell$ (where in Fig. 1 *d* is equal to $3 \cdot \ell$ then $3/2 \cdot \ell$ is *d*/2 in Fig. 2) we obtain

$$\mathbb{E}[dist^*] \approx 0.6585 \cdot \ell \tag{7}$$

And so, returning to the original formula of the Gap Statistic (Eq. 2) with an analogous reasoning we had with W_2 (Eq. 3), and considering an high enough number of Monte Carlo simulations in order to be able to apply the law of large numbers:

$$W_2^* \approx 2 \cdot \mathbb{E}[dist^*] \approx 2 \cdot 0.6585 \cdot \ell \tag{8}$$

Comparing $\mathbb{E}[dist^*]$ with the average distance $\mathbb{E}[dist]$ between two generic points on the perimeter of the square, it is smaller:

$$0.6585 \cdot \ell < 0.7351 \cdot \ell \tag{9}$$

which we can rewrite as

$$\mathbb{E}[dist^*] < \mathbb{E}[dist] \Rightarrow W_2^* < W_2 \tag{10}$$

Then from the previous observations

$$\frac{1}{B}\sum_{b=1}^{B}[log(W_{2,b}^*)] < log(W_2) \tag{11}$$

This means that the Gap Statistic yields negative values when both B (the number of Monte Carlo simulations) and n (the number of data points) are sufficiently large, since the empirical averages converge to their expected values. This shows that the negative outcome is not due to randomness, but is instead an inherent result of the underlying geometry of the dataset.

Now we want to show that the left element of (Eq. 2 is monotonically increasing with respect to the sides of the rectangle, and consequently we will have that as we approach between the two squares (keeping the side constant) the value of the Gap Statistic will only decrease. However, a much simpler heuristic reasoning of a geometric nature will suffice: if we extend the long side of the rectangle (b grows), we extend the range of possible distances between the two points and thus increase the average of the distances between the points inside the rectangle. We can therefore conclude that as the two squares get closer, the value of the Gap Statistic will only decrease (a "healthy" and "reasonable" behavior on the part of the metric).

This particular result depends on the choice of the bounding box null model, as suggested in the original formulation of the statistic [9]. We argue that this fundamental modeling choice introduces a structural limitation that should be explicitly acknowledged when applying the index.

We executed the present project in its entirety by utilising the open source R programming language, and made all our software code publicly available online on GitHub [32–34] ("Availability of software code" section) to grant full reproducibility [35]. Following a thorough evaluation of the available R packages, we decided to utilize the `index.Gap()` function of the `clusterSim` package [36] to calculate the Gap Statistic. Among the multiple software libraries employed, we also mention `ggplot2` [37] to produce the plots.

In the next section, we show the behavior of the Gap Statistic applied to the results of k-Means clustering obtained on artificial box-shaped data points.

3 Results

We show the results of the tests carried out on the cages in Figs. 3 and 4 through the k-Means method. The Gap Statistic value was consistently calculated using the k-Means algorithm with $k = 2$. In Fig. 3a, the two clusters are very close to each other, and calculating the value of the metric gives a negative result: -0.3969. In Fig. 3b, on the other hand, the cluster sizes have been reduced, increasing the distance between them. In this

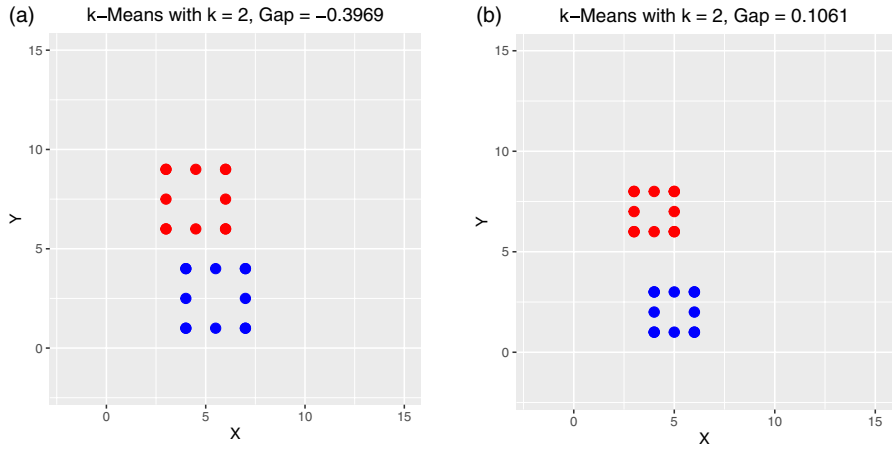

Fig. 3 Clustering application single example. Representation of points in the plane where we applied *k*-Means clustering with $k = 2$ and $B = 10$, with results coloured according to cluster membership and Gap Statistic calculated reported. Each cluster rectangle consists of 8 distinct points. In the **a** case the clusters are larger than in the **b** case

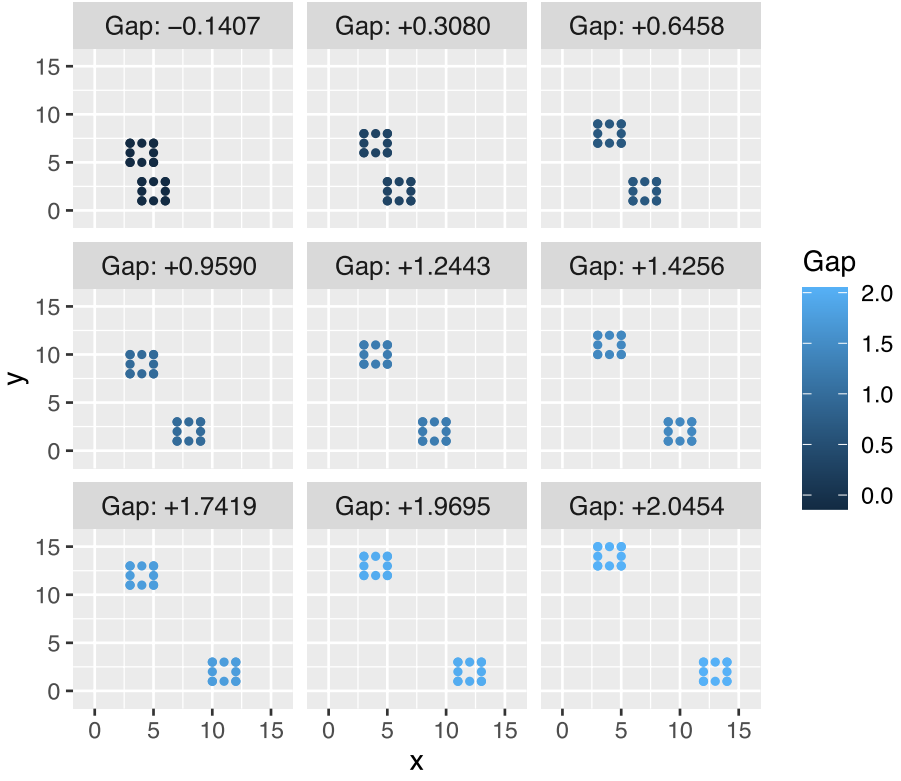

Fig. 4 Clustering application examples with sets of points moving apart. Representation of points in the plane where we applied *k*-Means with $k = 2$ and $B = 10$, and calculated the value of the Gap Statistic as the two clusters moved further apart. Each cluster rectangle consists of 8 distinct points. The colours of the two clusters reflect the resulting value of the metric, as indicated in the legend on the right

case, the Gap Statistic is positive: $+0.1061$. These results were unexpected. In fact, as the original article explains [9], if the clusters are well separated and free of outliers, the Gap Statistic is expected to be positive. In all the results reported, there are no outliers between the two clusters, as all the points belonging to each cluster lie on its perimeter.

By reducing the size of the two groups (and thus increasing the distance between them) the value of the metric considered becomes positive (Fig. 3b).

In order to determine whether this behavior of the metric is an isolated case or a consistent pattern, we created a grid illustrating nine different configurations of the clusters (Fig. 4), where in the first panel the two clusters are very close to each other and progressively move farther apart, reaching a state of maximum separation in the last panel. We calculated the Gap Statistic for each panel. The lowest value obtained was -0.1407, while the highest was $+2.0454$. In the first case, the two groups are close together (Fig. 3a). In the second case, they are slightly further apart: we added 1 to the y-coordinate of the first cluster and to the x-coordinate of the second. We repeated this procedure until, in the last case of the grid, the two clusters are clearly separated. After calculating the Gap Statistic for each configuration, we see that it starts with a negative value and gradually becomes positive until it reaches a value of 2 (Fig. 4). This behavior is in line with our expectations and therefore supports our hypothesis.

However, what remains unexpected is the negative value observed when the clusters are close but still well separated and free of outliers. In simple words, it is clear that the clusters identified by k-Means in Figs. 3 and 4 are well separated and, therefore, we expect the Gap Statistic to produce a positive value. That is not the case though.

To further verify whether the results shown in Fig. 4 were an isolated case, we repeated the same test ten times using ten different random seeds, and we reported the outcomes in Supplementary Figs. S1 and S2. The Gap Statistic values across all ten tests confirm the same trend observed in Fig. 4: an unexpected negative Gap value occurs when clusters are extremely close to each other yet still well-separated.

These negative Gap values seem to result from a mismatch between the support of the observed data and that of the reference distribution. This mismatch is inherent in the design of the Gap Statistic and is the key issue our study aims to highlight. The original authors of this score [9] deliberately focused on the within-cluster dispersion of the data, thereby shifting attention away from cluster shape and inter-cluster separation.

Even if the examples shown are theoretical, rectangular and box-shaped clusters can be commonly found in several real-world applications: for example in cluster analysis of geographic data [38], urban planning data [39], retail data [40], and environmental monitoring [41].

4 Discussion and conclusions

The Gap Statistic can be deceptive. The Gap Statistic was first introduced by Robert Tibshirani et al. [9] in the early 2000 s and has since been applied and adapted in various studies to assess the results of clustering tests and experiments. As previously mentioned, this metric is often used in comparison with other internal validation measures across different projects to identify the most reliable measure and better understand how they behave in diverse scenarios. In our case, the Gap Statistic produced results that, to our knowledge, had never been obtained or documented before. Specifically, when two clusters are clearly separated with no noise between them, the Gap Statistic performs well.. However, the metric tends to fail when the clusters are slightly or fully overlapping, often producing negative values that are clearly incorrect.

It is important to emphasise that, in this study, our focus was solely on the final value of the Gap Statistic rather than on the optimal number of clusters it suggests. The Gap

Statistic loses robustness and accuracy in the presence of outliers between clusters [23], which is a limitation that has been recognised since the Gap Statistic's original publication [9]. However, what we highlight here is a novel and previously unobserved behavior: when clusters are close to each other, neither overlapping nor separated by outliers, the metric unexpectedly yields a negative value despite such a configuration falling within the category of well-separated clusters.

In applied settings, a negative result can be misleading, especially in scenarios where a clearly optimal clustering solution exists. This situation highlights a misconception in the original authors' interpretation of what constitutes "good" clustering from a mathematical perspective. Serious problems might appear when the clusters are well separated but the Gap statistic has a negative value: just by looking at its value, a distracted reader might think the cluster analysis was of poor quality.

We illustrated this behavior in Fig. 3. To gain a more comprehensive understanding of the performance of the Gap Statistic, Fig. 4 offers further insight. Indeed, our analysis demonstrates that in instances where the two clusters are in close proximity to each other, the value of the metric in question becomes negative. As the separation between the two clusters increases, the metric exhibits a progressive increase in positivity. As previously documented [9], well-separated clusters have been shown to yield superior Gap Statistic performance in comparison to overlapping clusters or those surrounded by outliers. However, the aforementioned study implicitly suggests that a high Gap value requires the clusters not to be in contact.

This particular issue was the focus of our study. As demonstrated earlier (Figs. 3 and 4), the two clusters do not intersect; they are merely proximate, with no aberrant data points observed. However, the Gap Statistic, which is predicated on the assumption that physical separation alone is sufficient to produce a high Gap value, unexpectedly returns a negative value. We recognize that this issue is a special-case structural phenomenon rather than a universal limitation of the Gap Statistic, but we believe it is important to warn readers about this misbehavior of the score.

A potentially wrong estimate of the number of underlying clusters. As demonstrated by our findings, the Gap Statistic can yield negative values in clustering results containing two well-defined, rectangular-shaped clusters, thereby providing a misleading indication. This deceptive behavior of the Gap Statistic can lead to several adverse consequences, including the incorrect conclusion that a specific k value is not an appropriate number of clusters for a given scientific problem.

For instance, if one attempts to determine the optimal number of clusters by applying k-means with two clusters to a specific dataset and observes a negative Gap Statistic (as in Fig. 3a), they might be inclined to dismiss $k = 2$ as a viable option. However, as we have shown, the clusters identified could be well-separated, making two clusters a valid choice for that analysis. Yet, the Gap Statistic may mislead the clustering analyst into believing that k-means with 2 clusters has produced a poor result, potentially compromising the entire study.

Robert Tibshirani, Guenther Walther, and Trevor Hastie invented the Gap Statistic for estimating the (optimal) number of clusters in a dataset [9]. However, we have observed that it can fail in achieving this original goal when clusters are situated close to each other.

In conclusion, after several tests, we can say that the Gap Statistic can be an unreliable metric in some cases. In fact, the Gap Statistic may be sensitive to the geometry of the bounding reference distribution in specific configurations. It is unadvisable to rely only on its resulting value to determine the structure of the input dataset, as a negative value may misleadingly suggest the presence of numerous outliers, even when this is not the case, as demonstrated in this study. In a nutshell, the Gap Statistic is sensitive to geometric configurations, and negative values do not universally imply poor clustering.

We recommend to the readers to avoid using the Gap Statistic alone for clustering result assessment, to employ the Silhouette coefficient and the Davies-Bouldin index for convex-shaped clusters [42] and the DBCV index for concave-shaped clusters [13, 14] instead.

Limitations. Regarding limitations, we have tested this behavior only on artificial datasets shaped as squares, but not on other geometric shapes or real-world datasets. In this project, the *k*-Means algorithm was the sole algorithm we utilized, as the objective was to specify the number of clusters manually. Moreover, here we only used $k = 2$ for our tests and did not employ any other number of clusters. We intuitively believe that our claims should hold for any number of clusters, but we do not demonstrate this outcome in the present study.

Future developments. As future work, we plan to explore modifications to the original formulation of the Gap Statistic to investigate its behavior. A potential avenue for future research would be to utilise the DBSCAN algorithm [43] to analyse the results that it might produce. An initial scenario might involve the utilisation of DBSCAN as a preliminary processing step to identify outliers, followed by the implementation of *k*-Means and the Gap Statistic on the filtered dataset.

A further potential research direction could involve the utilisation of the Fuzzy Gap Statistic [25]. A particularly intriguing avenue for future research would be to analyse the behavior of this metric in scenarios where multiple points belong to more than one cluster. Furthermore, adopting the approach outlined by Mojgan Mohajer et al. [28], one could assess the efficacy of the Gap Statistic under diverse testing conditions, employing both the original formulation and its modified versions.

We also plan to develop a new study where to compare the Gap Statistic with BIC (Bayesian Information Criterion) [44], BKPlot (Best-K Plot) [45], and CCPI (Co-Cluster Profile Index) [46].

Some readers might be interested in exploring what happens when other null distributions are used to generate the reference datasets or when a similar setup is considered in higher dimensions. Mathematically, this aspect would present an interesting discussion. In higher dimensions, the curse of dimensionality greatly amplifies distances within the observed data, while the concentration of measure pushes the reference dataset increasingly toward the boundary of the hyperrectangle defined by the index–regardless of the chosen null distribution. However, the main focus of this study is to highlight the mismatch between the support of the observed data and that of the reference distribution, which we consider a structural limitation in the design of the Gap statistic. We will consider alternative null distributions and higher dimension setups in future developments.

Abbreviations

B	Number of Monte Carlo simulations
BIC	Bayesian Information Criterion
BKPlot	Best-K Plot

CCPI	Co-Cluster Profile Index
CDbw	Composed density between and within clusters
CI	Cluster
CVDD	Cluster validity index based on density-involved distance
CVNN	Clustering validation index based on nearest neighbors
DBCV	Density-based clustering validation
DBSCAN	Density-based spatial clustering of applications with noise
DCSI	Density cluster separability index
DISCO	Density-based internal score for clustering outcomes
dist	Distance
EGS	Enhanced Gap Statistic
k	Number of clusters
n	Number of data points
TGS	Temporal Gap Statistic
VIASCKDE	Validity index for arbitrary-shaped clusters based on the kernel density estimation

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1007/s44163-026-01195-2.

Supplementary file 1.

Acknowledgements
The authors thank the reviewers of the CIBB 2025 conference for their feedback.

Author contributions
E.M.M. performed the tests, conducted the literature review, prepared the plots, and contributed to writing the manuscript. A.F. wrote the mathematical analysis, conducted the literature review, and contributed to writing the manuscript. D.C. conceived and supervised the study, and contributed to writing the manuscript.

Funding
The work of D.C. is partially funded by the Italian Ministero Italiano delle Imprese e del Made in Italy under the Digital Intervention in Psychiatric and Psychologist Services (DIPPS) program (project code F/310240/01-04/X56), within the framework "Innovation Agreements" (Accordi per l'Innovazione) and is partially supported by the Ministero dell'Università e della Ricerca of Italy under the "Dipartimenti di Eccellenza 2023–2027" ReGAInS grant assigned to the Dipartimento di Informatica Sistemistica e Comunicazione at Università di Milano-Bicocca. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data availability
Our R software code and the generated data are available under the GPL–3.0 license at: https://github.com/Elisamerigo/GapStatisticArticle

Declarations

Ethics approval and consent to participate
Not applicable.

Consent to publication
All authors consent to the publication of the present article.

Competing interests
The authors declare that they have no competing interests.

Received: 13 October 2025 / Accepted: 20 March 2026
Published online: 15 April 2026

References

1. Rand WM. Objective criteria for the evaluation of clustering methods. J Am Stat Assoc. 1971;66(336):846–50. https://doi.org/10.1080/01621459.1971.10482356.
2. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochem Biophys Acta (BBA) – Protein Struct. 1975;405(2):442–51. https://doi.org/10.1016/0005-2795(75)90109-9.
3. Chicco D, Jurman G. An invitation to greater use of Matthews correlation coefficient in robotics and artificial intelligence. Front Robot AI. 2022;9:876814. https://doi.org/10.3389/frobt.2022.876814.
4. de Souto MCP, Coelho ALV, Faceli K, Sakata TC, Bonadia V, Costa IG. A comparison of external clustering evaluation indices in the context of imbalanced data sets. In: Proceedings of SBRN 2012 – the 2012 brazilian symposium on neural networks. IEEE; 2012. p. 49–54. https://doi.org/10.1109/SBRN.2012.25.
5. Calinski T, Harabasz J. A dendrite method for cluster analysis. Commun Stat Theory Methods. 1974;3(1):1–27. https://doi.org/10.1080/03610927408827101.
6. Dunn JC. Well-separated clusters and optimal fuzzy partitions. Cybern Syst. 1974;4(1):95–104. https://doi.org/10.1080/01969727408546059.

7. Davies DL, Bouldin DW. A cluster separation measure. IEEE Trans Pattern Anal Mach Intell. 1979;PAMI–1(2):224–7. https://doi.org/10.1109/tpami.1979.4766909.
8. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53–65. https://doi.org/10.1016/0377-0427(87)90125-7.
9. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. J R Stat Soc Ser B (Stat Methodol). 2001;63(2):411–23. https://doi.org/10.1111/1467-9868.00293.
10. Liu Y, Li Z, Xiong H, Gao X, Junjie W, Sen W. Understanding and enhancement of internal clustering validation measures. IEEE Trans Cybern. 2013;43(3):982–94. https://doi.org/10.1109/tsmcb.2012.2220543.
11. Halkidi M, Vazirgiannis M. Clustering validity assessment using multi representatives. In: Proceedings of SETN 2002 – the 2nd hellenic conference on artificial intelligence, p.237–249; 2002. http://lpis.csd.auth.gr/setn02/poster_papers/237.pdf.
12. Halkidi M, Vazirgiannis M. A density-based cluster validity approach using multi-representatives. Patt Recogn Lett. 2008;29(6):773–86. https://doi.org/10.1016/j.patrec.2007.12.011.
13. Moulavi D, Jaskowiak PA, Campello RJGB, Zimek A, Sander J. Density-based clustering validation. In: Proceedings of SDM24 – the 2014 SIAM international conference on data mining, p. 839–847. SIAM; 2014. https://doi.org/10.1137/1.9781611973440.96.
14. Chicco D, Sabino G, Oneto L, Jurman G. The DBCV index is more informative than DCSI, CDbw, and VIASCKDE indices for unsupervised clustering internal assessment of concave-shaped and density-based clusters. PeerJ Computer Science. 2025;11:e3095. https://doi.org/10.7717/peerj-cs.3095.
15. Cheng D, Zhu Q, Huang J, Quanwang W, Yang L. A novel cluster validity index based on local cores. IEEE Trans Neural Netw Learn Syst. 2018;30(4):985–99. https://doi.org/10.1109/tnnls.2018.2853710.
16. Lianyu H, Zhong C. An internal validity index based on density-involved distance. IEEE Access. 2019;7:40038–51. https://doi.org/10.1109/access.2019.2906949.
17. şenol. A. VIASCKDE index: a novel internal cluster validity index for arbitrary-shaped clusters based on the kernel density estimation. Comput Intell Neurosci. 2022;2022:1–20. https://doi.org/10.1155/2022/4059302.
18. Gauss J, Scheipl F, Herrmann M. DCSI – an improved measure of cluster separability based on separation and connectedness. 2023. arXiv:2310.12806.
19. Beer A, Krieger L, Weber P, Ritzert M, Assent I, Plant C. DISCO: internal evaluation of density-based clustering. 2025. arXiv:2503.00127.
20. Yang J, Lee J-Y, Choi M, Joo Y. A new approach to determine the optimal number of clusters based on the gap statistic. In Proceedings of MLN 2019 – the 2nd IFIP TC 6 international conference on machine learning for networking, volume 12081 of Springer Lecture Notes in Computer Science (LNCS), p. 227–239. 2020. https://doi.org/10.1007/978-3-030-45778-5_15.
21. Santos JM, Embrechts M. A family of two-dimensional benchmark data sets and its application to comparing different cluster validation indices. In: Proceedings of MCPR 2014 – the 6th mexican conference on pattern recognition, volume 8495 of Springer Lecture Notes in Computer Science (LNCS), p. 41–50; 2014. https://doi.org/10.1007/978-3-319-07491-7_5.
22. Yuan C, Yang H. Research on k-value selection method of k-means clustering algorithm. J. 2019;2(2):226–35. https://doi.org/10.3390/j2020016.
23. Kingrani SK, Levene M, Zhang D. Estimating the number of clusters using diversity. Artif Intell Res. 2017;7(1):15. https://doi.org/10.5430/air.v7n1p15.
24. Ribeiro RG, Rios R. Temporal Gap Statistic: a new internal index to validate time series clustering. Chaos, Solitons Fractals. 2021;142:110326. https://doi.org/10.1016/j.chaos.2020.110326.
25. Arima C, Hakamada K, Okamoto M, Hanai T. Modified fuzzy Gap Statistic for estimating preferable number of clusters in fuzzy k-means clustering. J Biosci Bioeng. 2008;105(3):273–81. https://doi.org/10.1263/jbb.105.273.
26. Yan M, Ye K. Determining the number of clusters using the weighted Gap Statistic. Biometrics. 2007;63(4):1031–7. https://doi.org/10.1111/j.1541-0420.2007.00784.x.
27. El-Mandouh AM, Abd-Elmegid LA, Mahmoud HA, Haggag MH. Optimized k-means clustering model based on Gap Statistic. Int J Adv Comput Sci Appl. 2019;10(1). https://doi.org/10.14569/ijacsa.2019.0100124.
28. Mohajer M, Englmeier K-H, Schmid VJ. A comparison of gap statistic definitions with and without logarithm function. Chaos, Solitons Fractals. 2010. https://doi.org/10.5282/ubm/epub.11920.
29. Khan IK, Daud HB, Zainuddin NB, Sokkalingam R, Abdussamad AM, Inayat A. Addressing limitations of the k-means clustering algorithm: outliers, non-spherical data, and optimal cluster selection. AIMS Math. 2024;9(9):25070–97. https://doi.org/10.3934/math.20241222.
30. Khan IK, Daud H, Zainuddin N, Sokkalingam R. Standardizing reference data in gap statistic for selection optimal number of cluster in K-means algorithm. Alex Eng J. 2025;118:246–60. https://doi.org/10.1016/j.aej.2025.01.034.
31. Mathai AM, Moschopoulos P, Pederzoli G. Random points associated with rectangles. Rendiconti del Circolo Matematico di Palermo. 1999;48:163–90. https://doi.org/10.1007/bf02844387.
32. Blischak JD, Davenport ER, Wilson G. A quick introduction to version control with Git and GitHub. PLoS Comput Biol. 2016;12(1):e1004668. https://doi.org/10.1371/journal.pcbi.1004668.
33. Perez-Riverol Y, Gatto L, Wang R, Sachsenberg T, Uszkoreit J, da Veiga F, et al. Ten simple rules for taking advantage of Git and GitHub. PLoS Comput Biol. 2016;12(7):e1004947. https://doi.org/10.1371/journal.pcbi.1004947.
34. Noble WS. A quick guide to organizing computational biology projects. PLoS Comput Biol. 2009;5(7):e1000424. https://doi.org/10.1371/journal.pcbi.1000424.
35. Sandve GK, Nekrutenko A, Taylor J, Hovig E. Ten simple rules for reproducible computational research. PLoS Comput Biol. 2013;9(10):e1003285. https://doi.org/10.1371/journal.pcbi.1003285.
36. Walesiak M, Dudek A. clusterSim: searching for optimal clustering procedure for a data set. *CRAN: Contributed Packages*. 2016. https://doi.org/10.32614/cran.package.clustersim.
37. Wickham H, Chang W, Henry L, Pedersen TL, Takahashi K, Wilke C, Woo K, Yutani H, Dunnington D, van den Brand T. ggplot2: elegant graphics for data analysis. *CRAN: Contributed Packages*. 2016. https://doi.org/10.32614/cran.package.ggplot2.
38. Han J, Lee J-G, Kamber M. An overview of clustering methods in geographic data analysis. Geogr Data Min Knowl Discov. 2009;2:149–70. https://doi.org/10.1201/9781420073980.

39. Fazlollahi S, Girardin L, Maréchal F. Clustering urban areas for optimizing the design and the operation of district energy systems. In *Computer Aided Chemical Engineering*, volume 33, p. 1291–1296. Elsevier; 2014. https://doi.org/10.1016/B978-0-444-63455-9.50050-7.
40. Holỳ V, Sokol O, Černỳ M. Clustering retail products based on customer behaviour. Appl Soft Comput. 2017;60:752–62. https://doi.org/10.1016/j.asoc.2017.02.004.
41. Mou W, Tan L, Xiong N. Data prediction, compression, and recovery in clustered wireless sensor networks for environmental monitoring applications. Inf Sci. 2016;329:800–18. https://doi.org/10.1016/j.ins.2015.10.004.
42. Chicco D, Campagner A, Spagnolo A, Ciucci D, Jurman G. The Silhouette coefficient and the Davies-Bouldin index are more informative than Dunn index, Calinski-Harabasz index, Shannon entropy, and Gap statistic for unsupervised clustering internal evaluation of two convex clusters. PeerJ Comp Sci. 2025;11:e3309. https://doi.org/10.7717/peerj-cs.3309.
43. Schubert E, Sander J, Ester M, Kriegel HP, Xiaowei X. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. ACM Trans Database Syst. 2017;42(3):1–21. https://doi.org/10.1145/3068335.
44. Neath AA, Cavanaugh JE. The Bayesian information criterion: background, derivation, and applications. Wiley Interdiscipl Rev Comput Stat. 2012;4(2):199–203. https://doi.org/10.1002/wics.199.
45. Chen K, Liu L. "Best K": critical clustering structures in categorical datasets. Knowl Inf Syst. 2009;20(1):1–33. https://doi.org/10.1007/s10115-008-0159-x.
46. Liu X, Lianyu H, Jiang M, He Z. Clustering validation via sample pair co-cluster testing. IEEE Trans Knowl Data Eng. 2025. https://doi.org/10.1109/TKDE.2025.3631331.

Publisher's Note