Optimal number of clusters to rank a model-based index

Mariaelena Bottazzi Schenone[1], Elena Grimaccia[2], Maurizio Vichi[1]

[1] Department of Statistical Sciences, Sapienza University, Rome (Italy)
[2] ISTAT - Italian National Institute of Statistics, Rome (Italy)

Abstract. This study proposes a new method to define the optimal number of groups in cluster analysis, in cases when the clusters' order is relevant. In this work, the clustering method of k-means is applied to a univariate index, resulting from a Structural Equation Model (SEM). In contrast to the majority of conventional procedures for choosing the number of clusters, the new methodology looks for the greatest number of clearly distinct clusters rather than a more parsimonious one. This method enables the construction of a granular ranking of the units in clusters starting from an index and minimizes the information loss caused by clustering with a low number of groups.
Indeed, the classification adds more information to the mere ordering of units: it aids in locating homogeneous groups of elements for which the index value can be considered the same. Namely, it helps in identifying units perceived as similar each other, which should be considered as "ties" in the ranking since they have substantially the same index value. This methodology works well when the goal is to rank units in groups from "the best" to "the worst", according to a particular measure.
The clusters' number has been chosen, considering the maximum number of significantly different clusters, according to the non-parametric Wilcoxon "rank-sum" test. Since there exists an ordering between clusters, the test compares each cluster with the closest one. An *"ad-hoc"* algorithm is proposed to define the ideal number of clusters.
In this paper, k-means clustering is applied to an index measuring air pollution across European urban areas. A clustering of cities for different air pollution levels is graphically represented. The analysis' results provide essential information to develop locally tailored policies aimed at the reduction of air pollution in metropolitan areas.

Keywords: Clusters ranking, Wilcoxon rank-sum test, Multidimensional index, Air Pollution, Metropolitan areas

1 Introduction

Cluster analysis is a technique used to identify meaningful groups within a data-set. Typically, it focuses on finding a parsimonious number of clusters that adequately represent the underlying data structure. However, in certain cases, the order of the clusters holds significance and there exists a ranking among them. In these cases, a tailored approach is required. This paper presents a novel procedure to tackle this problem. Indeed, the proposed method seeks to identify the maximum number of clearly distinct clusters. By doing so, the information loss due to clustering with few groups is minimized and a detailed ranking of units in clusters based on a specific index is obtained. In fact, this approach goes beyond simple unit ordering and enables the identification of homogeneous groups of elements within each cluster. All the elements in each cluster are perceived as similar and can be considered "ties" in the units' ranking. Therefore, our method is particularly valuable when the goal is to classify units from "the best" to "the worst", according to a given measure.

In this study, we apply the k-means clustering technique to a univariate index derived from a Structural Equation Model based analysis ([1], [2], [3]).
This index measures air pollution by simultaneously considering the six main pollutants more frequently addressed in the literature (PM10 and PM2.5, Sulphate Dioxide, Nitrogen Dioxide, Carbon Monoxide, and ground-level Ozone). Additionally, it includes significant determinants of air pollution, which better explain the pollutants levels. Our index aims at overcoming the limitations of monitoring individual gases and traditional composite indices, which often rely on strong normative assumptions and neglect the multivariate relationships among contaminants ([4]).
Using k-means clustering, we group cities based on their air pollution levels, as measured by the new index. We determine the optimal number of clusters by selecting the maximum possible number for which all the identified groups are distinct. This is assessed by the non-parametric Wilcoxon rank-sum test ([5]). The considered approach differs from commonly used methods, which prioritize identifying a more parsimonious number of clusters. Some of these methods are the gap statistics ([6]), the silhouette method ([7]), the elbow method ([8]) and the pseudoF method ([9]).
In the air pollution context, the obtained results are useful to design effective strategies to mitigate pollution and promote environmental well-being in urban settings. Indeed, the ordered clustering provides a basis for understanding the heterogeneity among cities and facilitates the identification of specific clusters with different degrees of air pollution.

The paper is structured as follows: Section 2 describes the databases used for the empirical study, Section 3 presents the adopted methodologies. In Section 4, an application on air quality in Europe is provided, while in Section 5 concluding remarks are drawn.

2 Data

The proposed methodology is applied to a data-set which refers to 130 metropolitan areas within the European Union. The cities included in the analysis are depicted in Fig. 3 and Fig. 4. The analysed data-set includes Worldwide Air Quality data from the Global Air Quality Index (AQI) platform (https://aqicn.org/). This global data-set provides standardized information on pollutants and atmospheric conditions worldwide ([10]). The six main pollutants have been included in the analysis (Carbon Oxide, Sulphur Dioxide, Particulate Matter 2.5, Particulate Matter 10, Nitrogen Dioxide and Ozone), together with the following meteorological and atmospheric covariates: air temperature, humidity, air pressure, wind-gust (m/s) and wind-speed (m/s) ([11]). In addition, to take into account the socio-economic features of the analysed cities, variables such as the GDP per capita, population density, elderly and youth dependency ratios, employment, unemployment and participation rates ([12], [13], [14]) have been included from the Organisation for Economic Co-operation and Development (OECD) Metropolitan database ([15]).
Traffic-related emissions have become a global environmental problem, most of all in urban areas ([16]). Therefore, the motorization rate (Number of passenger cars per thousand inhabitants, available at country level) and the Number of registered cars per 1000 population (at city level) have been included in the study from the Eurostat metropolitan regions (NUTS3) database ([17]).
Furthermore, two geographical covariates (Latitude and Longitude) have been considered in the analysis, to take into account the spatial configuration of the phenomenon of air pollution ([18], [19]).
To ensure the selection of the most meaningful and significant variables, a Multivariate Random Forest (MRF) regression ([20]) is employed to identify key predictors of the multivariate outcome,

which consists of the six pollutants. This rigorous variable selection process helps capture the complex relationships among gases while considering the specific social, demographic, economic, and meteorological features of the cities within the data-set.

3 Methodology

This section illustrates the methodology used to obtain a granular ranking among clusters with respect to a given index.
First, to obtain a measure with respect to which the clustering is done, Structural Equation Modelling (SEM) is employed. SEM is a powerful statistical technique that allows researchers to simultaneously model and estimate complex relationships among multiple dependent and independent variables. The concepts under consideration are typically unobservable and measured indirectly by multiple indicators. In this application, SEM is employed to obtain a single air pollution index.
As a second step, the cities' classification based on this latent construct is presented, estimating heterogeneity by k-means cluster analysis. This classification is useful to identify distinct classes of cities with similar air pollution characteristics. To establish a fine-grained ranking among these groups, clusters are ordered based on their centroid values. The cluster rank is assigned to all the units within that cluster.
The goal is to determine the highest feasible number of clusters, ensuring that all pairs of consecutive groups are distinct, according to the Wilcoxon rank-sum test.
We rank the European cities from the least to the most polluted group, providing a useful basis for pollution reduction strategies.

3.1 Structural Equation Modelling theoretical framework and model specification

The SEM model-based approach allows splitting the relations among latent and manifest variables, endogenous (endo) and exogenous (exo), in two parts ([21]). At first, there is a structural model that studies the formative (causal) relationships between the latent endogenous and exogenous variables (LVs), corresponding to latent concepts ([22]).
Furthermore, there are two measurement models for the LVs, which display the reflective relationships between the constructs and the manifest variables (MVs). All these relations are estimated simultaneously ([23], [24]). Given n multivariate observations, J exogenous MVs and M endogenous MVs, with $n > J + M$, the SEM can be formulated in a compact matrix form:

$$\begin{cases} \mathbf{N} = \mathbf{N}\mathbf{B}' + \mathbf{F}\boldsymbol{\Gamma}' + \mathbf{Z} \\ \mathbf{Y} = \mathbf{N}\mathbf{C}' + \mathbf{E} \\ \mathbf{X} = \mathbf{F}\mathbf{A}' + \mathbf{D} \end{cases} \tag{1}$$

where:

- $\mathbf{N}$ $(n \times L)$ and $\mathbf{F}$ $(n \times H)$ are the score matrices of endo and exo LVs, respectively (L is the number of endo LVs, H is the number of exo LVs);
- $\mathbf{B}$ $(L \times L)$ and $\boldsymbol{\Gamma}$ $(H \times L)$ are the regression coefficients' matrices of the endo and exo LVs, respectively;
- $\mathbf{Z}$ is the $(n \times L)$ matrix of the errors of the regression models;
- $\mathbf{Y}$ $(n \times M)$ and $\mathbf{X}$ $(n \times J)$ are the score matrices of endo and exo MVs, respectively;

- $\mathbf{C}$ $(M \times L)$ and $\mathbf{A}$ $(J \times H)$ are orthogonal matrices of coefficients (covariances or correlations) between MVs and endo and exo LVs, respectively;
- $\mathbf{E}$ $(n \times M)$ and $\mathbf{D}$ $(n \times J)$ are the matrices of the errors.

Recalling that the correlations between LVs and errors and between errors are for hypothesis null and given $\mathbf{\Sigma_{XX}}, \mathbf{\Sigma_{XY}}, \mathbf{\Sigma_{YY}}$ are the variance and covariance matrices of the measurement and structural models, let us define $\mathbf{S}$ the variance and covariance matrix of the MVs $[\mathbf{X}, \mathbf{Y}]$: $\mathbf{S} = \begin{bmatrix} \mathbf{S_{XX}} & \mathbf{S_{XY}} \\ \mathbf{S_{YX}} & \mathbf{S_{YY}} \end{bmatrix}$ and recall the estimators $\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma_{XX}} & \mathbf{\Sigma_{XY}} \\ \mathbf{\Sigma_{YX}} & \mathbf{\Sigma_{YY}} \end{bmatrix}$.

In this study, a Maximum Likelihood estimation has been applied and the discrepancy function to minimize is:

$$F_{ML}[\mathbf{\Sigma}] = \log|\mathbf{\Sigma}| + \mathrm{tr}[\mathbf{S}\mathbf{\Sigma}^{-1}] - \log|\mathbf{S}| - (J + M) \tag{2}$$

where $(J + M)$ is the number of MVs. The Likelihood is distributed according to chi-square with degree of freedom: $df = 1/2(J + M)(M + J + 1) - t$ where t is the number of parameters to be estimated.

3.2 Cluster analysis and choice of the number of clusters

In this paper, cluster analysis allows grouping units according to a given univariate index. The Centroid-based model of k-means has been employed ([25]). The k-means method assumes that each observation is equal to one of the K centroids. All the observations assigned to each centroid, perturbed by error in measuring the features, form a cluster. This clustering model can be written as:

$$\mathbf{X} = \mathbf{U}_K\mathbf{M}_K + \mathbf{E}_K \tag{3}$$

where $\mathbf{U}_K$, $\mathbf{M}_K$ and $\mathbf{E}_K$ are matrices of membership, prototype and error corresponding to K clusters, respectively. The Least Squares estimation of this model leads to the following optimization problem:

$$||\mathbf{X} - \mathbf{U}_K\mathbf{M}_K||^2 = \mathrm{tr}[(\mathbf{X} - \mathbf{U}_K\mathbf{M}_K)'(\mathbf{X} - \mathbf{U}_K\mathbf{M}_K)] \rightarrow \min_{\mathbf{U}_K, \mathbf{M}_K} \tag{4}$$

Subject to:

$$u_{ik} \in \{0, 1\}, \forall i, k \tag{5}$$

$$\mathbf{U}_K\mathbf{1}_K = \mathbf{1}_n, \tag{6}$$

where the rows of $\mathbf{M}_K$ are the cluster centroids, i.e., the data points averages in the clusters.
Once the index has been constructed (Sect. 4.1), the clustering goal is to partition the units in a disjoint set of K clusters in such a way that the dissimilarity between clusters is maximized.[3]

In this particular univariate clustering problem, the choice of the optimal number of clusters, denoted with K^*, is approached differently with respect to the most widely used methods. Instead of choosing a parsimonious number of isolated and homogeneous clusters, the goal is to determine the highest possible value for K^* while ensuring that consecutive groups remain statistically distinct. This departure from the conventional approach is motivated by the goal of obtaining a non-parsimonious

[3] A centroid, in this study, is the mean value of the index for the units in a cluster.

ordered partition of the units. Indeed, only units belonging to separate clusters are considered sufficiently dissimilar to have distinct ranks, while units within the same cluster are perceived as similar and assigned the same rank.

The statistical method used to identify K^* is explained below. For a given number of clusters k, the partitioning algorithm is run and each cluster is compared with the closest one to assess if they are significantly separated. This evaluation is conducted by means of a significance test, known as the Wilcoxon rank-sum test ([26]).
This test is a non-parametric, distribution-free, statistical hypothesis test used to compare two samples. It serves an alternative to the t-test, when the assumption of normal distribution is not met in the population. The test determines whether two samples are drawn from populations with the same underlying distribution. The alternative hypothesis of the Wilcoxon test states that there is a significant difference between the rankings of the two population groups being compared, while the null hypothesis suggests no difference.
The Wilcoxon rank-sum test should be preferred with respect to the t-test in situations where:

1. The sample distribution is not normal.
2. There are outliers.

The Wilcoxon rank-sum test is implemented as follows:

1. Take as input a set of values of two different groups (group 1 and group 2) with cardinalities n and m, respectively.
2. Compute all the possible $(n \times m)$ differences (D_i) between an element in group 1 and an element in group 2.
3. Rank the differences' absolute values from the smallest to the highest (differences of zero are dropped from the analysis).
4. Reassign the signs of these differences to their respective ranks (R_i).
5. Define the Wilcoxon test statistic as the smaller value W between W+ and W-, which are the sum of all the ranks (R_i) such that the corresponding difference (D_i) is positive or negative, respectively. If the null hypothesis is true, we expect W+ and W- to be similar.
6. Determine a critical value of the random variable W so that, if the observed value w of W is less than - or equal to - that critical value, H_0 is rejected in favour of H_1, and consequently the two groups can be considered statistically different one from the other. If w exceeds the critical value, H_0 is not rejected[4]

In this particular study, the Wilcoxon test compares the values of the units within a cluster (group 1) with those of the "consecutive" cluster (group 2). If these values are sufficiently different, group 1 is well separated not only from its nearest neighboring cluster (group 2) but also from all other clusters in the analysis.

[4] Equivalently: $Prob(w \leq W) = Prob(w - W \leq 0) = Prob(-W \leq -w) = Prob(W \geq w) = 1 - Prob(W < w) = 1 - pvalue$. Therefore, if $1 - pvalue$ is high (high probability that w is $\leq$ W) then pvalue is small and therefore reject the null hypothesis.

The critical value of W, for specific sample size and significance level (a for a one-tail test, $2a$ for a two-tail test), can be found in Wilcoxon distribution's tables. In fact, the Wilcoxon test statistics W is a random variable that follows the Wilcoxon distribution. This distribution is discrete, bell-shaped and non-negative. It is parameterized by two non-negative numbers: n and m.

The test is implemented for all pairs of consecutive clusters and therefore it is worthy to note that, for each k, *k-1* simultaneous significance tests are applied. Thus, to have an overall confidence level of $(1-a)\%$ for a specific value of k, the single test confidence level must be equal to $(1-a\times(k-1)\%)$, where *(k-1)* is the number of contemporaneous tests to be implemented. This correction technique is due to Bonferroni ([27]).

The clustering algorithm is run for different values of k, starting from $k = 2$. If the Wilcoxon tests assesses that the k clusters can be considered well separated, k is increased by one and the partitioning algorithm is run again. The procedure is iterated until a non-significant difference between "consecutive" clusters is observed. To facilitate clusters' comparison, it is essential to order them based on their centroid values, arranged from the smallest to the largest. This ordering allows for the efficient identification of consecutive clusters to be compared.
To perform the partitioning algorithm, a centroid-based 1-dimensional k-means model is employed to classify units according to the index built my means of SEM. The usual k-means algorithm, which is widely used for cluster analysis, does not guarantee the optimality of the final solution. However, in the particular case of one-dimensional clustering, an optimal dynamic programming algorithm has been developed by Froese et al. ([28]). The algorithm is implemented as an R package called *Ckmeans.1d.dp* ([29]).
Given $(x_1,\dots,x_N)$ the $(N\times 1)$ vector of observations, K the maximum number of clusters to be tested, ϵ the arbitrary constant for k-means convergence, a the confidence level of the Wilcoxon test and W a random variable following the Wilcoxon distribution, it is now possible to show the complete procedure (Algorithm 1).

Algorithm 1

for $k=2$ **to** K **do**
  //Start k-means
  $t \leftarrow 0$
  Randomly initialize k ordered centroids: $m_1^t,\dots,m_k^t$
  $\mathbf{C}_s \leftarrow \emptyset\ \forall s = 1,...,k$ //Ordered clusters
  repeat
    $t \leftarrow t+1$
    for $j=1$ **to** N **do**
      $s^* \leftarrow argmin_s ||x_j - m_s^t||^2$ //Assign x_j to the closest centroid
      $\mathbf{C}_s^* \leftarrow \mathbf{C}_s \cup x_j$
      for $i^*=1$ **to** k **do**
        $m_i^t \leftarrow \frac{1}{\mathbf{C}_s^*}\sum_{x_j \in \mathbf{C}_s^*} x_j$ //Centroid update step
      end for
    end for
  until $\sum_{s=1}^{k} ||m_s^t - m_s^{t+1}||^2 \leq \epsilon$
//End k-means
  ntest $\leftarrow$ k-1 //Number of simultaneous tests to be implemented
  for $s=1$ **to** ntest **do**
    $group1 \leftarrow x_j\ such\ that\ x_j \in \mathbf{C}_s^*$
    $n \leftarrow |group1|$
    $group2 \leftarrow x_j\ such\ that\ x_j \in \mathbf{C}_{s+1}^*$
    $m \leftarrow |group2|$

```
      for p = 1 to n do
        for q = 1 to m do
          D_i ← abs(group1[p] - group2[q])      //Compute the absolute value of the differences
          R_i ← rank(order(D_i))      //D_is are ranked from the smallest to the highest, their signs are
                                      // reassigned to their respective R_is
          W+ ← ∑ R_i such that D_i is positive
          W− ← ∑ R_i such that D_i is negative
          w ← min(W+, W−)      //Observed Wicoxon statistics
        end for
      end for
      pvalue ← Prob(W < w)
      if pvalue > a/ntest then
        stop      //Accept H_0, non-significant difference between consecutive clusters
        return k - 1      //Optimal (maximum) number of clusters
      end if
    end for
  end for
```

4 Application to urban air pollution in Europe

4.1 Structural Equation Model based air pollution index

In this study, a multidimensional hierarchical SEM has been employed to estimate an index of air pollution ([30]). This modeling specification has the advantages of simultaneously consider multiple levels within the model and leverage the information provided by relevant explanatory variables. Based on the results on an Explanatory Factor Analysis conducted on the six pollutants, a hierarchical, two latent factors model with exogenous MVs is estimated.
The air pollution index, which serves as a basis to rank cities, has been estimated employing the "sem" function from the *R "lavaan"* package ([31]). This function automatically standardizes the input variables, assigning negative weights to the variables that the factors reconstruct in the opposite direction from the others (e.g., Ozone). The resulting Model Based-Air Pollution Index (MB-API) is then normalized in $[0-1]$.
To determine the optimal set of explanatory variables for estimating the latent concept of air pollution, a MRF regression is used ([32], [33]). The MRF regression considers all the possible explanatory variables presented in Sect. 2, to simultaneously predict the six pollutants, and provides a score of importance for each covariate. The 8 most influencing covariates are selected and shown in Equation 7 and in the path diagram of the final SEM (Fig. 1).

$$\begin{cases} f1 = 0.41PM2.5 + 0.45PM10 - 0.23O3 + 0.24SO2 \\ f2 = CO - 0.3NO2 \\ MB\text{-}API = f1 + 0.3f2 \\ MB\text{-}API \sim -0.63lat + 0.31long + 0.63cars - 0.38temp \\ -0.61prs - 0.51\,YDR - 0.49part.rate - 0.22wind \end{cases} \tag{7}$$

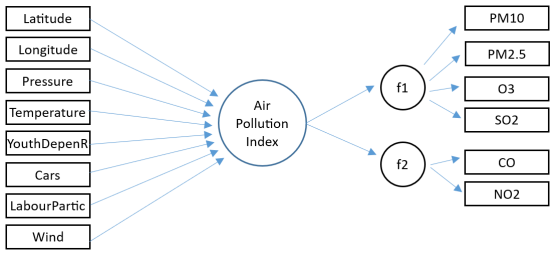

Fig. 1: SEM path diagram.

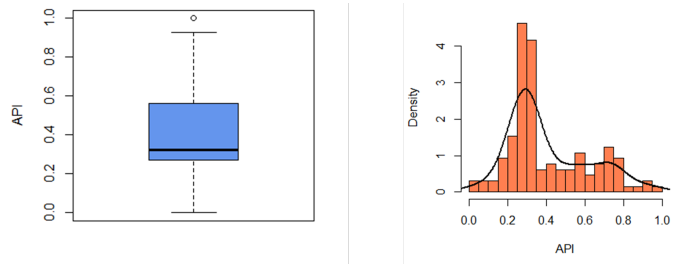

Fig. 2: Boxplot of MB-API values in EU cities in 2019 (panel A) and MB-API distribution (panel B).

To study the normality of the MB-API, a Shapiro-Wilks test is conducted. The test statistic is equal to 0.9, with a pvalue of $9.931e^{-08}$. Based on these results, it is possible to conclude that the index distribution deviates significantly from normality (Fig. 2). This finding further supports the choice of the non-parametric Wilcoxon test to determine the ideal number of clusters. It is worth noting that air pollution indexes often include cases with exceptionally high pollution values (outliers), which contribute to the departure from normality.

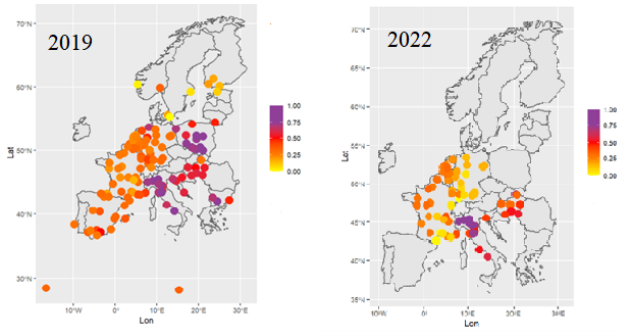

Fig. 3: MB-API values across EU cities: 130 cities in 2019 (panel A) and 70 cities in 2022 (panel B).

The map in panel A of Fig. 3 shows the different air pollution levels across the 130 European cities, (the two points located in the Mediterranean Sea are Santa Cruz de Tenerife and Las Palmas de Gran Canaria) according to the MB-API. Panel B reports the air pollution levels computed for the year 2022.

4.2 Cluster analysis

Cluster analysis is employed to identify groups of cities homogeneous in terms of air pollution level. The 130 European cities are grouped into clusters, each represented by a centroid that corresponds to a MB-API value. Two cities are included in the same cluster only if their pollution levels are closely similar. For this reason, having a large number of clusters ensures that cities within each cluster show high similarity and avoids clusters of cities with significant differences in air pollution levels.

Groups are ranked from 1 to K^* (K^* = total, maximum number of clusters) considering the centroids' values from the lowest to the highest: rank 1 corresponds to the smallest centroid and therefore to the group of less air polluted cities. In this application, according to the Wilcoxon rank-sum test and considering the Bonferroni correction for simultaneous testing, the ideal number of clusters is 11. Clusters for 2019 are shown in Tab. 1.

Furthermore, a comparison is made between the cities' air pollution levels in 2019 and 2022. The 70 cities included in the 2022 analysis are assigned to one of the 11 clusters individuated for the year 2019, basing on the minimum distance between the MB-API value in 2022 and the centroid value computed for 2019 (Tab. 2). This supervised clustering approach allows for the identification of cities that have improved or worsened their ranking, indicating corresponding changes in air quality. For instance, an examination of the results reveals that many French cities reduced their air pollution level with respect to 2019 (e.g., Perpignan moved from rank 5 to rank 1, indicating a significant improvement in air quality).

This approach enables a comprehensive understanding of the changes in air pollution levels across different cities over the specified time period.

Lastly, the left panel of the map in Fig. 4 shows cities ranking in 11 groups with a similar situation in terms of air pollution level in 2019. The ranking of the 70 cities for 2022, according to the 2019 categories, is shown in the panel on the right of Fig. 4.

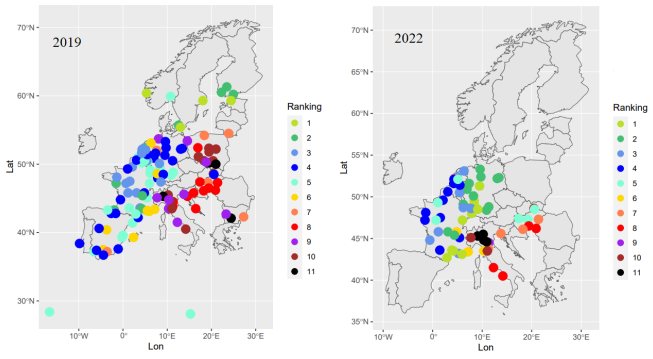

Fig. 4: Cities' ranking for 2019 (panel A) and 2022 (panel B).

Table 1: Cities' ranking according to the clustering (2019).

Rank	**Mean MB-API**	**Range MB-API**	**Cities within the cluster**
1	0.046	0-0.098	Malmo, Bergen, Stockholm, Tallin
2	0.179	0.136-0.207	Helsinki, Saint-Etienne, Tampere, Donostia, Copenhagen, Nimes, Amsterdam, Turku, Nantes
3	0.243	0.216-0.262	Limoges, Rotterdam, Orleans, The Hague, Zurich, Wiesbaden, Clermont-Ferrand, Montpellier, Utrecht, Rennes, Tours, Amiens, Maastricht, Breda, Haarlem
4	0.287	0.267-0.306	Sevilla, Paris, Rouen, Berlin, Malaga, Lyon, Nijmegen, Kassel, Brussels, Lille, Salamanca, Gasteiz, Hamburg, Koln, Eindhoven, Hannover, Antwerpen, Munster, Lisbon, Potsdam, Pamplona, Murcia, Bordeaux, Burgos, Freiburg, Stuttgart, Dodrecht, Dusseldorf
5	0.328	0.310-0.362	Augsburg, Besancon, Darmstadt, Grenoble, Bilbao, Perpignan, Barcelona, Santander, Metz, Marseille, Munich, Namur, Valencia, Santa Cruz de Tenerife, Nancy, Karlsruhe, Charleroi, Liege, Toulouse, Castellon de la Plana, Huelva, Oslo, Gent, Las Palmas de Gran Canaria
6	0.404	0.377-0.435	Madrid, Palma, Toulon, Groningen, Oviedo, Cordoba, Strasbourg, Nice
7	0.486	0.473-0.502	Muenster, Burgas, Gdansk, Kaunas, Budapest, Granada
8	0.572	0.533-0.607	Zagreb, Debrecen, Gyor, Szeged, Pecs, Rijeka, Poznan, Kecskemet, Bologna, Florence
9	0.686	0.639-0.713	Rome, Trieste, Sofia, Livorno, Bydgoszcz, Szczecin, Parma, Zabrze, Turin
10	0.760	0.734-0.789	Lodz, Rybnik, Wroclaw, Miskolc, Modena, Plock, Naples, Warsaw, Brescia, Kielce, Prato
11	0.917	0.848-1	Cracow, Milan, Tarnow, Plovdiv, Katowice

Table 2: Cities assigned to the 2019 clusters basing on 2022 air pollution level. Values in brackets show decrease or increase in the ranking of the cities of 2022 with respect to the ranks of 2019.

Rank	**Cities within the cluster**
1	Freiburg(+3), Perpignan(+4), Montpellier(+2), Besancon(+4), Kassel(+3), Nimes(+1), Darmstadt(+4), Stuttgart(+3), Karlsruhe(+4), Toulon(+5)
2	Clermont-Ferrand(+1), Wiesbaden(+1), Munich(+3), Hamburg(+2), Augsburg(+3), Hannover(+2), Saint-Etienne, Potsdam(+2), Berlin, Metz(+3), Munster(+2), Koln(+2), Dusseldorf(+2)
3	Groningen(+3), Maastricht, Limoges, Breda, Marseille(+2), Amsterdam(-1), Nancy(+2), Bordeaux(+1), Strasbourg(+3)
4	Orleans(-1), Nijmegen, Amiens(-1), The Hague(-1), Rotterdam(-1), Rennes(-1), Grenoble(+1), Toulouse(+1), Nantes(-2), Lille
5	Budapest(+2), Tours(-2), Utrecht(-2), Rouen(-1), Paris(-1), Eindhoven(-1), Gyorf(+3)
6	Lyon(-2), Dodrecht(-2), Nice, Livorno(+3)
7	Trieste(+2), Haarlem(-4), Pecs(+1), Debrecen(+1), Florence(+1)
8	Rome(+1), Kecskemet, Szeged, Naples(+2)
9	Bologna(-1)
10	Miskolc, Prato, Turin(-1)
11	Parma(-2), Brescia(-1), Modena(-1), Milan

5 Concluding remarks

This paper presents a novel approach to rank objects with respect to a given measure. In this field, the typical practice is to rank units based on their value of a given indicator. However, the proposed technique uses clustering to assign the same rank to units that belong to the same cluster.
Indeed, the primary objective in this paper is to achieve a granular, non-parsimonious, ranking of units within homogeneous groups that represent equivalence classes and for this reason the number of clusters should be large. This approach ensures that units with similar index values are grouped together, while also minimizing the inclusion of units with significantly different index values within the same cluster.
The presented technique identifies the maximum number of well-distinguished clusters according to the Wilcoxon rank-sum test.
In particular, the procedure to find the clusters optimal number requires an algorithm that iterates the 1-dimensional k-means clustering and the Wilcoxon test until the test returns, for a specific k, the acceptance of the null hypothesis of distributions' equality (thus, the algorithm stops and returns $K^* = k\text{-}1$).

This new approach is applied considering clustering with respect to an air pollution index resulting from a model-based SEM.
In the application, 130 European cities are ranked in eleven clusters with respect to their air pollution level.

Future developments will involve comparing this methodology with other techniques designed for analyzing and modeling ranked data. Additionally, further exploration will focus on the multivariate extension of the methodology. The multivariate partition of units into clusters should address the following properties: units within clusters are similar (homogeneous) and units between clusters are dissimilar (isolated), clusters are statistically distinct according to a test, clusters are ranked.

References

1. Garrido M., Hansen S.k., Yaari R., Hawlena H. : A model selection approach to structural equation modelling: A critical evaluation and a road map for egologists. Methods in Ecology and Evolution, 13, 42-53 (2021).
2. Fan, Y., Chen, J., Shirkey, G. : Applications of structural equation modeling (SEM) in ecological studies: an updated review. Ecol Process, 5, 19 (2016).
3. Budtz-Jørgensen E., Debes F., Weihe P., Grandjean P. : Structural equation models for meta-analysis in environmental risk assessment. Environmetrics, 21(5), 510–527 (2010).
4. Bruno F., Cocchi D. : Recovering information from synthetic air quality indices. Environmetrics, 18, 345-359 (2007).
5. Jiang Y., He X., Lee M.L.T., Yan J. : "Wilcoxon Rank-Based Tests for Clustered Data with R Package clusrank". Journal of Statistical Software, 96 (2017).
6. Tibshirani R., Walther, G., Hastie, T. : 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 (2001).
7. Rousseeuw P.J. : Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math., 20, 53–65 (1987).
8. Shi C., Wei B., Wei S. Wang W., Liu H., Liu J. : A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. EURASIP Journal on Wireless Communications and Networking, 2021(1), 1-16 (2021).

9. Vogel M.A. and Wong A. K. C. : PFS Clustering Method. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-1, 3, 237-245 (1979).
10. Boaz R. M., Lawson A. B., Pearce J. L. : Multivariate air pollution prediction modelling with partial missingness. Environmetrics, 30(7): e2592 (2019).
11. Liu, Y., Zhou, Y., Lu, J. : Exploring the relationship between air pollution and meteorological conditions in China under environmental governance. Sci Rep 10, 14518 (2020).
12. Martori J.C., Lagonigro R., Pascual R.I. : Sustainable Cities and Society Social status and air quality in Barcelona: A socio-ecological approach. Sustainable Cities and Society, 87, 104210 (2022).
13. Davis M. E. : Recessions and Health: The Impact of Economic Trends on Air Pollution in California. Am J Public Health, 102(10), 1951–1956 (2012).
14. Chen B., Kan H. : Air pollution and population health: a global challenge, Environ. Health Prev. Med, 13(2), 94-101 (2008).
15. OECD (2012) Redefining "urban". A new way to measure metropolitan areas.
16. Choma E. F., Evansb J. S., Gomez-Ibanezc J. A., Did Q., Schwartzb J. D, Hammitte, J. K., Spenglerb J. D. : Health benefits of decreases in on-road transportation emissions in the United States from 2008 to 2017. PNAS, 118 (51) (2021).
17. Eurostat : Methodological manual on territorial typologies. Luxembourg (2019).
18. Eurostat : How polluted is the air in urban areas? EDN-20210603-1 (2021)..
19. Urdangarin A., Goicoa T. and Ugarte M.D. : Evaluating recent methods to overcome spatial confounding. Revista Matemática Complutense, 333–360 (2022).
20. Genuer, R., Poggi J.-M., Tuleau-Malot C. : Variable selection using random forests. Pattern Recognition Letters, 31(14), 2225-2236 (2010).
21. Landis R. S., Beal D. J., Tesluk P.E. : Comparison of Approaches to Forming Composite Measures in Structural Equation Models. Organizational Research Methods, 3: 186 (2000).
22. Bollen K.A. : Evaluating Effect, Composite, and Causal Indicators in Structural Equation Models. MIS Quarterly, 35(2), 359-372 (2011).
23. Hair J. F., Sarstedt M. : Explanation plus prediction – The logical focus of project management research. Project Management Journal, 52(4), 319–322 (2021).
24. Tarka P. : An overview of structural equation modeling: its beginnings, historical development, usefulness and controversies in the social sciences. Quality & Quantity, 52, 313–354 (2018).
25. Vichi M., Cavicchia C., Groenen P. J. F. : Hierarchical Means Clustering, Journal of Classification, 39(3), 553-577 (2022).
26. Wilcoxon F. : Individual comparisons by ranking methods. Biometrics Bulletin 1 (1945).
27. Armstrong R. : When to use the Bonferroni correction. Ophthalmic Physiol, 34(5):502-8 (2014).
28. Froese R., Klassen J. W., Leung C. K. and Loewen T. S. : The Border K-Means Clustering Algorithm for One Dimensional Data. IEEE International Conference on Big Data and SmartComputing, 35-42 (2022).
29. Wang H. and Song M. : Ckmeans.1d.dp: Optimal k-means Clustering in One Dimension by Dynamic Programming. The R Journal Vol. 3/2 (2011).
30. Cavicchia C., Vichi M. : Second-order disjoint factor analysis. Psychometrika, 87 (1), 289–309 (2022).
31. Rosseel Y. : lavaan: An R Package for Structural Equation Modeling, Journal of Statistical Software, 48 (2) (2012).
32. Grömping U. : Variable Importance Assessment in Regression: Linear Regression versus Random Forest. The American Statistician, 63:4, 308-319 (2009).
33. Breiman L. : Random Forests. Machine Learn, 45, 5–32 (2001).