# Computational assessment of k-means clustering on a Structural Equation Model based index

Mariaelena Bottazzi Schenone[a], Elena Grimaccia[b], Maurizio Vichi[a]

[a] Department of Statistical Sciences, Sapienza University, Rome; mariaelena.bottazzischenone@uniroma1.it,maurizio.vichi@uniroma1.it

[b] ISTAT - Italian National Institute of Statistics, Rome; elgrimac@istat.it

**Abstract**

This paper proposes an alternative method for the choice of the number of centroids in a cluster analysis, when the groups' order is relevant. Differently from commonly used approaches, aimed at finding the minimum number of clusters, the illustrated method aims at finding the maximum one. Given that the clusters are ordered, this allows to define a granular ranking among them. The *k*-means partitioning algorithm is applied to an index resulting from a Structural Equation Model. The procedure is implemented on a measure of air pollution in urban areas: a clustering of main Italian cities, according to the optimal number of air pollution levels, is the final result. The analysis' interpretation provides useful information to design policies aimed at reducing air pollution.

**Key words:** Cluster analysis, Computational assessment, Structural Equation Models, Air Pollution, Latent variable models, Environmental statistical models.

## 1. Introduction

Most of the commonly used methods to identify the optimal number of clusters in a cluster analysis aim at finding the minimum *k* (Gap statistics [16]; Silhouette method [13]; Elbow method [14]). In this paper, the idea behind the choice of *k* is the opposite: the goal is to find the maximum number of "*distinguished*" clusters: what are the "well-distinguished" clusters is defined by means of bootstrap confidence intervals for clusters' centroids. Given a sample of units and a number of clusters *k*, these intervals can be computed by bootstrapping those units a number B of times and computing each time the *k*-means. The results are B vectors of *k* centroids. The final estimates of the *k* clusters' centroids as well as their empirical distributions can be obtained computing the mean and plotting the histograms of the bootstrap replicates, respectively. Given the *k* centroids' point estimates with the corresponding $\alpha/2$ and $(1-\alpha/2)$ percentile estimates, it is possible to build *k* percentile confidence intervals of the desired confidence level $\alpha$. If some of these confidence intervals do overlap by more than a fixed small constant $\varepsilon$, then the clusters are not "well-distinguished". The optimal number of clusters $k^*$ will be the maximum *k* such that none of the *k* intervals do overlap by more than $\varepsilon$.

A new air quality index has been estimated by means of the application of a Structural Equation Model (SEM) to pollutants variables, as it takes into account the multivariate relationships among contaminants ([2]). The analysis of the new index distribution among the Italian metropolitan areas can be useful to draw policy conclusions devoted to reducing air pollution levels.

The clustering analysis homogeneously groups Italian cities with respect to different air pollution levels. Assigning a rank to the cities within the same cluster, it is possible to classify them from the most to the less polluted.

The paper is structured as follows: Section 2 presents data used for the empirical study, Section 3 introduces SEM's specifications and modelling and the cluster analysis technique employed. In Section 4, an application on air quality for Italian cities is provided. In Section 5 concluding remarks are drawn.

## 2. Data

The dataset on which the study is conducted comes from the European Environmental Agency (EEA) and refers to 6 pollutants' emissions of 106 Italian provinces in 2022. For each city, the pollutant emission is computed as the average over the 365 days of daily median emissions. These 6 gases are the ones that the Environmental Protection Agency (EPA) individuated as major air pollutants and they are regulated by the Clean Air Act. They are: Ground-level ozone (O3), Particle pollution (also known as Particulate Matter (PM), including PM2.5 and PM10), Carbon monoxide (CO), Sulfur dioxide (SO2) and Nitrogen dioxide (NO2).

## 3. Methods: Structural Equation Model and Cluster analysis

In this paper, an index is built employing a Structural Equation Model. This model combines Confirmatory Factor Analysis and Multiple Regression Analysis into a comprehensive modelling framework that involves both endogenous and exogenous variables. In the so called "measurement part" of the model, the relations between the observed (manifest) variables (MVs) and the latent factors (LVs) can be studied. Moreover, the "structural part" of the model studies the causal relationships of the latent constructs among themselves. All these relations are estimated simultaneously ([8]; [3]; [15]; [6]).

In order to rank units with respect to the SEM-based index, the centroid-based model of *k*-means ([17]) is employed. 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, forms a cluster. The clustering goal is to partition the units in a disjoint set of *k* clusters to maximise the dissimilarity between centroids of the clusters.

In this analysis, a centroid of a cluster is the mean value of the index. For this univariate clustering problem, the optimal number of clusters $k^*$ is chosen to be the highest possible, as long as centroids are statistically different according to a desired confidence level α. The idea is to rank clusters by their centroids, so that units belonging to a cluster have the rank of that cluster. Hence, $k^*$ is the maximum value for which all $k^*$ centroids' confidence intervals at significance level (1-α) do not overlap by more than a fixed constant ε.

However, because of its deterministic nature, *k*-means does not yield confidence information about centroids' distribution and estimated cluster memberships, although this could be useful for inferential purposes. It is possible to obtain such information by means of a non-parametric bootstrap procedure. This procedure provides centroids' distributions ([9]) which can be used to derive probabilistic membership information on each object from all bootstrap samples. It also yields confidence information about the centroids in the form of confidence intervals ([7]).

Given a sample of size n, for a given *k*, the partitioning algorithm is run. The corresponding *k* centroids' estimates and corresponding confidence intervals are built applying bootstrap to that sample of n units.

For each of the B bootstrap iterations, the *k* centroids are ordered from the smallest to the largest. This allows to easily match the centroids over different bootstrap samples. Once the matching is done, the final estimate of each of the *k* centroids is obtained as the mean of the corresponding B centroids' values. The α/2% and (1- α/2)% centroids' percentiles can be estimated in a similar way and the (1-α)% percentile bootstrap confidence intervals can be computed as follows: $\left[percentile_{\frac{\alpha}{2}}\,;\, percentile_{\frac{1-\alpha}{2}}\right]$.

The partitioning algorithm chosen is a centroid-based 1-dimensional *k*-means and units are classified according to the SEM-based index. In this particular, unidimensional case, an optimal dynamic programming *k*-means algorithm has been developed by Froese et al. ([5]). The algorithm is implemented in the Ckmeans.1d.dp R package ([18]).

The clustering algorithm is run for different values of *k*, starting from $k = 2$. At each iteration, if the *k* bootstrap confidence intervals do not overlap by more than ε, the *k* clusters can be considered well separated. Then, *k* is increased by one and the partitioning algorithm is run again. The procedure is

iterated until two overlapping confidence intervals are found. A crucial point is the need of ordering the clusters with respect to their centroid value, from the smallest to the largest. This allows to find the consecutive clusters' confidence intervals to be compared.

## 4. Application to urban air pollution for Italian metropolitan areas

In this study, a multidimensional index to measure air pollution is built by means of a hierarchical SEM ([4]). This flexible model has the advantages of taking simultaneously into account a number of levels in the hierarchy and to exploit the information available in meaningful explanatory variables.

Based on the results of a preliminary Explanatory Factor Analysis on the six main pollutants, a hierarchical, two latent factors model is estimated using the "sem" function from the R "lavaan" package ([12]). This function automatically standardizes the six pollutants, assigning negative weights to the variables that the factor reconstructs in the opposite direction from the others.

The resulting index, called Model Based-Air Pollution Index (MB-API), is normalized in 0-1. This index allows to rank cities with respect to their air pollution level.

The estimated model can be written as follows.

$$\begin{cases} f1 = 0.64PM2.5 + 0.62PM10 - 0.37O3 + 0.41SO2 \\ f2 = CO - 0.65NO2 \\ MBAPI = f1 + 0.79f2 \end{cases} \quad (1)$$

Fig.1 shows the different levels of air pollution, according to the MB-API, for the 106 Italian cities. The most polluted areas are in Piedmont, Lombardy, Veneto and Emilia-Romagna, while the less polluted cities are in the islands and in Calabria.

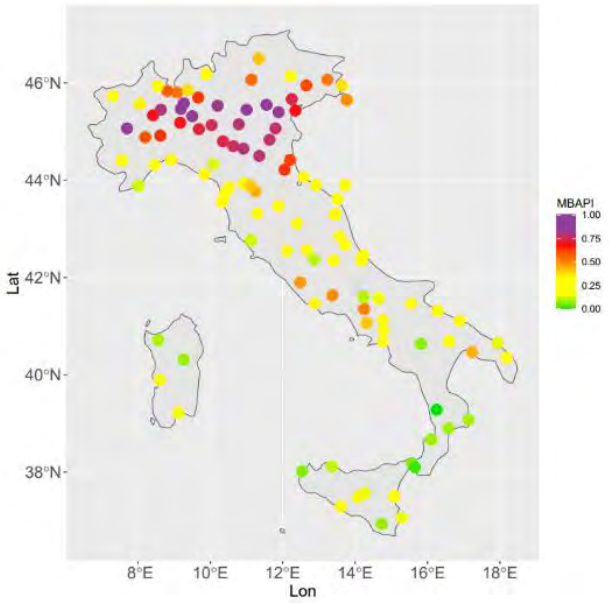


Figure 1: Air pollution index values in Italian metropolitan areas.

Cluster analysis is then applied to find groups of cities homogeneous with respect to the air pollution level. The Italian cities are grouped into clusters, each represented by a centroid that corresponds to an index value. It is important to note that to allow cities' ranking, cluster must be ordered with respect to the corresponding centroids. The clustering algorithm is run for $k = 2$ up to 10. The bootstrap procedure (with a number of bootstrap replicates equal to 10000) is used to compute the corresponding centroids' confidence intervals at 90% shown in Fig. 2.

Two clusters are considered well separated if the difference between the upper bound of a cluster and the lower bound of the consecutive one is smaller than a constant $\varepsilon = 0.01$ (1% of the index range, equal to 1). It is possible to note that for $k = 6$ the clusters do overlap by more than ε and therefore the optimal $k$ should be 5. However, for $k = 7$ the clusters are well distinguished and therefore the optimal maximum number of non-overlapping clusters $k^*$ is 7.

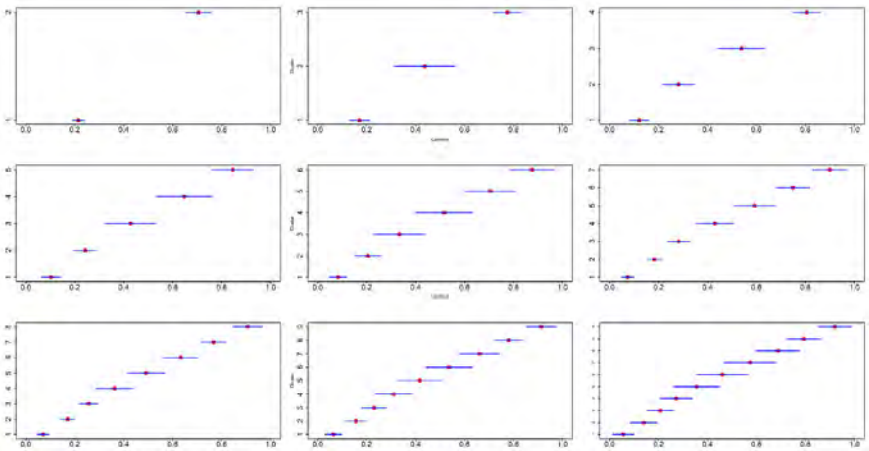

Figure 2: 90% bootstrap confidence intervals for *k*-means centroids. *k* ranges in 2-10.

Of course, the optimal number $k^*$ depends on the α and ε values chosen *a priori*. It is worthy to note that this method of choosing *k* is very different with respect to the classical Elbow, Silhouette and Gap Statistics methods, whose aim is to find the minimum (and not the maximum) *k* such that the clusters are well defined. In all 3 cases, in fact, $k^*=2$ (as shown in the plots of Fig.3).

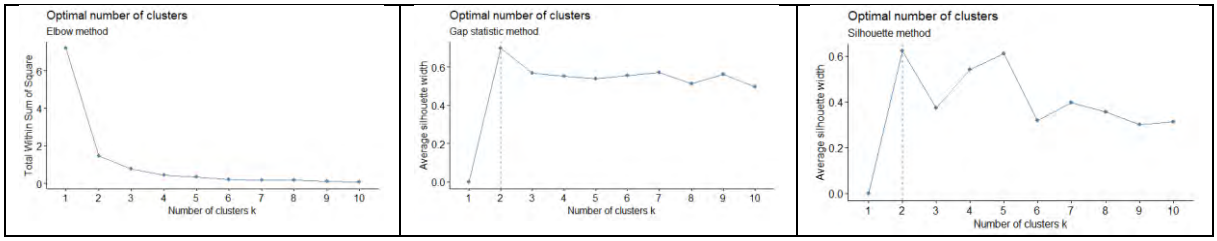


Figure 3: Choice of the optimal number of clusters according to the most widely used methods.

The maximum optimal number of clusters can be found also considering the maximum number of clusters whose centroids are simultaneously significantly different, according to nonparametric Wilcoxon tests ([10]), with confidence level chosen following the Bonferroni's correction ([1]). In this case, $k^*=10$ and therefore the more parsimonious solution $k^* = 7$ seems even more reasonable.

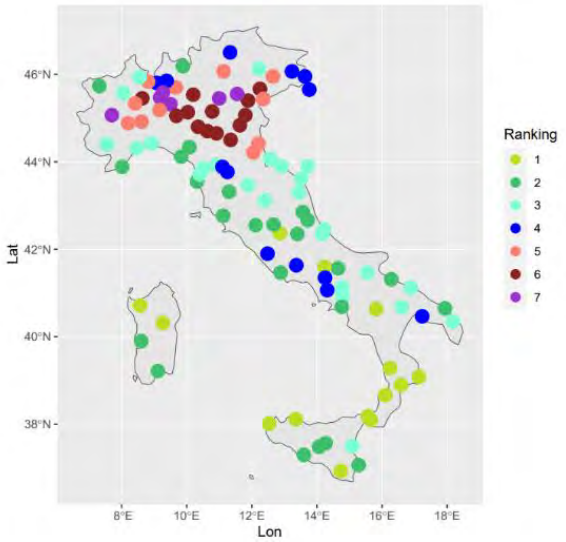


Figure 4: Choice of the optimal number of clusters according to the most widely used methods.

Basing on the previous results, groups are ranked from 1 to 7 considering the centroids' values from the highest to the lowest: rank 1 corresponds to the lowest centroid and therefore to the group of less air polluted cities.

The map in Fig. 4 shows air pollution distribution in Italy in 2022, highlighting groups of cities with a similar situation in terms of air pollution levels. It is possible to note that close points tend to have the same colour: cities in the same region often have a similar air pollution level.

## 5. Concluding remarks

The novelty of our paper resides in the fact that the *k*-means cluster analysis is employed in order to estimate the maximum number of significantly different centroids. The significance is assessed by means of percentile confidence intervals, built with a bootstrap procedure.

The performance of the proposed method is shown by an air pollution analysis: a measure of air quality in metropolitan areas has been developed based on a structural equation model. The procedure obtain a ranking of Italian cities, according to the optimal number of clusters of air pollution.

As a possible extensions of the study, the air pollution index can be improved considering also exogenous explanatory variables in the Structural Equation Model, such as the number of cars in the city or the percentage of green spaces. The model immediately integrates the new covariates. In case of adding more covariates, a different multidimensional clustering technique must be used and the obtained results can be compared with the unidimensional case.

Furthermore, a sensitivity and robustness analysis of the ranking can be conducted in a simulation framework, for instance computing average absolute shift in ranking index.

Thanks to this granular ranking, policy makers can easily identify the most polluted metropolitan areas and employ the estimated index as a unique measure of air pollution.

### References


[1] Armstrong R. : When to use the Bonferroni correction. Ophthalmic Physiol, 34(5):502-8 (2014).
[2] Boaz R. M., Lawson A. B., Pearce J. L. : Multivariate air pollution prediction modelling with partial missingness. Environmetrics, 30(7): e2592 (2019).
[3] Bollen K.A. : Evaluating Effect, Composite, and Causal Indicators in Structural Equation Models. MIS Quarterly, 35(2), 359-372 (2011).
[4] Cavicchia C., Vichi M. : Second-order disjoint factor analysis. Psychometrika, 87 (1), 289–309 (2022).
[5] 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 Smart Computing, pp. 35-42 (2022).
[6] Hair J. F., Sarstedt M. : Explanation plus prediction – The logical focus of project management research. Project Management Journal, 52(4), 319–322 (2021).
[7] Hofmans J. : On the Added Value of Bootstrap Analysis for K-Means Clustering (2015).
[8] Landis R. S., Beal D. J., Tesluk P.E. : Comparison of Approaches to Forming Composite Measures in Structural Equation Models. Organizational Research Methods 2000 3: 186 (2000).
[9] Martella F., Vichi M. : Clustering microarray data using model-based double K-means (2012).
[10] Rey D., Neuhäuser M. : Wilcoxon-Signed-Rank Test. International Encyclopedia of Statistical Science. Springer (2011).
[11] Rizzo M. : Statistical Computing with R. Computer Science and Data Analysis Series. Chapman & Hall/CRC The R Series. p. 198 (2008).
[12] Rosseel Y. : lavaan: An R Package for Structural Equation Modeling, Journal of Statistical Software, 48 (2) (2012).
[13] Rousseeuw P.J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65 (1987).
[14] 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, 1-16 (2021).
[15] 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).
[16] 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).
[17] Vichi M., Cavicchia C., Groenen P. J. F. : Hierarchical Means Clustering, Journal of Classification. https://doi.org/10.1007/s00357-022-09419-7 (2022).
[18] Wang H. and Song M. : Ckmeans.1d.dp: Optimal k-means Clustering in One Dimension by Dynamic Programming. The R Journal Vol. 3/2 (2018).