Assessing environmental quality by clustering a structural equation model based index: An application to European cities air pollution

Una misura di qualità ambientale ottenuta clusterizzando un indice basato su modelli ad equazioni strutturali: una applicazione alla qualità dell'aria nelle principali città europee

Mariaelena Bottazzi Schenone, Elena Grimaccia and Maurizio Vichi

Abstract This paper proposes an innovative computational procedure to determine the optimal number of clusters. The aim is to identify the maximum number of significantly distinct clusters, when the centroids are orderable and order is relevant. The insight is that ranking according to this optimal number of clusters allows to better classify units in order to assess their quality with regard to a variable of interest. By means of bootstrap confidence intervals estimated on clusters' centroids, the procedure allows to identify the optimal number of "well-separated" groups. The centroids are obtained applying a unidimensional k-means clustering and they allow to classify and rank the measure of an Index based on a Structural Equation Model. The procedure ranks European cities according to their level of air pollution.
Abstract *Il lavoro propone una procedura computazionale innovativa per determinare il numero ottimale di cluster. Lo scopo è identificare il numero massimo di cluster significativamente distinti, quando i centroidi sono ordinabili e l'ordine è rilevante. L'intuizione è che la classificazione in base a questo numero ottimale di cluster consente di classificare le unità al fine di valutarne la qualità rispetto a una variabile di interesse. La procedura consente di identificare il numero ottimale di gruppi "ben separati", mediante intervalli di confidenza bootstrap. I centroidi sono ottenuti applicando un clustering k-medie unidimensionale e permettono di classificare un Indice stimato con un Modello ad Equazioni Strutturali. La procedura consente di classificare le principali città europee in base al loro livello di inquinamento atmosferico.*

Mariaelena Bottazzi Schenone
Department of Statistical Sciences, Sapienza University, Rome (Italy), e-mail: mariaelena.bottazzischenone@uniroma1.it

Elena Grimaccia
ISTAT - Italian National Institute of Statistics, Rome (Italy), e-mail: elgrimac@istat.it

Maurizio Vichi
Department of Statistical Sciences, Sapienza University, Rome (Italy), e-mail: maurizio.vichi@uniroma1.it

Key words: Bootstrap confidence intervals, Simulation study, Environmental quality, Structural Equation Models, Cluster analysis, Air Pollution.

1 Introduction

Most of the commonly employed clustering methods aims at identifying the optimal minimum number k of centroids [16,14,15]. However, when the aim of the study is the ranking of units according to a measure, it would be useful to identify the maximum number of clusters. This paper presents a procedure aimed at finding the optimal maximum number of well separated clusters, classifying an index resulting from a Structural Equation Model (SEM). In the application, the k-means clustering is applied to an index that measures Air Pollution, taking into account simultaneously the six main pollutants usually considered in the literature. The estimation of the SEM accounts also for several meaningful socio economic and climate-related covariates that enhance the significance of the estimates. The clustering of European metropolitan areas with respect to different air pollution levels is presented. The number of centroids has been chosen, considering the maximum number of clusters whose (1-α)% confidence intervals do not overlap by more than α. The ranking of the main 130 European cities provides useful information to design policies, aimed at reducing urban air pollution [5].

2 Data

The six main air pollutants identified by the Environmental Protection Agency (EPA) are: Ground-level ozone (O_3), Particle pollution (also known as Particulate Matter (PM), including PM2.5 and PM10), Carbon monoxide (CO), Sulphur dioxide (SO_2) and Nitrogen dioxide (NO_2). Data on pollutants in 130 metropolitan areas in the European Union are obtained from the Worldwide Air Quality data (https://aqicn.org/), which cover pollutants and atmospheric conditions around the world [1]. In addition, socio-economic features of cities (GDP per capita, population density, elderly and youth dependency ratios, employment, unemployment and participation rates) are included [12,4] together with meteorological and atmospheric covariates: air temperature, humidity, air pressure, wind-gust (m/s) and wind-speed (m/s), according to [10]. Traffic-related air pollutant emissions have become a global environmental problem, most of all in urban areas [3]. 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. Geographical covariates (Latitude and Longitude) have been included in the analysis, in order to take into account the spatial configuration of the phenomenon of air pollution [17].

3 Methods

The Air Pollution Index employed for measuring atmospheric environmental quality is built with a Structural Equation Mode that takes into account both endogenous and exogenous variables [7]. The relationships between observed (manifest) variables and latent factors, and among latent constructs themselves are estimated simultaneously [8]. In order to rank units with respect to the SEM-based index, the centroid-based model of k-means [18] 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. 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 achive such information by means of a non-parametric bootstrap procedure. This procedure provides centroids' distributions [11] 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 [9]. Given a sample of units and a number of clusters k, this can be done bootstrapping those units a number B of times. 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. 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 α. If some of these confidence intervals do overlap by more than α, then the clusters are not "well-separated". The optimal number of clusters k* will be the maximum k such that none of the k intervals do overlap by more than α. Given a sample of size n, for a given k, the partitioning algorithm is run. The corresponding k centroids' confidence intervals are built applying bootstrap to that sample of n units. Bootstrap allows to estimate each of the k centroids as the mean of the centroids' values for all the bootstrap replicates. This technique allows also to estimate the corresponding centroid standard error and therefore compute the $(1-\alpha)\%$ percentile confidence interval [13]. The clustering algorithm has been applied for different values of k, starting from k = 2. If the k bootstrap confidence intervals do not overlap, 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 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. The partitioning algorithm chosen is a centroid-based 1-dimensional k-means and units are classified according to an index built my means of SEM. In this particular unidimensional case, an optimal dynamic programming algorithm has been developed by [6].

4 Results

In this study, a multidimensional index to measure air pollution is built by means of a hierarchical SEM [2]. This 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 [7]. We called the resulting index Air Pollution Index (API). Cluster analysis is then applied to find groups of cities homogeneous with respect to the air pollution level. European cities are grouped into clusters, each represented by a centroid that corresponds to an API value. It is important to note that to allow cities' ranking, clusters must be ordered with respect to the corresponding centroids. The clustering technique of 1 dimensional k-means is applied using the R function "Ckmeans.1d.dp" of the homonymous R package [19]. 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.1.

When the difference between the upper bound of a cluster and the lower bound of the consecutive one is smaller than $\alpha = 0.05$, then they are considered well separated. In this application, for k = 8 the clusters do overlap by more than α, and therefore the optimal k* is 7. It is worth mentioning that this method of choosing k is very different compared to the Elbow, Silhouette or Gap Statistics methods, whose aim is to find the minimum k, such as the units are optimally allocated in separate clusters. According to all these 3 methods, in fact, k*=2.

Based 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. Fig. 2 shows groups of European cities with a similar situation in terms of air pollution levels in 2022. It

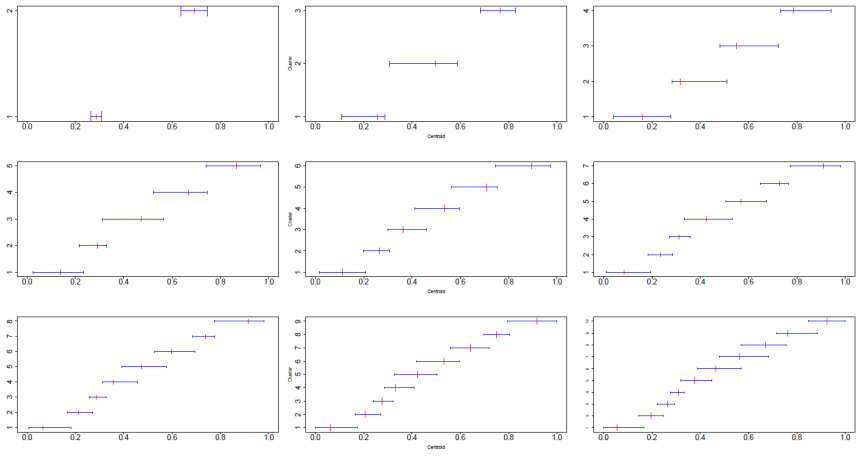

Fig. 1 95% bootstrap confidence intervals for k-means centroids. k ranges in 2-10.

is possible to note that close points tend to have the same colour: cities in the same country mostly have a similar air pollution level.

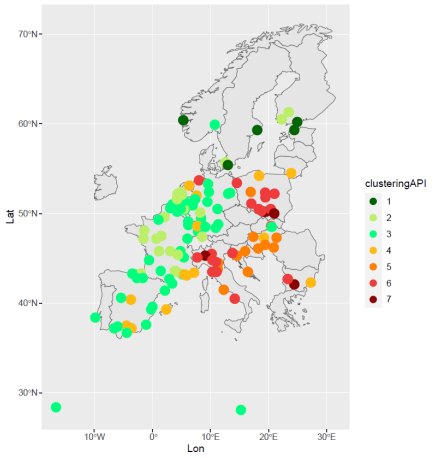

Fig. 2 Clusters of European cities according to API.

5 Concluding remarks

In this paper, European cities are grouped in clusters and ranked with respect to their air pollution level, measured by a structural equation model based index (API). The optimal number of clusters is the maximum number of significantly different centroids, according to centroids' percentile confidence intervals built by means of bootstrap. In this way, an innovative procedure aimed at identified the best maximum number of cluster is exemplified. Moreover, European cities are more granularly classified in seven clusters, with a gain in information, compared to the much smaller number of clusters suggested by other classical methods. The procedure could be improved employing a multidimensional clustering technique, to be compared to this unidimensional approach.

References

1. Boaz R. M., Lawson A. B., Pearce J. L.: Multivariate air pollution prediction modelling with partial missingness. Environmetrics, 30(7): e2592 (2019)
2. Cavicchia C., Vichi M.: Second-order disjoint factor analysis. Psychometrika, 87 (1), 289–309 (2022)
3. 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)
4. 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)
5. Dominici F., Samet J. M., Zegeral S. L.: Combining Evidence of Air Pollution and Daily Mortality from the 20 Largest US Cities: A Hierarchical Modelling Strategy. Journal of the Royal Statistical Society Series A. (2000)
6. 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)
7. Grimaccia E., Bottazzi-Schenone M., Vichi M.: Structural-Equation-Model-based assessment of Pollution in European urban Areas. Conference of European Statistics Stakeholders 2022. Available via https://drive.google.com/file/d/1V3BN-K9SY66Q7mB5UnGuQvJSGYV7f2dC/view Cited 10 March 2023
8. Hair J. F. and Sarstedt M.: Explanation plus prediction – The logical focus of project management research. Project Management Journal, 52(4), 319–322. (2021)
9. Hofmans J.: On the Added Value of Bootstrap Analysis for K-Means Clustering. Journal of Classification (2015)
10. Liu, Y., Zhou, Y., Lu, J.: Exploring the relationship between air pollution and meteorological conditions in China under environmental governance. Sci Rep 10, 14518. https://doi.org/10.1038/s41598-020-71338-7 (2020)
11. Martella F., Vichi M.: Clustering microarray data using model-based double K-means. Journal of Applied Statistics (2012)
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. Rizzo M.: Statistical Computing with R. Computer Science and Data Analysis Series. Chapman and Hall/CRC The R Series. p. 198. (2008)
14. Rousseeuw P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65. (1987)
15. 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), 1-16 (2021)
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. Urdangarin A., Goicoa T. and Ugarte M.D.: Evaluating recent methods to overcome spatial confounding. Revista Matemática Complutense 10.1007/s13163-022-00449-8. (2022)
18. Vichi M., and Kiers H. A. L.: Factorial k-means analysis for two-way data. Computational Statistics and Data Analysis, 37, 49–64 (2001)
19. 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)