UCL DEPARTMENT OF GEOGRAPHY

YEAR 2020-21

MODULE CODE:	**GEOG0114**
MODULE NAME:	**Principles of Spatial Analysis**
COURSE PAPER TITLE:	**K-means clustering algorithm review for air pollution prediction affected by meteorological conditions**
WORD COUNT:	**1494**

Your essay, appropriately anonymised, may be used to help future students prepare for assessment. Double click this box to opt out of this ☒

Introduction

Over the past few decades, the decrease of global air quality has seriously affected human health, resulting in disastrous weather phenomenon and acute respiratory and cardiovascular disease. Considering the reasons, the widely noticed one is the long-term pollutant emission from non-environmental-friendly human activities. However, the occurrence is also affected by the short-term impact of regional meteorological conditions, such as wind elements, temperature and humidity (Franceschi, Cobo & Figueredo, 2018). According to the concentration of pollutants, it is possible to identify the types of pollution weather and analyze the meteorological sensitive factors that affect the transport and diffusion of pollutants under different meteorological conditions. Furthermore, it is significant for detecting and understanding the casual relationship between meteorological conditions and air pollution concentrations, which could contribute to accurate weather forecast and air pollution management.

Considering this issue, clustering analysis, as an explorative data analysis technique used for identifying similar observations by similar characteristics (Govender & Sivakumar, 2020), has gradually been applied. This strategy has been successfully contributed to atmospheric science since 1970s, including climate and meteorological investigation, while applied to air pollution investigation since 1980s (ibid.). Among the studies, k-means has become the most widely used clustering algorithm to investigate the relationship between local synoptic meteorology and air pollutant behavior. As Figure 1 represents, k-means accounts for high proportions of research on the air pollution clustering measurements for both ground based analyses and formation trajectories. Therefore, this article will analyze the application principles of k-means in the air pollutants behavior related studies in detail, followed by the discussion on the applications' advantages and disadvantages, from data preparation, computing and result interpretation perspectives, for further improvement on the methodology.

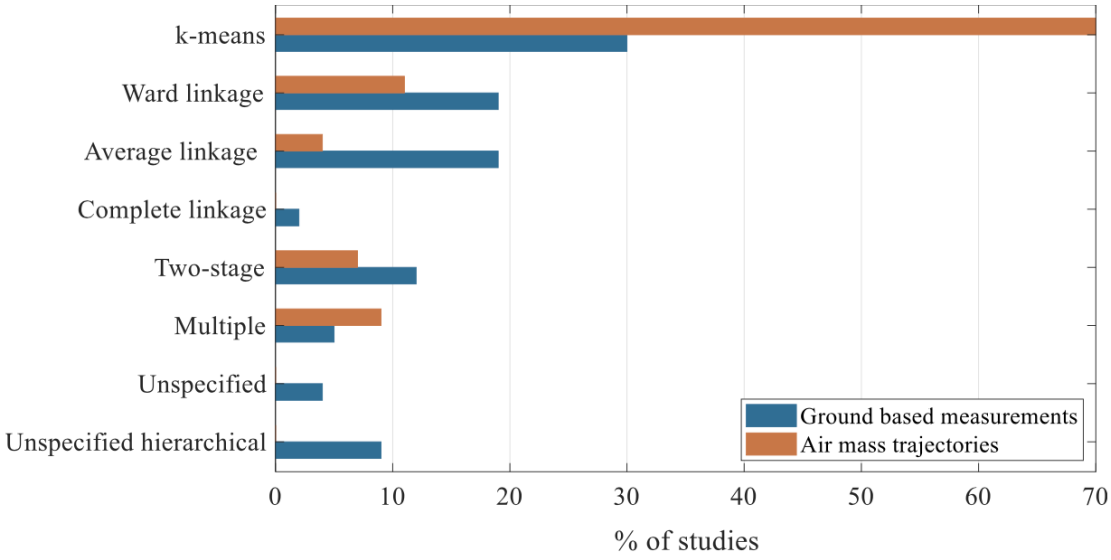

Figure 1: Application proportion of clustering algorithms in previous research (Govender & Sivakumar, 2020)

k-means clustering algorithm for synoptic situation classification

According to Hartigan and Wong (1979), k-means strategy aims to divide several observations in n dimensions into k ($k \leqslant n$) clusters or partitions with the nearest mean, where the sum of squares of within-cluster members should be minimized. Given a set observations, the cluster is composed by a collection of data points with specific similarities in properties or characteristics aggregated together. The primary aim of k-means is dividing the observations into k groups where each of them contains more than one observation. Following the subsequent steps, the k-means algorithm can minimize the total within-cluster variance and optimize the clustering results:

Step 1: Identify a k value and randomly choose k distinct points as cluster center
Step 2: Assign each observations to the cluster with the nearest center.
Step 3: Select the new centers for formed clusters, by calculating the mean.
Step 4: Repeat the last two steps with mean value becoming new cluster center until no change for the last iteration (Chu et al., 2012).

When analyzing the relationship between local synoptic meteorology and air pollutant behavior, the required meteorological element information might include the pressure, wind speed, temperature, relative humidity, evaporation, rainfall, sunshine hours, low-level cloud fraction of each station (Sfetsos & Vlachogiannis, 2010). After computing based on the previously mentioned steps, clusters with k centers could be identified by the similarity of all relevant meteorological attribute values. Finally, according to the characteristics of each cluster, the pollution weather could be predicted.

Data preparation

Due to the limitation of conditions, the change of weather station address or other reasons, some meteorological elements information might not complete. However, the distance-based clustering algorithm is sensitive to data values. In order to ensure effective clustering results, investigating the statistical information such as the data distribution information is necessary for clarifying the data availability and the impact on the results. Especially, a large number of missing values and outliers will cause serious deviation and mislead the prediction (Austin et al., 2012). In addition, if the value range of different attributes is different, the influence on the result would be inconsistent when participating in the distance-based k-means algorithm. In other words, the influence weight of the attributes with larger numerical change on the clustering result is far greater than that of the attributes with smaller range (Dabbura, 2018).

In order to correct this effect, normalization or standardization of the input data would be required. However, if the distribution normality is not considered in the process, uniform quantization of data could provide values tending to be the same after normalization. In contrast, outliers with greater impact on data normalization should be removed, in case they further influence the following data mining process. For instance, Grubbs test for detecting the single outlier in a univariate dataset was applied for the outlier removal process before conducting the traditional k-means algorithm for pollution prediction (Sfetsos & Vlachogiannis, 2010). In addition, missing value cannot participate in distance operation; therefore, meteorological

experts should be consulted. Following their instruction, interpolation or manual correction methods could be applied. However, in many real-world practices, researchers preferred to simply ignore the incomplete datasets, although repeating the analysis with those contributors could probably improve the clustering results and interpretability (Austin et al., 2012).

Data computing

This method with the ability to handle large datasets could be conducted with low complexity and computationally fast (Steinley 2006). Therefore, k-means algorithm has been prevailingly applied to air pollution investigation, although there are natural defects within this method. For k-means, the selection of k value and initial centers should be cautiously specified in advance so that the clustering process could be conducted based on the number of clusters, which are major obstacles.

Indeed, there are several approaches which could contribute to this selection and improve the final clustering results, such as elbow method calculating the additional variation that would be explained by additional clusters (Dabbura, 2018). Sometimes, multiple choices of k values should be tested if the related domain knowledge of the pollutant profiles is limited. For example, Austin et al. (2012) ran the k-means algorithm for k values in the range of 2 to 8 to identify the best choice afterwards. However, there is few methodologies paying much attention on the choice of initial centers. According to Sahafizadeh & Ahmadi (2009), different initial centroids' location could bring different clustering results, which should be cunningly placed. In that case, placing them as far away from each other as possible might be a better choice, but there is still a need to explore more robust approaches.

Result interpretation

The identification of clustering patterns for different regions could be seen as a guidance to adjust the execution of relevant policies. In that case, better buffering approaches could probably mitigate the influence of air pollutants. For example,

temperature adjustment measures could alleviate urban heat island effect and accelerate airflow, while artificial rainfall means could promote wet deposition (Chang, Ni & Li, 2020). However, mistaken result interpretation might lead to policy failure. Therefore, the k-means clustering results require an appropriate method to determine the attributes of each cluster and then make corresponding predictions. For example, Sahafizadeh & Ahmadi (2009) indicated that the decision tree could be applied for the decision analysis, since it is a classification algorithm for calculating conditional probabilities. Adopting this approach, they have computed the decision rules for predictive dusty days based on relationship between air pressure, humidity and monthly dusty days.

From the result interpretation perspective, not only the clustering results should be explained properly, but also its accuracy and significance. In that case, sensitivity analysis has been mentioned by Austin et al. (2012) to determine whether the clustering results have been significantly affected by outlier observations. During the experiment, the complete dataset and random test dataset are clustered in the same way and compared to identify misclassification percentage (ibid.). In that case, the cluster membership could and should be adjusted in k-means algorithm if necessary. In addition, benefiting from the significance calculation, Franceschi, Cobo & Figueredo (2018) were able to comprehensively interpret the forecasting models, recognizing that the k-means clustering results only effective ($p < 0.05$) for high pollution areas, which is useful for formulating reasonable policies for different regions.

Conclusion

In conclusion, the essay reviewed the application of k-means algorithm in air pollution investigation aspect, especially based on meteorological factors. The k-means principle for synoptic situation classification has been explained first. After that, the limitations on the required datasets (such as missing values and outliers) in data preparation process have been identified. Besides, although several solutions have been promoted, they should better be put into practice in the future. Regarding the data computing aspect, easy operating and excellent computing speed are main

advantages of k-means, although the selection of k and initial centers might become a potential risk for final prediction. Finally, from the result interpretation perspective, it is possible to introduce other professional method for decision analysis. Besides, investigating the accuracy and significance of k-means results is also significant. Therefore, the future k-means practice in air pollution investigation should further test the rationality and accuracy of various methods, as well as encourage the introduction of other methods to improve k-means algorithm.

Reference list

Austin, E. et al. (2012). 'A framework for identifying distinct multipollutant profiles in air pollution data', *Environment International*, 45, pp. 112-121.

Chang, V., Ni, P., & Li, Y. (2020). 'K-clustering methods for investigating social-environmental and natural-environmental features based on air quality index*', IT Professional*, 22 (4), pp. 28-34.

Chu, H. et al. (2012). 'Integration of fuzzy cluster analysis and kernel density estimation for tracking typhoon trajectories in the Taiwan region', *Expert systems with applications*, 39 (10), pp. 9451-9457.

Dabbura, I. (2018). *K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks*. Available at: https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a (Accessed: 28 December 2020).

Franceschi, F., Cobo, M., & Figueredo, M. (2018). 'Discovering relationships and forecasting PM10 and PM2.5 concentrations in Bogotá, Colombia, using Artificial Neural Networks, Principal Component Analysis, and k-means clustering', *Atmospheric pollution research*, 9 (5), pp. 912-922.

Govender, P., & Sivakumar, V. (2020). 'Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019)', *Atmospheric Pollution Research*, 11 (1), pp. 40-56.

Hartigan, J., & Wong, M. (1979). 'Algorithm AS 136: A k-means clustering algorithm', *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28 (1), pp. 100-108.

Sahafizadeh, E., & Ahmadi, E. (2009). 'Prediction of air pollution of Boushehr City using data mining', *2009 Second International Conference on Environmental and Computer Science*. pp. 33-36.

Sfetsos, A., & Vlachogiannis, D. (2010). ‘A new approach to discovering the causal relationship between meteorological patterns and PM10 exceedances. *Atmospheric research*, 98(2-4), pp.500–511.

Steinley, D. (2006). ‘K-means clustering: A half-century synthesis’, *British journal of mathematical & statistical psychology*, 59 (1), pp. 1-34.