Atmospheric Pollution Research 11 (2020) 40–56

Contents lists available at ScienceDirect

Atmospheric Pollution Research

journal homepage: www.elsevier.com/locate/apr

Review Article

Application of *k*-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019)

P. Govender*, V. Sivakumar

University of KwaZulu-Natal, School of Chemistry and Physics, Durban, 4000, South Africa

ARTICLE INFO

Keywords:
Hierarchical and *k*-means clustering
Air pollution
Air mass trajectories
Particulate matter

ABSTRACT

Clustering is an explorative data analysis technique used for investigating the underlying structure in the data. It described as the grouping of objects, where the objects share similar characteristics. Over the past 50 years, clustering has been widely applied to atmospheric science data in particular, climate and meteorological data. Since the 1980's, air pollution studies began employing clustering techniques, and has since been successful, and the aim of this paper is to provide a review of such studies. In particular, two well known and commonly used clustering methods i.e. *k*-means and hierarchical agglomerative, that have been applied in air pollution studies have been reviewed. Air pollution data from two sources i.e. ground-based monitoring stations and air mass trajectories depicting pollutant pathways, have been included. Research works that have focused on spatio-temporal characteristics of air pollutants, pollutant behavior in terms of source, transport pathways, apportionment and links to meteorological conditions, comprise much of the research works reviewed. A total of 100 research articles were included during the period of 1980–2019. The purpose of the clustering approach, the specific technique used and the data to which it was applied constitute much of the discussion presented in this review. Overall, the *k*-means technique has been extensively used among the studies, while average and Ward linkages were the most frequently applied hierarchical clustering techniques. Reviews of clustering techniques applied in air pollution studies are currently lacking and this paper aims to fill that gap. In addition, and to the best of the authors' knowledge, this is the first review dedicated to clustering applications in air pollution studies, and the first that covers the longest time span (1980–2019).

1. Introduction

Over the past fifty years, global air quality has shown a decline, a direct effect of human activities such as biomass burning, industrial operations and vehicle emissions (Adame et al., 2012). The presence of aerosols or particulate matter (PM) suspended in the air has become of great concern for a long time, owing to the adverse effects on human health. According to Fullerton et al. (2008), air pollution is a significant cause of morbidity and mortality. In addition, several studies (Zhang and Smith, 2007; Harinath and Murthy, 2012; Laumbach and Kipen, 2012; Bergstra et al., 2018) have demonstrated the association between particulate matter and acute respiratory and cardiovascular diseases, among many others. Due to this, many research works have been undertaken related to air pollution and air quality monitoring, identification of sources, long-range pollutant transport pathways and the development and implementation of effective control and mitigation strategies. Within many of these studies, cluster analysis has become an effective tool for the analysis of air pollutants.

Cluster analysis, or more commonly referred to as “clustering”, is a technique used for the grouping similar observations, data points or feature vectors based on their similar characteristics (Jain et al., 1999). According to Kaufman and Rousseeuw (1990), it is “the art of finding groups in data”. In general, the aim of cluster analysis is to identify groups of similar objects, where objects in a cluster are more similar to each other than objects in different clusters. Clustering can be used for the identification of interesting patterns and distributions and yield possible insights into the underlying structure of the data (Halkidi et al., 2001). Therefore, cluster analysis is a useful technique for discovering and extracting information that may have been previously unnoticed. Cluster analysis was proposed as early as 1930, however its application gained popularity only much later in the 1960s. Clustering techniques have found application across a wide variety of disciplines such as biology, social science, medicine and geography during the 1970s, and in atmospheric science in the 1980s (Gong and Richman, 1995).

Peer review under responsibility of Turkish National Committee for Air Pollution Research and Control.
* Corresponding author.
E-mail addresses: paulenegovender@gmail.com, GovenderP5@ukzn.ac.za (P. Govender).

https://doi.org/10.1016/j.apr.2019.09.009
Received 9 May 2019; Received in revised form 16 September 2019; Accepted 16 September 2019
Available online 19 September 2019
1309-1042/ © 2019 Turkish National Committee for Air Pollution Research and Control. Production and hosting by Elsevier B.V. All rights reserved.

Application of clustering, in particular *k*-means and hierarchical agglomerative techniques, to air pollution data has been conducted since the 1980's, and has since gained considerable attention.

Previous reviews on clustering applications, such as those by Gong and Richman (1995) and Jolliffe and Philipp (2010), have focused mainly on climate and precipitation and to a lesser extent on air pollution. Given that the effects of PM exposure is dangerous to human health, gaining a deeper understanding on the temporal and spatial behavior and dynamics of air pollutants is certainly required. Aghabozorgi et al. (2014) presented a review on spatial and temporal clustering of air pollution, however this study was restricted to Malaysia and did not provide an overview on a global scale. This review aims to present an overview of two commonly used clustering techniques i.e. *k*-means and hierarchical, that have been applied in air pollution studies, particularly those studies which focused on finding patterns and investigating underlying structure of the data, which is the fundamental purpose of cluster analysis. Mainly those studies involving the use of ground-based measurements and air mass trajectories have been included. More specifically, significant contributions presenting clustering applications in air pollution studies reviewed here are categorized as one or more of the following subjects of interest:

(1) Identification of spatial and temporal patterns of air pollutants
(2) Air pollutant exposure and air quality management
(3) Relating air pollutant behavior with local synoptic meteorology
(4) Transport pathways and source apportionment

A review of the literature, mainly published research articles, was undertaken and where the selection of articles was guided by the themes defined above. Google scholar was the main search engine used. The following keywords in different combinations were used in the search: “cluster analysis”, “*k*-means”, “hierarchical”, “clustering”, “air pollution”, “air pollutants”, “spatial”, “temporal”, “variation”, “air mass trajectories” and “air pollution behavior”. In addition, studies were also obtained through cross-referencing. Based on the subjects of interest and together with the keywords, studies during the years 1980–2019 were selected for consideration. Upon evaluation of the abstracts, only those falling into one of the four categories listed above and utilising either of the two clustering methods, were retained for further review. Original research articles available in peer-reviewed journals were considered eligible for inclusion. The literature was also screened for earlier review articles relevant to the present topic. Together with the keywords, inclusion criteria were air pollution studies that have employed clustering of pollutant data obtained through ground-based measurements (category 1), and those that have employed clustering of air mass trajectories used for depicting air pollutant transport pathways (category 2). The literature search for inclusion in this review has been conducted up until April 2019. A total of 103 (57 and 46 in category 1 and 2, respectively) research articles were reviewed.

The review has been organized such that it begins with an overview and a brief discussion of the two commonly used clustering techniques i.e. hierarchical agglomerative (including Ward, single, average, centroid and complete linkages) and *k*-means. Thereafter, the purpose of the clustering approach, the specific technique used and the data to which it was applied constitute much of the discussion. The paper is structured as follows: Section 2 gives a description of the *k*-means and hierarchical clustering methods. Section 3 discusses the choice of the optimal number of clusters, a common issue in the clustering procedure. Some additional well-known clustering methods are discussed in Section 4. Section 5 presents a discussion on the application of the clustering methods under two categories i.e. air pollution data from ground based measurements and air mass trajectories, both of which are related to air pollution studies. Section 6 contains a summary, followed by a brief list of recommendations in Section 7 and some concluding remarks in Section 8.

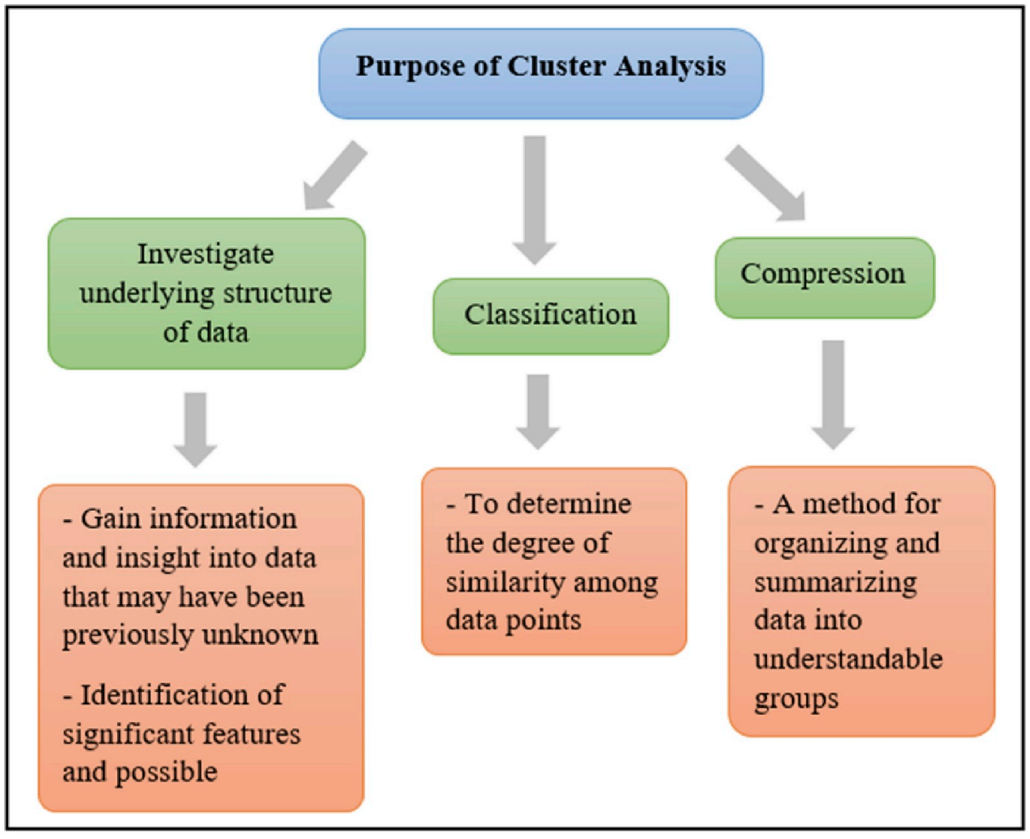

Fig. 1. Schematic diagram describing three main objectives of cluster analysis and the outcome of each objective.

2. Description of cluster analysis techniques

According to Jain (2010), cluster analysis can be categorized into three main objectives. A schematic of these objectives with a description of the outcomes are presented in Fig. 1. Cluster analysis techniques can be broadly classified as hierarchical and non-hierarchical, where the latter is often referred to as partitional. Although there are several cluster analysis techniques, this review is restricted to two that are well known and most commonly used i.e. agglomerative hierarchical and iterative partitional clustering. In particular, single, average, complete, centroid and Ward linkages for hierarchical, and the *k*-means technique for partitional clustering are discussed.

2.1. Hierarchical clustering: single, average, complete, centroid and Ward linkages

Hierarchical clustering techniques recursively find nested clusters either in an agglomerative or divisive manner. Agglomerative clustering is one where each data point starts off in its own cluster and thereafter merges the similar pair of clusters successively resulting in a hierarchy. Alternatively, divisive clustering starts with all the data points in one cluster and repeatedly divides each cluster into smaller ones. Once divisions or fusions are made, they cannot be reversed and thus re-adjustment is not possible with hierarchical clustering (Liao, 2005). An agglomerative hierarchical clustering algorithm is generally implemented as follows:

Step 1: Each observation is considered to be initial clusters.

Step 2: Distances between clusters are computed.

Step 3: Two clusters that have minimum distance are combined and replaced by a single cluster. The distance matrix is then recomputed in order to reflect this merging process.

Step 4: Repeat Step 2 and 3 until there is only a single cluster containing all observations.

The distance or proximity measure is needed to quantify the similarity between objects. The most common is the Euclidean distance. Others include Manhattan, Minkowski and Hamming distances. The output of a hierarchical clustering algorithm is a dendrogram, which is a two-dimensional tree-like structure depicting the sequence of nested clusters (Dubes and Jain, 1976). The distance of each fusion (or division) is also represented on the structure. Cutting the dendrogram at a desired level results in a set of disjoint groups (or clusters). Fig. 2 shows an illustration of a dendrogram showing divisions or fusions at different stages. There are different proximity measures used for combining clusters in hierarchical algorithms. Common ones include single, complete, average, centroid and Ward linkage. These linkage methods are described below.

2.1.1. Single

Single linkage computes the smallest dissimilarity between two objects. The proximity of two clusters is defined by the minimum distance between any two objects belonging to the different clusters. Single linkage is also known as nearest neighbour. This type of linkage method is appropriate for handling non-elliptical shapes, but tends to be sensitive to outliers.

2.1.2. Complete

Complete linkage, the opposite of Single linkage, computes the largest dissimilarity between two objects. The proximity of two clusters is defined by the maximum distance between any two objects belonging to the different clusters. Complete linkage is also known as farthest neighbour. This type of linkage method tends to produce compact clusters and is less sensitive to outliers.

2.1.3. Average

Average linkage is the intermediate between the maximum and minimum distance methods. More specifically, the distances between each object in the first cluster and each object in the second cluster are determined. Thereafter, the average of these distances between all pairs of objects is computed. Average linkage is superior to that of single and complete linkages since it allows for the minimization of within-cluster variance and maximization of the between-cluster variance (Kalkstein et al., 1987).

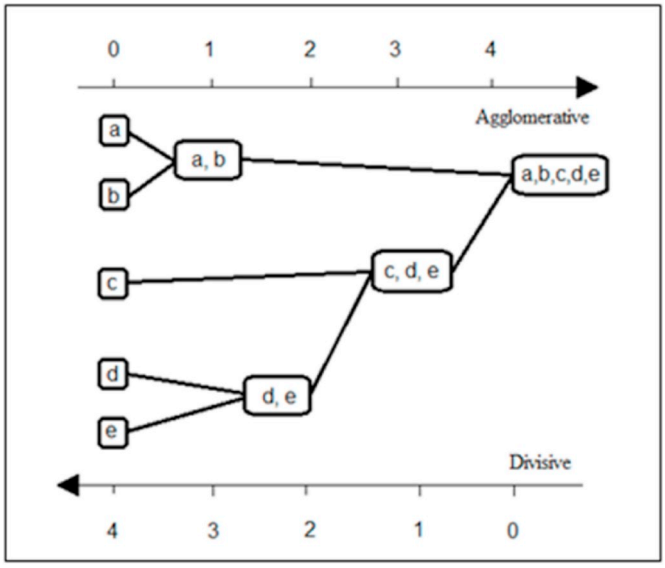

Fig. 2. Illustration of a dendrogram i.e. a two-dimensional structure showing divisions or fusions at different stages. (Adapted from Finding Groups in Data, Kaufman and Rousseeuw, 1990, with permission of the publisher, John Wiley & Sons, Inc.).

2.1.4. Centroid

Centroid linkage is defined as the distance between centres of gravity (centroids) of two clusters. Upon the addition or removal of an object, the centroid is recomputed. This linkage method is more robust to outliers, and tends to perform better than others when dealing with clusters of different sizes (Everitt et al., 2011).

2.1.5. Ward

Ward linkage (or Ward minimum variance method) (Ward, 1963), is defined as the smallest increase in the within-cluster sum of squares due to the merging of two clusters. The Ward's distance between two clusters A and B having centres a and b and frequencies n_A and n_B, is given by

$$d(A, B) = \frac{d(a, b)^2}{n_A^{-1} + n_B^{-1}}, \quad (1)$$

where a and b are the centroids of clusters A and B, respectively (Tuffery, 2011). According to Tuffery (2011), Ward linkage is the one which most closely matches the purpose of clustering and thus the most effective. According to Jolliffe and Philipp (2010), the Ward linkage is a frequently used hierarchical method. A representation of single, complete and average linkages is presented in Fig. 3. For further details on the different linkage methods the reader is referred to Anderberg (1973).

2.2. Partitional clustering: k-means

Non-hierarchical or partitional clustering methods create all the clusters simultaneously by partitioning the data. The discovery of the *k*-means clustering algorithm more than 50 years ago by Steinhaus (1956), (and later by Ball and Hall, 1965; MacQueen, 1967), has led to its application in a variety of fields such as psychology, marketing research, medicine and biology. Although several other clustering

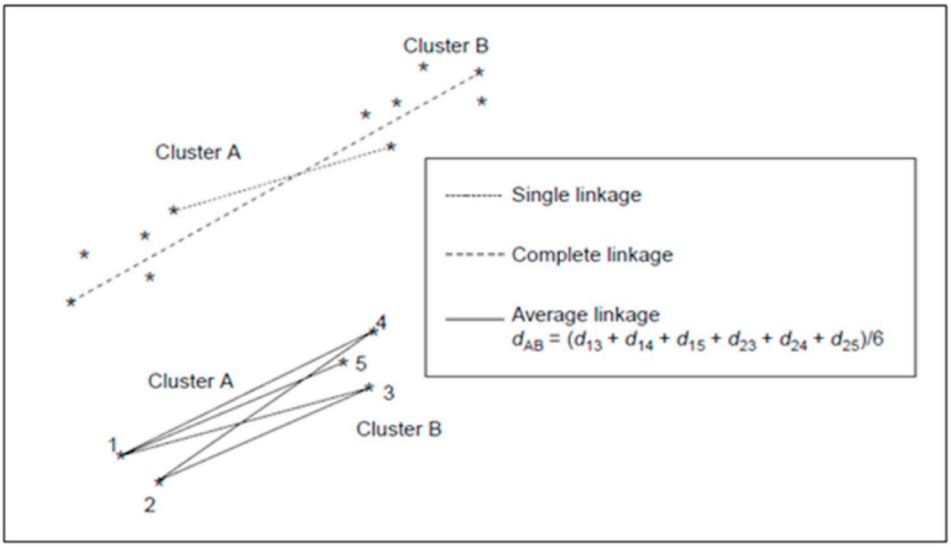

Fig. 3. Representation of single, complete and average linkages (Adapted from Cluster Analysis, Everitt et al., 2011, with permission of the publisher, John Wiley & Sons, Inc.).

algorithms have been developed since then, *k*-means remains one of the most widely used methods owing to its simplicity, ease of implementation and efficiency (Jain, 2010).

According to Hartigan and Wong (1979), the aim of the *k*-means algorithm is to divide *m* objects in *n* dimensions into *k* (where $k \leq n$) partitions (or clusters) such that the within-cluster sum of squares is minimized. As opposed to hierarchical techniques, *k*-means produces a flat clustering structure. The similarity between a pair of objects is defined by their distance, and where among those available, the Euclidean distance is often used as a distance measure. The partition divides the data into *k* groups such that each group contains at least one object. Given a set objects, the primary aim of *k*-means clustering is to optimize the following objective function:

$$J = \sum_{j=1}^{k} \sum_{i \in C_j} ||x_i - c_j||^2, \tag{2}$$

where J is the criterion function, x_i is the ith observation, c_j is the jth cluster center, C_j is the object set of the jth cluster and k represents the number of clusters. Any norm representing the distance between the data object and the center of the cluster is denoted by $||*||$ (Chu et al., 2012). The criterion function attempts to minimize the distance of each point from the center of the cluster to which the point belongs (Halkidi et al., 2001). In general, the *k*-means iterative clustering method is implemented as follows:

Step 1: Choose a *k* value. Use these as the initial set of *k* centroids.

Step 2: Assign each of the objects to the cluster with the nearest centroid.

Step 3: Determine the new centroids of the *k* clusters, by computing the mean of the cluster members.

Step 4: Repeat steps 3 and 4 until there is no change in the criterion function after an iteration (Chu et al., 2012).

There is no guarantee that *k*-means finds the global minimum, but it does find a local minimum for a given initial choice of centroids. In order to check for variation in clustering due to different initial centroids, *k*-means is run several times. Furthermore, the *k*-means algorithm belongs to the family of clustering algorithms which require the a priori specification of a desired number of clusters.

2.3. Two-stage clustering

Two-stage clustering, in this review, refers to the combined use of two clustering methods, where the output from the first is used as input into the second. More specifically, hierarchical clustering is the first stage and *k*-means is the second. The reason for the combination of two such methods and in this order is to use the dendrogram for choosing the number of clusters to be used as "seeds" in the *k*-means algorithm. This approach combines the strengths of both methods, where no a priori specification of the number of clusters is required in the first step and the speed of the second step. Ward, average and centroid linkages usually precede the *k*-means algorithm, and it is often the case where Ward's linkage is one which is commonly used (Tuffery, 2011).

2.4. Other clustering methods

In addition to the clustering methods discussed above, some of the other well-known methods include partitioning around medoids (PAM) or *k*-medoids, hidden Markov model (HMM), mixture models and fuzzy *c*-means which are briefly described. Table 1 provides a comparison of the advantages and disadvantages of the different clustering methods.

2.4.1. Partitioning around medoids

Instead of using the mean value of the objects in a cluster as a reference point, an actual object can be used. The aim is to find the most centrally located object within the cluster, which is referred to as the medoid, and objects that are closest are assigned to the medoid to create clusters (Madulatha, 2012; Omran et al., 2007). According to Han et al. (2012), the absolute-error criterion is defined as

$$E = \sum_{i=1}^{k} \sum_{\boldsymbol{p} \in C_i} dist(p, o_i) \tag{3}$$

where E is the sum of the absolute error for all objects p in the data set, and o_i is the representative object of C_i. This is the basis for the *k*-medoids method, which groups *n* objects into *k* clusters by minimizing the absolute error.

2.4.2. Hidden Markov models

Hidden Markov models belong to the group of model-based clustering, and is useful for time series data. Using a hidden Markov model, each data point represents the observed value of a time series at time *t*. The model consists of two parts: (i) the time series observation and (ii) the unobserved state variables through which the time series observations are generated.

The model is characterized by a set of states, an initial probability distribution for the first state, a transition probability matrix linking

Table 1
Comparison of the advantages and disadvantages of the clustering methods (Bhagat et al., 2016; Namratha and Prajwala, 2012, Omran et al., 2007).

Clustering method	Advantages	Disadvantages
***k*-means**	• Low complexity • Computationally fast • Ability to handle large data sets • Cluster membership can be adjusted	• Number of clusters need to be specified in advance • Sensitive to outliers • Inability to deal with non-convex clusters of varying size and density • Sensitive to scale of the data set • Different initial centroids produce different results
***k*-medoids**	• Less sensitive outliers • Cluster membership can be adjusted	• Number of clusters need to be specified in advance • Different initial centroids produce different results
Hierarchical (single, complete, average, centroid and Ward linkage)	• Does not require the number of clusters to be specified in advance • Dendrogram provide graphical representation • Ability to detect clusters of varying shapes and sizes	• High complexity • Computationally slow • Once clusters are formed no adjustments can be made • The level of cutting of the dendrogram may be difficult to decide • Clusters dependent on the distance metric used
Hidden Markov model	• Flexibility in handling various types of data	• Requires many parameters • Requires large data sets
Mixture models	• Clusters can be characterized by a small number of parameters	• Computationally expensive if the number of distributions is large or the data set contains very few observed data points. • Requires large data sets • Difficult to estimate the number of clusters
Fuzzy *c*-means	• Allows for cluster assignment flexibility • More realistic in terms of being able to give the probability of belonging to a cluster	• High complexity • Number of clusters need to be specified in advance • May converge to a local optima

successive states, and state-dependent probability distributions responsible for generating the time series data. Only the time series observations are visible to the observer while the state variables are hidden. The hidden Markov model provides statistics such the mean, standard deviation and weight values for a cluster, according to the observations comprising the cluster (Gómez-Losada et al., 2014; Gómez-Losada et al., 2018).

2.4.3. Mixture models

In some cases it is insufficient to describe a data set using one distribution. This is particularly common when the data set is derived from two or more sub-populations. Therefore, it becomes necessary to fit a composition of distributions to the underlying data set, where such distributions are called mixture models. Mixture models are defined by the parameters specific to each component and the proportion in which the mixed components occur. Clustering of data objects is achieved through determining the parameters of the components and thus classifying each data object by the respective component. Variables such as the mean and variance are used to characterize these models. Several techniques may be used to fit the mixture model distributions for e.g. graphical methods, the method of moments, maximum likelihood estimation (MLE) and Bayesian approaches. The most widely-used approach is the Expectation-Maximization (EM) algorithm for the MLE of mixture model distribution (Gómez-Losada et al., 2014).

2.4.4. Fuzzy c-means

Traditional clustering approaches create partitions, where within the partition an object belongs to one and only one cluster. These are therefore "hard" clusters. Fuzzy clustering was intended to alleviate this problem by providing an object to be associated with a cluster using a membership function. Fuzzy clustering can be converted into hard clustering by assigning each object to the cluster with the largest membership value (Jain et al., 1999). The fuzzy *c*-means is the most well-known fuzzy clustering algorithm. The algorithm attempts to minimize the objective function, called the *c*-means function defined as:

$$J_m = \sum_{j=1}^{k} \sum_{i=1}^{N} u_{ij}^{m} \, || x_i - c_j ||^2 \tag{4}$$

$$u_{ij} = \left\{ \sum_{i=1}^{k} \left(|| x_i - c_j || / || x_i - c_l || \right)^{2/(m-1)} \right\}^{-1}, \tag{5}$$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} X_i}{\sum_{i=1}^{N} u_{ij}^{m}}, \tag{6}$$

where N is the number of observations, m is any real number greater than 1, referred to as the fuzziness coefficient and u_{ij} represents the degree of membership of x_i in jth the cluster (Chu et al., 2012).

3. Choice of the optimal number of clusters and cluster validity

One of the primary difficulties with the application of clustering algorithms for the extraction of meaningful information and patterns from data, is the choice of the optimal number of clusters. This has been a topic of much debate and is still considered to be an important, yet largely unresolved, aspect of clustering analysis. The first comprehensive investigation of several procedures for determining the number of clusters was by Milligan and Cooper (1985), where 30 different evaluation criteria were tested on artificial data sets.

For hierarchical and partitional clustering techniques, there are several criteria that can assist the researcher in choosing and validating the optimal number of clusters. These include silhouette index, Dunn index, Calinski-Harabz (CH), Davies-Bouldin (DB), Rand index, Bayes Information Criterion (BIC), Akiake Information Criterion (AIC) and Gap statistics, within-cluster sum of squares (WCSS), between-cluster sum of squares (BCSS) and semi-R^2, to name a few. In the air pollution studies by Beddows et al. (2009), Wegner et al. (2012) and von Bismarck-Osten and Weber (2014), the silhouette index and Dunn index were mostly applied. The silhouette index is a measure of cluster compactness and separation. Silhouette index values range from −1 to 1, indicating not well-clustered and well-clustered observations, respectively. Observations with value 0 indicates those on the border of two clusters. The Dunn index is defined as the ratio of the minimum distance between two observations belonging to different clusters, divided by the maximum distance between two observations belonging to one cluster. A higher Dunn index value indicates a more optimal clustering solution (Wegner et al., 2012; von Bismarck-Osten and Weber, 2014). Further details on cluster validation can be found in Halkidi et al. (2001), Kaufman and Rousseeuw (1990) and Tuffery (2011). Despite the several criteria available to assist the user in selecting the appropriate number of clusters, the choice of the optimal cluster number also depends on the application and the number that will yield the most meaningful clusters. Therefore, it is left to the researcher to evaluate more than one method in order to find which is best for the data at hand.

4. Particulate matter exposure and associated risk to human health

Hazardous chemicals expelled into the environment through a variety of natural and anthropogenic activities result in adverse effects on human health and the environment. The various commonly-encountered pollutants such as carbon monoxide (CO), sulphur dioxide (SO_2), nitrogen oxides (NOx), volatile organic compounds (VOCs), ozone (O_3), $PM_{2.5}$ and PM_{10}, show differences in their chemical composition, reaction properties and ability to be diffused either over long or short distances.

Pollutant exposure has both acute and chronic effects on human health, including respiratory, lung and cardiovascular diseases. Moreover, premature mortality and reduced life expectancy have also been associated with short-term and long-term pollutant exposures (Kampa and Castanas, 2008).

For many years now the health effects associated with PM exposure has been widely recognized. Suspended in the air, these particles vary in size and composition and consist of mixtures that are complex and variable. Industrial processes, factory and power plant operations, motor vehicles, construction activities, fires and dust constitute some of the major contributors to PM sources. PM can be classified into three categories based on their size in terms of aerodynamic diameter: (1) ultrafine particles (smaller than 0.1 μm), (2) fine particles, (smaller than 1 μm) and (3) coarse particles (larger than 1 μm). Deposition of the PM particle within the respiratory tract is dependent on its size. PM_{10} particles accumulate mainly in the region of the upper respiratory tract, while fine and ultra fine particles have the ability to penetrate into the region of the lung. In addition, ultra fine and fine particles have been found to have worse effects than coarse particles, in terms of mortality, cardiovascular and respiratory effects (Kampa and Castanas, 2008). Despite the size, the presence of PM in the air is nevertheless a significant concern to public health.

Environmental monitoring of air pollutant levels is therefore important to understand to pollutant behavior. In addition, efforts aimed at reducing pollutant levels are required to be greatly intensified, which includes analysing and understanding the temporal and spatial patterns of pollutant behavior, and implementation of appropriate and effective short and long-term control and mitigation strategies. These factors are crucial to ensure that the risks to human health are minimized. Clustering analysis can be used as a tool in the analysis of many of the above aspects. This review highlights many of the studies where this has been successfully achieved.

5. Clustering applications

This section presents the applications of hierarchical (Ward, single, average, centroid and complete linkages) and partitional (*k*-means) clustering techniques in air pollution studies spanning almost 40 years. Particularly those investigations focusing on spatial and temporal patterns, exposure and air quality management, source apportionment, transport pathways and relating air pollutants to synoptic meteorology, are presented. The discussion is divided into two categories i.e. (1) ground-based air pollution measurements and (2) air mass trajectories. Generally, the clustering techniques are referred as hierarchical (including Ward, single, average, centroid and complete linkages), *k*-means, two-stage clustering (hierarchical followed by *k*-means) and multiple methods (the use of more than one clustering method).

5.1. Ground-based air pollution measurements

Several works (Zhang and Smith, 2007; Fullerton et al., 2008; Harinath and Murthy, 2012; Laumbach and Kipen, 2012; Bergstra et al., 2018) have highlighted the impact of aerosol pollutants on various respiratory-related illnesses. Therefore, identification and analysis of the spatial and temporal patterns of pollutants are important aspects for understanding the levels and behavior of the individual species, so that appropriate control and mitigation strategies could be implemented. Clustering has been a useful tool for this purpose, where hierarchical methods have been widely used (Kalkstein and Corrigan, 1986; Cheng et al., 1992; McGregor and Bamzelis, 1995; Greene et al., 1999; Saksena et al., 2003; Flemming et al., 2005; Beaver and Palazoglu, 2006; Giri et al., 2007; Grivas et al., 2008; Pires et al., 2008a; Pires et al., 2008b; Contini et al., 2010; Gao et al., 2011; Lu et al., 2011; Unal et al., 2011; Dominick et al., 2012; Latif et al., 2014; Pandey et al., 2014; Iizuka et al., 2014; Solazzo and Galamarini, 2015; Kahya et al., 2017; Kwon et al., 2018; Cakmak et al., 2018; Qiao et al., 2018; Soares et al., 2018), along with the *k*-means method (Crecelius et al., 1980; Gorham et al., 1984; Sausy et al., 1987; Sanchez et al., 1990; Comrie, 1996; Omar et al., 2005; Kim et al., 2008; Jin et al., 2011; Adame et al., 2012; Shi et al., 2014; Munir et al., 2015; Lyapina et al., 2016; Zhao et al. 2016, 2018; Davuliene et al., 2019). In this review, there have also been studies (Kalkstein et al., 1987; Cheng et al., 1992; McGregor and Bamzelis, 1995; Greene at al. 1999; Cakmak et al., 2018) investigating the relationship between aerosol pollutant behavior and climate, in particular, to identify local synoptic events associated with high (or low) pollutant levels. Two-stage clustering and the use of multiple methods are significantly fewer (Davis and Kalkstein, 1990; Davis and Gay, 1993; Eder et al., 1994; Davis et al., 1998; Lu et al., 2006; Beddows et al., 2009; Austin et al., 2012; Hsu and Cheng, 2016, 2019). Table 2 presents a summary of the literature that have applied hierarchical and *k*-means techniques to ground based measurements for air pollution studies. Each author is listed together with the clustered data/variable, method, and if more than one clustering method was applied, the method which was found to have the best performance was indicated.

5.1.1. Clustering techniques for analysis of spatial and temporal variation of pollutants

An analysis of air quality in Delhi, India was carried by Saksena et al. (2003), where the average linkage clustering approach was used for classification of sulphur dioxide (SO_2), nitrogen dioxide (NO_2) and suspended particulate matter (SPM) criteria pollutants, using data from 9 stations. Clusters represent spatial patterns of the pollutants. All pollutants were classified into 2 clusters. In general, results showed no statistically significant differences that existed in the mean concentration of all pollutants between stations belonging to different land-use types (residential and industrial). Giri et al. (2007) used Ward's clustering to gain an understanding of spatial air pollution, in particular PM_{10}, in the Kathmandu Valley of Nepal. Individual cluster types were based on PM_{10} concentrations from 6 air-quality monitoring sites, during pre-monsoon, monsoon, post-monsoon and winter seasons. Two seasonally independent clusters of similar PM_{10} concentration characteristics were identified. These two groups represent valley and urban backgrounds and main city area associated with high commercial and vehicular activity. Cluster characteristics during the monsoon, post-monsoon and winter seasons were analyzed. For example, a reduction in PM_{10} concentration was expected during the monsoon, however this reduction was only observed in certain areas, thus indicating the limited decreasing effect of the monsoon. Kim et al. (2008) examined the temporal $PM_{2.5}$ patterns across the U.S. for the characterization of spatially homogeneous regions. A total of 522 monitoring sites over a period of 5 years was used in the analysis. The types of *k*-means clusters identified were temporal (annual resolution) and seasonal trends of $PM_{2.5}$ concentration. Clustering produced 6 regions that exhibit homogenous temporal $PM_{2.5}$ concentration patterns which were: Central, Florida/Gulf Coast, Midwest, Northeast, Southeast, and West regions. Within each spatially homogenous region, distinct temporal patterns were observed. It was found that higher $PM_{2.5}$ concentrations occur in winter in the western part of the U.S., but in summer in the northeastern and southeastern regions. A study investigating air pollution in China was carried out by Gao et al. (2011). This study focused

Table 2
Summary of the literature that have applied hierarchical, *k*-means and two-stage clustering techniques to ground-based pollutant measurements in air pollution studies.

Year	Author	Data/variables	Clustering approach/method	Best performing method if more than one used
1980	Crecelius et al.	Particulate elements	*k*-means	
1984	Gorham et al.	Several ions related to air pollution, agriculture and sea spray	*k*-means	
1986	Kalkstein and Corrigan	Air and dew point temperature, pressure, wind speed, cloud cover, visibility, SO_2	Hierarchical	
1987	Sausy et al.	Particle elements	*k*-means	
1990	Davis and Kalkstein	Air and dew point temperature, pressure, wind components, cloud cover	Two stage: Average linkage and *k*-means	
1990	Sanchez et al.	PM concentrations and several meteorological variables.	*k*-means	
1992	Cheng et al.	O_3, TSP	Hierarchical	
1993	Davis and Gay	Air and dew point temperature, geopotential height, wind speed and direction, aerosol data	Two stage: Average linkage and *k*-means	
1994	Eder et al.	O_3, air and dew point temperature, pressure, cloud cover, wind speed	Two stage: Average linkage and *k*-means	
1995	McGregor and Bamzelis	SO_2,NO_2, O_3, NO, CO, PM_{10},	Hierarchical	
1996	Comrie	850 mb geopotential height	*k*-means	
1998	Davis et al.	Temperature, pressure, humidity, cloud cover, wind speed, O_3.	Two-stage: Average linkage and k-means	
1999	Greene et al.	Air and dew point temperature, cloud cover, pressure and wind speed and direction	Hierarchical	
2003	Saksena et al.	SO_2, NO_2, PM	Hierarchical	
2005	Flemming et al.	PM_{10}, SO_2, NO_2, O_3	Hierarchical	
2005	Omar et al.	Microphysical and optical properties of aerosols	*k*-means	
2006	Gramsch et al.	PM_{10}, O_3,	Hierarchical	
2006	Lu et al.	PM_{10}	Hierarchical, *k*-means, self-organizing maps	*k*-means, self-organizing maps
2007	Giri et al.	PM_{10}	Hierarchical	
2008	Beaver et al.	Wind, O_3	Hierarchical	
2008	Grivas et al.	PM_{10}	Hierarchical	
2008	Kim et al.	$PM_{2.5}$	*k*-means	
2008a	Pires et al.	SO_2, PM_{10}	Hierarchical	
2008b	Pires et al.	CO, NO_2, NO_x, O_3	Hierarchical	
2009	Beddows et al.	Particle size	*k*-means, *k*-median, fuzzy and model-based clustering	*k*-means
2010	Contini et al.	Several PM_{10} ionic species	Hierarchical	
2011	Jin et al.	O_3	*k*-means	
2011	Gao et al.	PM_{10}, SO_2, NO_2	Hierarchical	
2011	Lu et al.	SO_2, NO_2, RSP	Hierarchical	
2011	Unal et al.	PM_{10}, wind speed and direction, precipitation, pressure	Hierarchical	
2012	Adame et al.	O_3, NO_2 and SO_2	*k*-means	
2012	Dominick et al.	O_3, CO, NO, NO_2, SO_2, PM_{10}, temperature, humidity, wind speed	Hierarchical	
2012	Wegner et al.	Aerosol size data	*k*-means	
2013	Austin et al.	A set of 20 p.m.$_{2.5}$ components	Two-stage: Ward linkage and *k*-means	
2014	Hussein et al.	Particle size data	*k*-means	
2014	Iizuka et al.	NO*x*O*x*, PM, NmHC	Hierarchical	
2014	Latif et al.	O_3, CO, NO, NO_2, NOx, SO_2, PM_{10}, CH4, THC, NmHC	Hierarchical	
2014	Pandey et al.	PM_{10}, $PM_{2.5}$, $PM_{1.0}$, SO_2, NO_2	Hierarchical	
2014	Shi et al.	NO_2, SO_2	*k*-means	
2014	von Bismarck-Osten and Weber	Particle number size data	*k*-means	
2015	Huang et al.	$PM_{2.5}$	Hierarchical	
2015	Munir et al.	O_3	*k*-means	
2015	Solazzo and Galamarini	O_3	Hierarchical	
2016	Cakmak et al.	O_3, $PM_{2.5}$, air and dew point temperatures, pressure, wind speed, cloud cover	Hierarchical	
2016	Hsu and Cheng	Wind speed and direction, $PM_{2.5}$	Two-stage clustering: unspecified linkage and *k*-means	
2016	Wang et al.	Trajectories, $PM_{2.5}$, SO_2, NO_2, temperature, humidity, wind speed and direction	Clustering method not specified	
2016	Lyapina et al.	O_3	*k*-means	
2016	Zhao et al.	$PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, O_3	*k*-means	
2017	Kahya et al.	$PM_{2.5}$	Hierarchical	
2018	Cakmak et al.	Ozone, $PM_{2.5}$, air and dew point temperatures, pressure, wind speed, cloud cover	Hierarchical	
2018	Kwon et al.	GHG	Hierarchical	
2018	Qiao et al.	$PM_{2.5}$	Hierarchical	
2018	Soares et al.	SO_2, NO_2	Hierarchical	
2018	Zhao et al.	$PM_{2.5}$, O_3	*k*-means	
2019	Davuliene et al.	BC	*k*-means	
2019	Hsu and Cheng	Wind speed and direction, sea level pressure, $PM_{2.5}$, PM_{10}, O_3	Two-stage clustering: unspecified linkage and *k*-means	
2019	Zhang et al.	Trajectories at 500 m, Pb, CO, O_3, SO_2, NO_x	Clustering method not specified	

on air pollution on regional scale and where the daily air pollution index (DAPI) from 81 cities were studied. The DAPI was analyzed, using average linkage clustering, to yield cities with a similar distribution of pollution levels. At the first level the clusters represented two regimes i.e. north and south, that consisted of cities with similar DAPI. Thereafter, the two DAPI regimes were divided into 7 clusters where the clusters represented the frequency of the DAPI for the cities contained within them. Jin et al. (2011) demonstrated the use of *k*-means clustering to identify spatial ozone pollution regimes over San Joaquin Valley (SJV) of California. Clusters (or regimes) identified through the *k*-means approach represent locations of similar ozone spatial distribution. Of the six regimes identified, two corresponded to low, three to moderate, and one to high ozone clusters. Meteorological measurements were used to explain ozone spatial distributions, and their correlation to those in the San Francisco Basin (SFB). Currently, the existing measurement sites were able to capture spatial patterns of ozone within the SFB and Scaremento Valley, although those locations along the western part SJV were under-represented. To investigate spatio-temporal variations in PM_{10} concentrations and identify monitoring sites with similar pollutant behavior in Turkey, Instanbul, Unal et al. (2011) applied Ward's clustering to data recorded at 10 sites over a 5 year period. The clusters identified were representative of the temporal variability of daily PM_{10} concentrations. Winter, summer and annual clustering revealed 5 distinct PM_{10} regions, with significant variation across the city. Six days of the week were found to have higher concentrations, and winter and summer seasons are characterized by high and low concentrations, respectively. In addition, PM_{10} concentrations were highest for low wind speeds and winds originating from SW and ENE directions. Dominick et al. (2012) investigated air pollutant (O, O_3, PM_{10}, SO_2, NO_2, NO) sources and spatial patterns in Malaysia, using Ward's clustering technique. Clusters represent monitoring stations in order to identify spatial patterns of air quality. Measurements of pollutant species together with temperature humidity and wind speed from 8 ground based stations were analyzed. The stations were classified into 3 clusters, which comprised of main city centres (cluster 1), residential and commercial areas (cluster 2) and industrial areas (cluster 3). From all pollutants analyzed, it was found that PM_{10} contributed the highest pollution levels at all stations. A study investigating pollutant clustering at the city-scale was by Austin et al. (2013), where $PM_{2.5}$ data from 109 monitoring sites were clustered using by the two-stage approach i.e. Ward linkage followed by *k*-means. Cluster types from the two-stage approach were made up of the different geographical regions that showed similar ratios of a variety of $PM_{2.5}$ elements. The resulting groups represent broad differences in emissions. Similar to Dominick et al. (2012), Latif et al. (2014) also performed an analysis of air pollutant behavior in Malaysia using Ward's method. However, the study included additional organic pollutants such as methane (CH4), total hydrocarbon (THC) and non-methane hydrocarbon (NMHC) over a significantly longer period (15 years), and was restricted to the Malaysian peninsular. Four temporal variation pattern clusters were identified from hourly-resolution data of several air quality variables. Evaluation of spatial and seasonal variations in PM_{10}, $PM_{2.5}$, $PM_{1.0}$, SO_2 and NO_2 in a Jharkhand coalfield in India, was conducted by Pandey et al. (2014), using average linkage clustering. Cluster types were comprised of monitoring sites depicting similar behavior in terms of the pollutant dispersion and spatial variations. Five air quality monitoring sites were used and results showed that, for all the sites, concentrations of all types of PM were found to be the highest during the winter season, followed by summer and rainy seasons. Shi et al. (2014) conducted a similar study in Xiamen, China using the *k*-means approach however, only to NO_2 and SO_2, and in an urban coastal location. Cluster types consisted of diurnal pollutant variations, which were classified into 3 and 4 clusters for NO_2 and SO_2, respectively. While many solutions for *k* were tested, the present study could not provide new diurnal patterns compared to classifications made with smaller values of *k*. Nevertheless, the clustering results produced provide useful insights into air quality regimes and pollution levels in Xiamen. Characteristics of regional distribution of $PM_{2.5}$ in Xi'an, China were investigated by Huang et al. (2015), using a hierarchical method. Individual cluster types were based on $PM_{2.5}$ concentrations from 13 monitoring sites that were divided into 3 clusters. Some of the main findings include, temporal distribution of $PM_{2.5}$ concentration is higher in winter, followed by autumn, spring and summer. In terms of the spatial distribution characteristics, the highest concentration was found to be located within 3 of the 13 monitoring sites. In addition, it was found that the distribution of $PM_{2.5}$ was not related to the geographical locations, and variation in the concentration of $PM_{2.5}$ was mostly a result of industrial activities. Munir et al. (2015) studied the temporal variations of O_3 in Makkah, Saudi Arabia. Clusters of 4 and 12 diurnal cycles were produced. These clustering solutions correspond to seasonal and monthly cycles, respectively. Zhao et al. (2016) performed an assessment of the air quality in 31 Chinese cities, by analysis of 6 criteria pollutants ($PM_{2.5}$, PM_{10}, CO, NO_2, SO_2 and O_3) with respect to their annual and diurnal variation, in order to better understand the pollution situation in China. Annual, seasonal and diurnal variation cluster types were established separately for each of the 6 pollutants. In terms of annual variations, clustering of $PM_{2.5}$ and SO_2 concentrations divided cities into 3 and 5 groups, respectively, and based on PM_{10}, CO, NO_2 and O_3 divided cities into 4 groups. With the exception of O_3, CO and SO_2, the concentrations of pollutants in winter months were significantly higher than in other months. The most polluted cities were mainly located in North China Plain and northeastern China during April 2014 to March 2015 due to rapid economic growth and more industrial emissions within the regions. Slightly polluted cities were mainly in the southern region and those cities with high altitude. Diurnal variation of O_3 concentrations showed opposite trends to that of other pollutants. Kahya et al. (2017) applied Ward's clustering for investigating the spatial and temporal $PM_{2.5}$ distributions in Turkey. Cluster types consisted of 13 monitoring stations that were grouped into 5. Two sites exhibited the highest $PM_{2.5}$ concentrations which were attributed mainly to fossil fuel heating. Four urban stations were found to exceed the standard limit during the duration of the study period. Qiao et al. (2018) used average linkage clustering to investigate $PM_{2.5}$ source apportionment and contributions from 25 Chinese cities, derived from a Community Multiscale Air Quality (CMAQ) model. Nine clusters represented groups of cities possessing similar contributions. Results showed highest and lowest annual $PM_{2.5}$ concentrations for the northern and coastal southern and eastern cities, respectively. Model uncertainties in source apportionment and contribution estimations may be attributed to the unstandardized emission inventories available for China. Zhao et al. (2018) identified the potential source regions of $PM_{2.5}$ and O_3 reaching the Sichuan Basin in China, by applying *k*-means clustering to data from 22 sites. Cluster types comprised of hourly $PM_{2.5}$ and O_3 data grouped into 4 and 5 clusters, respectively. Highest concentration of $PM_{2.5}$ were mainly located in the western and southern regions of the Sichuan basin. For O_3, large differences in variations were observed among the cities, with high levels in the south and eastern parts of the basin. In addition, trajectories arriving at 500 m were assigned to clusters using Ward's method. Major transport pathways resulting in increased $PM_{2.5}$ and O_3 concentrations in winter and summer, respectively, were identified by combining trajectories with pollutant concentrations. The Tibetan Plateau was considered to be an important source region of high O_3, particularly for the western region. Other recent studies involving clustering for pollutant assessment in China includes that of Zhang et al. (2019). Davuliene et al. (2019) presented an analysis on black carbon (BC) variation in Preila, Lithuania over an 8-year period, using the backwards trajectory clustering scheme developed by Bycenkiene et al. (2014) (discussed later). Lowest and highest concentrations of black carbon were observed during the winter and summer seasons, respectively.

5.1.2. Clustering techniques for air quality monitoring and optimization of monitoring networks

Flemming et al. (2005) produced an air quality classification scheme for PM_{10}, O_3, SO_2 and NO_2 pollutants in Germany, using Ward linkage. Clusters represented the daily, weekly, and annual cycles of the 4 pollutants. Three clustering techniques, Ward's method, *k*-means, and a combination of self-organizing maps (SOMs) and *k*-means, were used by Lu et al. (2006) for wintertime PM_{10} concentrations in Taiwan. Comparison of all three techniques suggested that 71 stations comprising the air quality monitoring network can be classified into 5 air quality classes (or regions), with the combined approach yielding the best results. Individual clusters represented the different regions of the country, based on the PM_{10} distribution. Of the 5 regions, PM_{10} pollution levels were found to be the highest and the lowest for the southern and eastern regions of Taiwan, respectively. The use of clustering for the management of air quality monitoring stations in the Oporto metropolitan area of Portugal, was examined by Pires et al. (2008a). Focusing on sulfur SO_2 and PM_{10} concentrations, the aims were to identify areas exhibiting similar pollution trends and find the source of these emissions. From a set of 10 monitoring sites, clustering by the average linkage method was used to create 6 clusters, no more than 2 clusters for SO_2 and PM_{10}, respectively. Cluster types consisted of SO_2 and PM_{10} concentrations from the various monitoring sites. Findings showed that many areas are characterized by the same air pollution patterns, yet are covered by more than one monitoring station, suggesting ineffective management of the number of stations within the network. The additional stations could possibly be moved to areas which currently lack coverage, thus expanding the overall coverage of the network. One main source of emission was found for SO_2 while three sources were found for PM_{10}. Using the same clustering approach, Pires et al. (2008b) extended their previous work by considering concentrations of carbon monoxide CO, NO_2, and O_3 for the same location. Clustering revealed 3 different groups for CO and for NO_2 and 2 groups for O_3. Cluster types were similar to those presented in Pires et al. (2008a), but for different pollutants. Two of the monitoring sites showed differences in the behavior of the pollutants due to the location of dominant emission sources and location topography. Using average linkage clustering, Lu et al. (2011) conducted a similar study to Pires et al. (2008a, 2008b), for the assessment of an air quality monitoring network in Hong Kong. However, the analysis was restricted to SO_2, NO_2 and respirable suspended particulates (RSP). Similar to Pires et al. (2008a, 2008b), clusters represent pollutant (SO_2, NO_2, RSP) concentrations from the monitoring sites. Results of this study were intended to be used for optimization of the number of monitoring stations over the region. This was because there currently are redundant monitoring stations contributing to the ineffective management of air quality which required relocation. Iizuka et al. (2014) adopted the use of Ward's clustering for four pollutants (NO*x*, O*x*, PM and NMCH) in the Kanto region of Japan. Using air pollution measurements for 2 non-consecutive years, a total of 476 monitoring stations were grouped into 8. Monitoring stations based on the similarity of the individual pollutant behavior comprised the different cluster types. Reducing the number of monitoring stations was also explored, where 3 simple criteria were proposed. Upon application of this criteria, a significant portion of monitoring stations could be removed however, details of this removal method were not discussed. Clustering by complete linkage was employed by Kwon et al. (2018) for the analysis of greenhouse gas emissions (GHG) information from 24 Asian countries recorded for a decade. Six GHG groups (A-F) were established, thus providing a grouping of countries with similar major and minor GHG emissions characteristics. The results of the study have implications for the development of further climate mitigation strategies in Asia. Similar to Shi et al. (2014), Soares et al. (2018) also investigated NO_2 and SO_2 pollutants, but by average linkage clustering. One of the aims was to analyse and optimize air quality monitoring networks in Canada. Cluster types were based on spatial distributions of the individual pollutants.

5.1.3. Clustering techniques for correlating pollutant concentrations with specific synoptic conditions

Kalkstein and Corrigan (1986) used Ward's clustering method to investigate the relationship between SO_2 concentrations and meteorology for Wilmington, Delaware. Ten clusters were identified describing the synoptic categories linked to the different SO_2 concentration levels. Linear regression was used to examine the relationship between individual weather variables and SO_2 concentrations, however the relationships were found to be rather weak. Studies that involved the use of clustering techniques for relating air quality parameters, such as O_3, to meteorology was conducted by Eder et al. (1994) and Davis et al. (1998). Two-stage clustering i.e. average linkage followed by *k*-means was the preferred approach, since it showed to have the best performance in terms of cluster cohesion. Furthermore, it was successfully applied in previous studies (Davis, 1991; Davis and Walker, 1992; Davis and Gay, 1993) that have aimed to link air pollution and air quality with meteorology. Average linkage clustering was employed by Cheng et al. (1992) to investigate pollution concentrations O_3 and total suspended particles (TSP)) in Philadelphia during the summer season. The Temporal Synoptic Index (TSI) developed earlier by Kalkstein et al. (1987) was used again as the clustering variable to establish 10 cluster types of major summertime synoptic categories. Pollutant concentrations associated with each category showed the highest concentration for continental air mass, which is characterized by low amounts of cloud cover, high air temperature and pressure, a large dew point depression and moderate south-westerly winds. Alternatively, continental polar, related to condition of low air pressure and dew point temperature, had the lowest pollution concentration. Comrie (1996) identified 6 atmospheric circulation patterns within the U.S.-Mexican border region, with the aim of linking weather patterns with ozone pollution for the region. Application of *k*-means clustering to 850 mb geopotential height data over a 32-year period, produced 6 major circulation patterns. Cluster types represent the 6 climatologically characteristic atmospheric circulation patterns consisting of cyclonic and anticyclonic flow. Periods of high ozone pollution were found to be during the presence of high pressure systems and hence during summer. An assessment of synoptic weather patterns on pollutant concentrations recorded in summer for four US cities with varying climate conditions was conducted by Greene et al. (1999). Using a number of weather variables including air and dew point temperature, cloud cover, pressure and wind speed and direction, the aim was to develop a TSI that can be used be used to identify synoptic events associated with specific pollution episodes. Average linkage clustering was applied, producing between 8 and 10 synoptic categories for all cities. These cluster types represented synoptic categories that describe the different atmospheric conditions which usually occur at a given location and which were associated with specific meteorological characteristics. Overall, it was found that the four cities differ in their pollution loadings under the different synoptic patterns. Nevertheless, this study has further implications for the relationship between climate and pollution and its effect on human mortality. Gramsch et al. (2006) applied hierarchical clustering to O_3 and PM_{10} concentrations for Santiago, Chile, and demonstrated that these two pollutants had similar cluster characteristics, suggesting that the concentrations of these pollutants were controlled by meteorological and topographical factors. The four clusters established represent spatial similarity between stations based on the O_3 and PM_{10} pollutant patterns. Furthermore, pollution reduction for this particular region has to incorporate the entire city, since results show that local pollutant levels are not solely dependent on local emission sources. Adame et al. (2012) applied *k*-means clustering of daily patterns of O_3, NO_2 and SO_2, where the study region was one of heavy industrialized activity in Spain, and where four regimes could be defined. The four regime (or cluster) types describe the diurnal variation for each pollutant at hourly temporal resolution. Analysis of daily

variation of associated meteorological parameters such temperature and wind speed were used to characterize the conditions that are linked to the occurrence of these regimes. For daily variation of ozone, one of the clusters showed a peak in the early morning hours. The possible reasons for this could be the influence of industrial emissions, residual layer formation at night and intense mixing mechanisms during the day. Although, these reasons required verification possibly with the aid of mesoscale meteorological models. Hsu and Cheng (2016) demonstrated the use of the two-stage clustering approach for the classification of synoptic weather patterns with the aim of investigating the influence of meteorological characteristics on $PM_{2.5}$ in Yunlin, Taiwan. Major synoptic weather conditions comprised the different cluster types for e.g. continental anticyclone, monsoonal flow and Pacific subtropical high-pressure system. Wang et al. (2016) investigated $PM_{2.5}$ pollution characteristics for a 3-year period in Shanghai, China. Seasonal variations in $PM_{2.5}$ mass concentration were analyzed together with major regional transport patterns generated through the HYSPLIT model. Back trajectories over the Shanghai region were clustered and 4 air mass types were identified. Generally, air mass from the western region resulted in the highest pollution, whereas air mass from the East China sea had the lowest pollution due to maritime air. Moderate $PM_{2.5}$ pollution was associated with air mass from the north of China. Using hierarchical clustering with exposure and meteorological data, Cakmak et al. (2018) performed an assessment of the effects of long term $PM_{2.5}$ and ozone exposure, by establishing synoptic weather zones in Canada. Based on 5 parameters, 6 weather categories/zones were established (here and in the earlier study by Cakmak et al., 2016). Cluster types were spatially differentiated weather type zones. High $PM_{2.5}$ and ozone exposures were found to be correlated with different weather categories, providing useful insight on how weather patterns influence the impact of air pollutants on human health and mortality. In addition, occupational air pollution exposure was also an aspect investigated and was found to contribute to lung cancer. However, certain occupations were found to be at a higher risk than others, but this could not be fully explained by the present analysis. Similar to their earlier work, Hsu and Cheng (2019) performed a classification of synoptic weather patterns to assess the effects of meteorological conditions on PM_{10}, $PM_{2.5}$ and O_3 pollutants. Cluster types comprised wintertime weather types associated with high pollutant concentrations. Six weather patterns and the associated pollutant behavior were identified for the region.

5.1.4. Clustering techniques for other cluster types

Crecelius et al. (1980) performed one of the early studies on aerosols where *k*-means clustering was used to examine aerosol types and its associated formation process. Cluster types consisted of groups of elements that have a similar formation process. Classification of microphysical and optical properties of aerosols from 250 Aerosol Robotic Network stations using *k*-means clustering, was done by Omar et al. (2005). A set of 6 main clusters of aerosol types were categorized as desert dust, biomass burning, urban industrial pollution, rural background, polluted marine, and dirty pollution (i.e. pollution containing significant amounts of absorbing species). Examining the climatology of in situ measurements at individual sites was proposed as an appropriate validation of the clustering results. However, the availability of such data was currently restricted to few sites and furthermore, in situ measurements of the types of aerosols at AERONET stations was rather limited. Beddows et al. (2009) investigated the use of four i.e. fuzzy, *k*-means, *k*-median and model-based clustering techniques, for the characterization of particle size data from 4 monitoring stations in the UK. Upon application of an appropriate cluster validity index, *k*-means was found to produce the best clustering. Cluster types show distinct day and night-time trends in the particle size data, and revealed related information about their formation and dynamic processes that was essential for understanding its effect on human health. An extension of this analysis, by including data from additional particle sources around the city, would be required if source apportionment were to be examined. Wegner et al. (2012) analyzed aerosol number size distributions. Results showed 7 characteristic size distributions. Clustered size distributions were analyzed in terms of physical properties and the association to local meteorological conditions. Similar to Beddows et al. (2009), Hussein et al. (2014) applied *k*-means clustering to understand the fingerprints of urban aerosol particles in Helsinki, Finland. Seven fingerprints of aerosols were identified, some of which originating from local sources such as the transport sector and other anthropogenic activities. von Bismarck-Osten and Weber (2014) demonstrated the application of two *k*-means clustering approaches for analysing the particle number size distribution from 9 sites in middle-Europe. A specific signature was assigned to each cluster type based on the temporal and seasonal occurrence and the prevailing meteorological conditions.

5.2. Air mass trajectories

A trajectory, as described by Dorling and Davies (1995) is the “history” of an air parcel. Hafner et al. (2007) explains that trajectories are an approximation of the path that an air particle has taken before arriving at a particular site. Applying cluster analysis to trajectories will result in groups that have similar transport speeds and directions and are assumed to be representative of specific synoptic patterns. Within each cluster, individual trajectories may be averaged to produce a mean trajectory that represents the overall pattern of the cluster. A common source of trajectories include Hybrid Single-Particle Lagrangian Integrated Trajectory, (HYSPLIT) model of the Air Resources Laboratory (ARL) of the National Oceanic and Atmospheric Administration (NOAA). HYSPLIT is the most widely applied air trajectory model employed for establishing source-receptor relationships over long distances (Wang et al., 2010). Application of clustering to air mass trajectories for the analysis of atmospheric circulation patterns, relating air masses to synoptic/meteorological conditions and investigating pollutant behavior in terms of their source, transport pathways, source apportionment and air quality monitoring, were the common objectives in many of the studies reviewed here.

In recent years, air mass trajectory clustering has been widely used to identify homogeneous groups of transport patterns. In particular, the *k*-means techniques has been extensively applied (Brankov et al., 1988; Legras et al., 1988; Mo and Ghil, 1988; Dorling et al., 1992; Dorling and Davies, 1995; Mahura et al., 1999; Jorba et al., 2004; Abdalmogith and Harrison, 2005; Borge et al., 2007; Hafner et al., 2007; Pongkiatkul and Oanh, 2007; Toledano et al., 2009; Baker, 2010; Markou and Kassomenos, 2010; Rozwadowska et al., 2010; Wang et al., 2010; Wong et al., 2010; Cheng et al. 2011, 2013; Makra et al., 2011; Valenzuela et al., 2012; Bycenkiene et al., 2014; Yu et al., 2014; Donnelly et al., 2015; Lv et al., 2015; Luo and Chen, 2015; Terrouche et al., 2015; Fang et al., 2017; Ding et al., 2017; Wu et al., 2018; Skiles et al., 2018; Liu et al., 2019; Zhan et al., 2019). Other works (Moody and Galloway, 1988; Moody and Samson, 1989; Cheng and Wallace, 1993; Cape et al., 2000; Wang et al., 2004; Taubman et al., 2006; Tshela and Djolov, 2018) have employed Ward and average linkages, while fewer studies (Fernau and Samson, 1990a; b; Harris and Kahl, 1990; He et al., 2003; Kassomenos et al., 2010; Li et al., 2012) have relied on multiple methods and two-stage approaches. Table 3 presents a summary of the literature that have applied hierarchical and *k*-means techniques to air mass trajectories for air pollution studies. Each author is listed together with the clustered data/variable, method, and if more than one clustering method was applied, the method which was found to have the best performance was indicated.

5.2.1. Clustering techniques for linking air mass trajectories with synoptic meteorology

Using Ward's clustering, Moody and Galloway (1988) investigated the relationship between atmospheric transport and the precipitation composition for Bermuda. Trajectories for 850 mb and 700 mb levels were calculated using the Gridded Atmospheric Multilevel Backward

Table 3
Summary of the literature that have applied hierarchical, *k*-means and two-stage clustering techniques to air mass trajectories in air pollution studies.

Year	Author	Data/variables	Clustering approach/method	Best performing method if more than one used
1988	Legras et al.	500 hPa geopotential heights	*k*-means	
1988	Mo and Ghil	500 mb geopotential heights	*k*-means	
1988	Moody and Galloway	Trajectories at 850 mb, 700 mb	Hierarchical	
1989	Moody and Samson	Rawinsonde and precipitation chemistry data	Hierarchical	
1990a	Fernau and Samson	Trajectories at 300 m above ground level	Hierarchical	Ward linkage
1990b	Fernau and Samson	Trajectories at 300 m above ground level	Hierarchical	Ward linkage
1990	Harris and Kahl	Trajectories at 700 hPa and 500 hPa	Hierarchical, Two-stage clustering: Wards linkage and *k*-means	Two-stage
1992	Dorling et al.	Trajectories at 1000 hPa	*k*-means	
1993	Cheng and Wallace	500 hPa height fields	Hierarchical	
1995	Dorling and Davies	Trajectories at 1000 hPa	*k*-means	
1998	Brankov et al.	Trajectories at 200 m above ground level	*k*-means	
1999	Mahura et al.	Trajectories	*k*-means	
2000	Cape et al.	Trajectories for 900 hPa	Hierarchical	
2003	He et al.	Trajectories at 200 m, 500 m, 1000 m, 1500 m, 2000 m and 3000 m above ground level, PM2.5	Two-stage clustering: average linkage and *k*-means	
2004	Jorba et al.	Meteorological data, trajectories at 5500 m, 3000 m, 1500 m above mean sea level	*k*-means	
2004	Wang et al.	Trajectories at 1100 m above ground level	Hierarchical	
2005	Abdalmogith and Harrison	Trajectories at 900 hPa, PM_{10}	*k*-means	
2006	Taubman et al.	Trajectories at 1000 m, 2000 m, 3000 m above ground level	Hierarchical	
2007	Hafner et al.	Trajectories, $PM_{2.5}$	*k*-means	
2007	Borge et al.	Trajectories at 750 m, 1500 m and 3000 m above mean sea level	*k*-means	
2007	Pongkiatkul and Oanh	Trajectories at 1000 m above ground level, average mixing height, maximum mixing height, total cloudiness, dry and wet bulb temperature, surface pressure, daily rainfall, humidity, wind speed and direction	*k*-means	
2009	Toledano et al.	Trajectories at 500 m, 1500 m, 3000 m above mean sea level	*k*-means	
2010	Baker	Trajectories at 10 m, 400 m, 800 m above ground level	*k*-means	
2010	Kassomenos et al.	Trajectories at 10 m, 100 m and 500 m above ground level	Hierarchical, *k*-means and self-organizing maps	None indicated
2010	Markou and Kassomenos	Trajectories at 750 m, 1500 m, 3000 m above mean sea level	*k*-means	
2010	Rozwadowska et al.	Trajectories 1000 m 2500 m and 5000 m above mean sea level, AOT	*k*-means	
2010	Wang et al.	Trajectories at 300 m above ground level	*k*-means	
2010	Wong et al.	Trajectories at 500 m, 1000 m, 2000 m and 4000 m, aerosol data	Two-stage: Ward and *k*-means	
2011	Cheng et al.	Trajectories at 50 m above ground level, $PM_{2.5}$	*k*-means	
2011	Makra et al.	Trajectories at 500, 1500 and 3000 m amsl, $PM_{10,}$ and meteorological data	*k*-means	
2012	Li et al.	Meteorological data, $PM_{10,}$ trajectories at 200 m	Two-stage clustering: average linkage and *k*-means	
2012	Valenzuela et al.	AOD, AE, Trajectories at 1500 m, 3000 m above ground level	*k*-means	
2013	Cheng et al.	Trajectories at 30 m above mean sea level, PM_{10}	*k*-means	
2014	Bycenkiene et al.	Trajectories at 100 m above mean sea level, BC	*k*-means	
2014	Yu et al.	Trajectories at 100 m above ground level, $PM_{2.5}$, PM_{10}, O_3, NO_2, CO, SO_2	*k*-means	
2015	Donelly et al.	Trajectories, NO_2, PM_{10}	*k*-means	
2015	Lv et al.	Trajectories at 200 m above ground level, $PM_{2.5}$	*k*-means	
2015	Luo and Chen	Trajectories, $PM_{2.5}$	*k*-means	
2015	Terrouche et al.	Trajectories at 750 m	*k*-means	
2017	Fang et al.	Trajectories at 10 m above ground level, PM_{10}, $PM_{2.5}$	*k*-means	
2017	Ding et al.	Trajectories at 100 m above ground level, $PM_{2.5}$	*k*-means	
2018	Skiles et al.	Trajectories at 100 m, $PM_{2.5}$	*k*-means	
2018	Tshela and Djolov	Trajectories at 500 m above ground level	Hierarchical	
2018	Wu et al.	Trajectories at 100 m above ground level, $PM_{2.5}$	*k*-means	
2019	Liu et al.	Trajectories, AOD, wind speed and 850 hPa and 500 hPa geopotential heights	*k*-means	
2019	Zhan et al.	Trajectories at 1000 m, $PM_{2.5}$, PM_{10}, CO, NO_2, SO_2, $O_{3,}$ and meteorological data	*k*-means	

Isobaric Trajectory (GAMBIT) model. Dominant flow patterns by season were identified where there were 7 and 9 clusters found for the warm and cool seasons, respectively. Sulfate and nitrate concentrations were found to be higher in the warm season as compared to the cool. Although this study addressed the influence of atmospheric transport on precipitation composition, there are also other influencing factors such as wet and dry deposition. Ignoring these factors limits the amount of chemical variability that can be explained by only taking transport path under consideration. Similar to Moody and Galloway (1988), Moody and Samson (1989) analyzed back trajectories through Ward's hierarchical clustering to identify precipitations episodes that occurred under similar transport patterns in the mid-western United States. The results from the clustering provided insight into the precipitation characteristics for the region. For example, it was found that certain transport patterns were strongly correlated with higher ion concentrations. Further results also showed differences in transport patterns could be responsible for the variability in the composition of precipitation. There were also several other possible factors that could have resulted in differences in composition even under similar transport conditions, but these were not addressed. Hierarchical clustering by Ward's linkage applied to transport vector derived from back trajectories was investigated by Fernau and Samson (1990a). The aim was to

characterize air mass movements and identify periods of similar meteorology and precipitation chemistry for the eastern North American region. Cluster analysis of trajectories sampled at 10 sites produced 7 clusters, having distinct mean transport fields corresponding to high and low pressure systems observed on weather maps. In addition to Ward, average, centroid and median linkages were tested, but the results were shown to be unsatisfactory since one large cluster and smaller clusters referred to as "outlier clusters" were produced. This is in contrast with the clustering configuration produced by Ward's method which resulted in equally sized clusters. Fernau and Samson (1990b) extended their earlier work by investigating spatial precipitation and pollutant patterns for the same region. Ward's clustering was again used and identified weather patterns correlated with large amounts of pollution deposition. The analysis revealed some clusters that were associated with very dry conditions over the region, while some with very high precipitation levels. In addition, some clusters were associated with high levels of pollutant deposition. It was found that the highest pollutant depositions over the widest areas were as a result of mean transport patterns, with large areas of slow air mass movement over the regions of high sulfur emissions. Harris and Kahl (1990) carried out trajectory classification for Mauna Loa, Hawaii. Analysis of 8 years of 700 hPa and 500 hPa isobaric back trajectories were performed using Ward's clustering, after the average linkage and *k*-means methods. Ward's was combined with the *k*-means procedure, which revealed 6 clusters with distinctive features such easterly flow in the summer associated with trade winds, and strong westerly flow during winter. Average linkage clustering was eliminated as it produced outlier clusters containing few cluster members. Dorling et al. (1992) made use of *k*-means clustering for 1000 hPa isobaric back trajectories arriving at Eskdalemuir, Scotland, for investigating relationship between synoptic meteorology and pollutant concentrations. Eight clusters were found that describe distinct air flows with significantly different pollution source regions that were identified. Cheng and Wallace (1993) applied Ward's clustering to wintertime circulation patterns in the Northern Hemisphere. Several regimes were established through the clustering, three of which were found to be most reproducible in subsets of the data. The reproducibility of the larger clusters was compared by observing how well various ones were replicated when the procedure was repeated on randomly chosen halves of the dataset in an ensemble of 50 runs. However, despite this attempt at verifying reproducibility it was observed that membership of even the most reproducible clusters changed substantially. Dorling and Davies (1995) was an extension of the work of Dorling et al. (1992), whereby the same clustering technique was applied, and where five additional monitoring stations in northwest Europe have been included. McGregor and Bamzelis (1995) employed Ward's hierarchical clustering procedure to categorize a variety of meteorological variables with the objective of identifying days of similar weather conditions and different air mass types for Birmingham, UK. Using the multivariate synoptic index, 6 main air mass types were found and the associated meteorological characteristics were outlined. Air mass category of mixed maritime continental anticyclonic was found to be most frequently occurring. Pollutant characteristics, in particular of SO_2, NO_2, O_3, NO, CO, PM_{10}, associated with major airmass types were also presented. One of the findings was that anticyclonic activity is linked to a higher frequency of severe pollution events as compared to cyclonic activity. Due to the spatial variation of meteorological air pollution relationships, any synoptic index derived using station-specific data may not be applicable beyond the region for which it was developed. Application of average linkage clustering to back trajectories arriving at Mace Head, Ireland, was conducted by Cape et al. (2000). The 5 cluster types produced can be described as northerly, northwesterly, westerly, southwesterly and easterly. One of the main findings included the significant difference in ozone concentrations of the clusters. Even though the study was based on a very limited data set, different ozone concentrations were adequately captured through clustering. Jorba et al. (2004) clustered back trajectories arriving at 1500, 3000, and 5500 m above sea level for the Barcelona area, using 4 years of trajectory data. The main transport patterns were identified at 5500 m, which consisted of westerly, northwesterly, southwesterly flows, and regional re-circulations. Some of the interannual variability observed in the clustering patterns may be attributed to the North Atlantic Oscillation (NAO), although the reversal of the NAO (2000/01) was not clearly captured by the annual average transport patterns. Instead this feature was observed among winter patterns. Investigating summertime air pollution over the mid-Atlantic U.S. by Taubman et al. (2006) was done using the average linkage method. Eight clusters were identified that describe the trajectory densities, transport patterns and source regions of the clusters. Results show that areas of maximum trajectory density combined with wind speed are effective predictors of regional pollutant loadings. Similar to Jorba et al. (2004), Toledano et al. (2009) also followed the clustering approach of Dorling et al. (1992) for the classification of air masses reaching El Arenosillo, located on the southwestern coast of Spain. Aerosols arriving at the site were quantitatively characterized by Aerosol Optical Depth (AOD) and Angstrom Exponent (AE). Air mass trajectories at 1500 m could be grouped into 7 clusters. The study identified three main aerosols i.e. coastal marine, continental and desert dust, observed in the region. Baker (2010) conducted *k*-means clustering for the analysis long range air transport pathways and associated pollutant concentrations in Birmingham, UK. Six main trajectory clusters were identified as Arctic, strong-westerly, slow-easterly, westerly, south-westerly and slow-southerly. Results showed that highest pollutant concentrations were associated with the slow-easterly air mass cluster, while the lowest were associated with south-westerly and strong-westerly clusters. Kassomenos et al. (2010) performed a comparison of hierarchical, *k*-means and self-organizing maps clustering methods, for the classification of air mass trajectories arriving at Athens, Greece, for arrival heights of 10 m, 100 m and 500 m. All methods were found to be dependent on arrival height, but with varying degrees of dependence. Of all methods, *k*-means was found to be least dependent on trajectory arrival height. Sources of PM_{10} in Shanghai were investigated by Li et al. (2012), where two-stage clustering formed one of three methods for the study. This produced 7 clusters that were found to yield the best solution for air mass classification. Three clusters associated with the winter monsoon were found to be responsible for transporting a high concentration of PM_{10}. The remaining cluster corresponds to the monsoon transition period and showed an insignificant contribution to PM_{10} concentration levels in Shanghai. Investigation into desert dust events over Granada, Spain was carried out by Valenzuela et al. (2012). Clustering was applied to a set of 183 back-trajectories arriving over the region of Granada. Aerosol optical and microphysical properties were also included in the analysis. One of the findings included that, the transport of air masses from North Africa toward the southeastern region of the Iberian Peninsula have shown to follow three main paths. Six pollutants, as was used by Zhao et al. (2016) from 22 cities, meteorological data and air mass trajectories arriving at 1000 m were analyzed. Three polluted and five clean weather patterns were identified, which usually occurred in winter and summer, respectively. More recent studies that have employed *k*-means clustering of air mass trajectories in China for the analysis of pollutant transport were those by Fang et al. (2017), Ding et al. (2017) and Wu et al. (2018). Ding et al. (2017) investigated the spatial and temporal characteristics of $PM_{2.5}$ pollution in Hong Kong. Highest and lowest $PM_{2.5}$ levels were found to be in during the winter and summer seasons, respectively. In addition, six main source regions were identified to be responsible for $PM_{2.5}$ pollution. Fang et al. (2017) performed *k*-means clustering on trajectories arriving at Haikou, a coastal Chinese city, for the identification of PM_{10} and $PM_{2.5}$ sources during the winter and spring seasons. Results showed that pollutant concentrations were higher in winter than in spring. In addition, analysis of trajectories show pollutants were significantly affected by regional sources during the winter season as compared to the spring

season. Tshela and Djolov (2018) used Ward linkage to identify 5 major patterns in air trajectories arriving at six sites in Limpopo, South Africa. These corresponded to northerly to north-easterly, easterly to south-easterly, and south-westerly to north-westerly patterns. Wu et al. (2018) identified several Chinese cities (Beijing, Tianjin and Shandong, Henan and Hebei provinces) having the highest $PM_{2.5}$ pollution by clustering of trajectories. High $PM_{2.5}$ pollution levels during winter may be a result of the direct emissions and secondary formation of $PM_{2.5}$, and by the burning of biomass and fossil fuels for residential heating. Liu et al. (2019) analyzed the distribution, source and transport of aerosols over Central Asia during and 8-year period. In particular, dust and smoke events over the area were the main focus of the investigation. Clustering of trajectories revealed that during spring and summer, dust events were mainly transported from the northern Arabian Peninsular and northern African regions, while smoke events were mostly transported from Russia and Europe. During autumn and winter seasons majority of aerosol events were produced locally. Also for the Sichuan Basin, and similar to Zhao et al. (2016) and Zhao et al. (2018), Zhan et al. (2019) studied air pollution in the region by analysis of synoptic weather patterns and its association and impact to particle pollution.

5.2.2. Clustering techniques for linking air mass trajectories with pollutant patterns

He et al. (2003) used average linkage with *k*-means clustering to investigate the source-receptor relationship between pollutant concentrations and long-range air masses transport of $PM_{2.5}$ arriving at Taean, South Korea. Five trajectory groups were identified which contained 98% of the data, thus resulting in the other three groups being substantially less significant. Hafner et al. (2007) performed a clustering analysis on rainfall and fine aerosol ($PM_{2.5}$) data for three western U.S. sites, using the *k*-means algorithm. Using a 7-year period of daily trajectories, the aim was to assign rainfall amounts and pollutant concentrations to the arrival time for each trajectory, in order to understand the synoptic patterns and routes of atmospheric contamination, respectively. Results showed that trajectories can be grouped into 6 main patterns for each site. For each site trajectory clusters for 1, 5 and 10 days were computed to denote short, medium and long-range flow patterns, respectively. It was found that highest $PM_{2.5}$ concentrations were associated with slow moving clusters and lowest $PM_{2.5}$ concentrations with fast moving clusters. Overall, those clusters generated using single-day trajectories performed best at differentiating the clusters by rainfall and aerosol concentration, and hence are a better predictor for rainfall and $PM_{2.5}$ concentrations. This was followed by 5-day trajectories as the second best. Clusters not being able to recognize short-term, high concentration $PM_{2.5}$ events, for e.g. a forest fire or distant plumes, were highlighted as a drawback of the method. Borge et al. (2007) analyzed the long-range transport influences on urban PM_{10} pollutants by clustering of trajectories with 750, 1500 and 3000 m arrival heights, for Athens, Madrid and Birmingham. Results suggest that long-range transport from the North African and continental European regions have significant impact on the PM_{10} levels in Madrid and Birmingham, while a moderate effect for Athens. Instead, local emission sources play a larger role for PM_{10} levels in Athens. In addition, two indices based on PM_{10} exceedances were proposed for the assessment of PM_{10} in each cluster, and can be used to associate future exceedances with specific circulation patterns. Pongkiatkul and Oanh (2007) also assessed long-range transport of pollutants, in the Bangkok Metropolitan region, where clustering of HYSPLIT trajectories produced 6 main groups that were associated with PM_{10} and $PM_{2.5}$ levels. To assess the contribution of the long-range transport on pollutants and draw definitive conclusions, additional monitoring sites and more PM composition data are required. Characterization of PM_{10} concentration in Leece, Italy was studied by Contini et al. (2010), where Ward linkage comprised one of three methods used in the analysis. Markou and Kassomenos (2010) demonstrated the use of clustering for the classification of back trajectories arriving at Athens in Greece. Five years of trajectories arriving at three heights, 750, 1500 and 3000 m were clustered separately, producing 12 trajectory classes for each level. The clustering results could be used in further investigations such as identifying air masses over the area and analysing the influence of atmospheric transport patterns on pollutant concentrations. Analysis of air mass trajectories to investigate the transport of air pollutants, particularly PM_{10}, over Beijing was conducted by Wang et al. (2010). By applying *k*-means clustering to trajectories, the study aimed find a relationship between atmospheric transport patterns and PM_{10} levels affecting air quality. Similar to Wang et al. (2010), Cheng et al. (2011) have applied the *k*-means technique for the analysis of various pollutant sources and respective transport pathways in China, which was achieved by combining air pollutant measurements with air mass trajectory clustering. Cheng et al. (2013) performed a similar investigation to that of Wang et al. (2010) and Cheng et al. (2011), for the analysis of pollutant sources and transport pathways in China. Bycenkiene et al. (2014) applied *k*-means clustering to trajectories reaching Preila, Lithuania, with the aim of evaluating transport patterns of black carbon in the south-eastern Baltic region. Six clusters were identified and the carbon levels associated with each of them were analyzed. Days with high concentrations of black carbon were associated with air masses that originated and passed over the southern European regions prior to its arrival in Preila in winter. Other studies in China similar to Wang et al. (2010), Cheng et al. (2011) and Cheng et al. (2013) include those by Yu et al. (2014), Lv et al. (2015) and Luo and Chen (2015). Similar to Jin et al. (2011), Skiles et al. (2018) presented a source apportionment analysis using *k*-means clustering of air mass trajectories in the San Joaquin Valley however, in this case, for $PM_{2.5}$ organic carbon pollutants.

5.2.3. Clustering techniques for other cluster types

Wong et al. (2010) studied transport pathways in Hong Kong to investigate potential sources of different aerosols. Ward combined with *k*-means clustering was applied to aerosol properties such as aerosol optical thickness (AOT), single scattering albedo (SSA) and Angstrom exponent (AE), to produce a classification of 4 aerosol types i.e. mixed urban, polluted urban, dust and heavy pollution. Mixed urban aerosols were found to be most prevalent. The 4 types were associated with trajectories (500, 1000, 2000 and 4000 m) to identify the pollutant sources and pathways arriving in Hong Kong. Terrouche et al. (2015) used clustering of trajectories for the identification of potential distant sources that contribute to particulate pollution and metallic elements in Constantine, Algeria. Trajectories arriving at 750 m were clustered by the *k*-means algorithm, showing that the Sahara desert was a major source contributing to PM_{10} and Fe. Long-range transport of air masses originating from the north and south regions, in particular the Mediterranean sea, were the major contributors to Na, Mg, K and Ca. Elements such as Zn, Cu and Pb are of anthropogenic origin i.e. traffic and industrial activities.

6. Summary

The aim of this paper was to provide a review of cluster analysis applications to air pollution studies, in particular using ground-based pollution measurements and air mass trajectories depicting pollutant transport pathways. This review focused on the use of hierarchical (Ward, single, average, centroid and complete linkages) and partitional (*k*-means) clustering techniques. Based on the discussion presented in Section 5, the following can be concluded:

- Application of hierarchical and *k*-means clustering methods have been applied in air pollution studies for more than three decades, with one of the earliest research works appearing in 1980. Since then, a large number of research works have contributed to this topic. Air pollution studies have shown to be especially important for those locations experiencing high levels of pollutant concentrations, particularly from human activities such as industrial operations, vehicle emissions and biomass burning. Many of the studies reviewed in this paper have shown that clustering of air pollution measurements can enable efficient pollution monitoring,

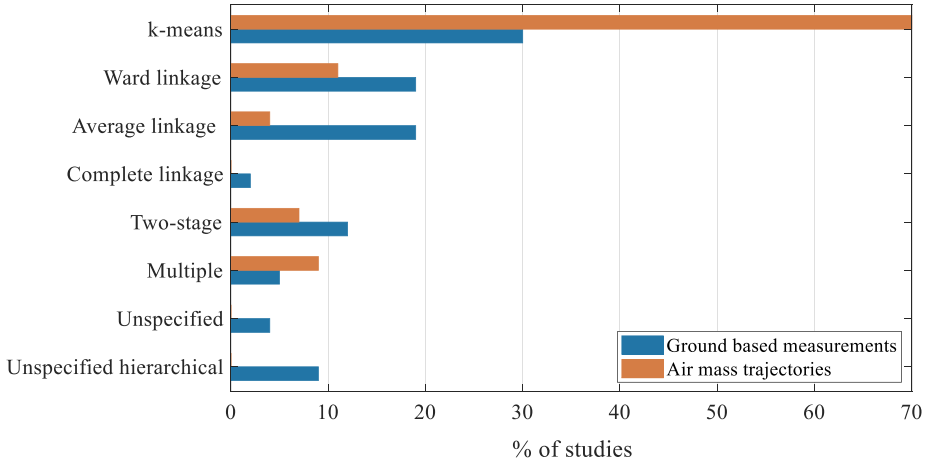

Fig. 4. The percentage of studies used by the *k*-means and hierarchical methods under each category. For both categories, *k*-means had the largest proportion. For category 1, Ward and average linkages are the second most commonly used methods. For category 2, Ward linkage is mostly used after *k*-means.

identification of sources and the development and implementation of effective control and mitigation strategies. Research involving these aspects are currently, and will remain, important due to the continuous increase of pollution emissions from these on-going activities.

- Many of the studies have focused on analyzing spatio-temporal pollutant characteristics, air mass origins and pathways, transport patterns associated with various air pollutants arriving at different geographical locations and establishing links between weather types and air pollutants, all of which are important for environmental purposes. An example of one such purpose is air quality monitoring, which currently is of major significance due to the adverse effects it poses on human health.
- As shown in Fig. 4, for the clustering of ground-based air pollution measurements (category 1), the *k*-means method has been used in 30% of studies in this review, and thus found to be most common. This is followed by Ward and average linkage clustering, for which each were 19%. Two-stage clustering and the use of multiple methods individually account for 13% and 5% of the studies, respectively. Some studies (4%) did not specify the clustering method, while other studies (9%) stated the use of a hierarchical method, but did not specify the type of linkage. For the clustering of air mass trajectories (category 2), the *k*-means method has been employed in 70% of the studies reviewed, followed by Ward's clustering which accounts for about 11%. Multiple methods were used in 9%, while two-stage clustering was used in 7% of studies. Average linkage was the least used method that accounted for 4%.
- Overall from this review, three techniques namely *k*-means, Ward and average linkage clustering have shown to be the most used for both categories, accounting for about 77% of all studies. In addition, of all the studies making use of hierarchical clustering, only the agglomerative method has been employed, with little attention to divisive methods. This may be due to researchers being guided by the literature where the agglomerative method is widely applied. Individually, *k*-means Ward's, and average methods constitute 48%, 16%, and 13%, of all studies respectively. Ward and average linkages have been popular choices, while other linkages were not.
- Fig. 5 shows the proportion of studies in terms of geographical location. More specifically, the location where the study was conducted. China and USA were found to have the highest percentage of studies (each with 24%) employing hierarchical and *k*-means clustering for air pollution studies. This was followed by the studies that focused on multiple locations (13%). The UK and Spain were similar with 7% and 6%, respectively. All other countries individually accounted for less than 5% of all studies.
- Given that the two-stage approach is supposed to assist in

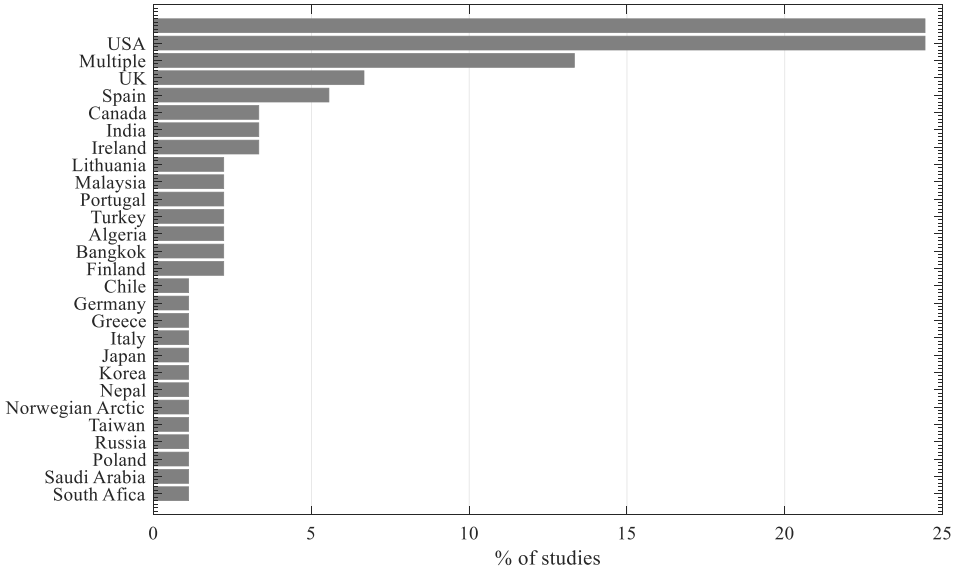

Fig. 5. Frequency of studies by country (where the study was conducted) based on the current review.

developing the initial cluster groupings for the selection of the appropriate '*k*' in the *k*-means procedure, it would be expected that many more studies employ this method, since selection of the number of clusters is one of the most well-known problems in cluster analysis, as pointed out by Jain (2010). However, it was noticed from the literature presented here that a majority of researchers continue to gravitate towards the use of single-approach techniques such as *k*-means, average and Ward clustering.

- There is no consensus on a single criteria for validating the number of clusters. Instead cluster validation criteria varied among the authors. Therefore, it would be best if the user performs a comparison using two or more validation criteria, in order to decide on the most appropriate for the data at hand.
- Many of the studies reviewed here have applied clustering techniques to long-term data records, in some cases in excess of 10 years. Therefore, this highlights and speaks to the importance of establishing and maintaining long-term high-quality measurement stations for recording various types of air pollution data, for the identification and monitoring of patterns, trends and anomalies.
- From the review the main software that has been used to perform the clustering was found to be SPSS. Other software included MATLAB, SAS, XLSTAT (an add-in for Excel) and R. For clustering of trajectories, the TrajStat software was often used.

Overall, each of the two clustering methods has its own advantages and disadvantages. Furthermore, the literature on air pollution studies presented here does not offer a clear consensus on the best method. Therefore, it is left to the researcher to explore and test different methods with the aim of finding the one most appropriate for the data. Tables 2 and 3 can be used to facilitate a comparison between the data or variables used and the different clustering methods, as well as to assist the researcher in selecting the most appropriate method under each of the two categories.

7. Recommendations

The following suggestions/recommendations are proposed based on the current review:

- Average and Ward linkages have been the most commonly used among studies, while other linkages have been less common. There should be studies undertaken to analyse and compare different linkage methods and their performance when applied to air pollution data and air mass trajectories.
- Most studies have analyzed the behavior of pollutants from industrial activities and urban aerosols. It would be interesting to investigate the pollution levels and transport pathways at locations where there are several large-scale power plants.
- Currently, China and USA have conducted majority of air pollution research, while for other countries this has been significantly fewer, especially for regions such as Africa and India where biomass burning is prevalent. Therefore more studies in locations experiencing high levels of pollution need to be undertaken.
- Given that the use of cluster analysis has been successful in the study of air pollutant behavior, in particular for the analysis of spatial and temporal characteristics, the results should be incorporated into research directed toward the development of air pollutant forecasting models.

8. Conclusions

The aim of this paper was to provide a review of the clustering applications, in particular by the use of hierarchical agglomerative and *k*-means methods, that have been applied in air pollution studies conducted over the past 40 years. More specifically, studies that focused on spatio-temporal characteristics of air pollutants, pollutant behavior in terms of their source, transport pathways, apportionment and links to meteorological conditions, were presented in this review. Reviews of clustering techniques applied to air pollution studies are currently lacking and this study aims to fill that gap. The use of two clustering methods were reviewed, particularly for studies involving ground based measurements of pollutants and air mass trajectories describing pollutant sources and pathways. For all studies involving ground based measurements of pollutants, the *k*-means method was employed in 28% of them, followed by the average (20%) and Ward linkages (20%). Two-stage methods were used in 13% of studies, while all other methods individually comprise less than 10%. For studies involving air mass trajectories describing pollutant sources and pathways, the *k*-means method comprised 70% of them, followed by Ward linkage (11%), multiple methods (9%) and two-stage clustering (6%). From the various hierarchical linkage methods, average and Ward were most commonly applied. In terms of geographical location, most studies (48%) reviewed have been conducted in China and USA. Each clustering technique has its own advantages and disadvantages and there is no one "best" method. Therefore, it is left to the researcher to explore the different clustering methods to find which best suits the data or the application at hand. Overall, this review intends to provide researchers with a guide to choosing the most appropriate cluster analysis method for application to some of the data commonly encountered in air pollution studies. In addition, and to the best of the authors' knowledge this is the first review that covers the longest time span (1980–2019) of clustering applications in air pollution studies.

Declarations of interest

None.

Acknowledgements

P. Govender would like to acknowledge the National Astrophysics and Space Science Programme (NASSP) at the University of KwaZulu-Natal, South Africa for financial support.

References

Abdalmogith, S.S., Harrison, R.M., 2005. The use of trajectory cluster analysis to examine the long–range transport of secondary inorganic aerosol in the UK. Atmos. Environ. 39, 6686–6695.
Adame, J.A., Notario, A., Villanueva, F., Albaladejo, J., 2012. Application of cluster analysis to surface ozone, NO_2 and SO_2 daily patterns in an industrial area in Central-Southern Spain measured with a DOAS system. Sci. Total Environ. 429, 281–291.
Aghabozorgi, S., Shirkhorshidi, A.S., Wah, T.Y., Soltanian, H., Herawan, T., 2014. Spatial and temporal clustering of air pollution in Malaysia: a review. In: International Conference on Agriculture, Environment and Biological Sciences (ICFAE 14), Antalya, Turkey.
Anderberg, M.R., 1973. Cluster Analysis for Applications. Academic Press, New York.
Austin, E., Coull, B., Thomas, D., Koutrakis, P., 2012. A framework for identifying distinct multi pollutant profiles in air pollution data. Environ. Int. 45, 112–121.
Baker, J., 2010. A cluster analysis of long range air transport pathways and associated pollutant concentrations within the UK. Atmos. Environ. 44, 563–571.
Ball, G.H., Hall, D.J., 1965. A clustering technique for summarizing multivariate data. Behav. Sci. 12, 153–155.
Beaver, S., Palazoglu, A., 2006. Cluster analysis of hourly wind measurements to reveal synoptic regimes affecting air quality. J. Appl. Meteorol. Climatol. 45, 1710–1726.
Beddows, D.C.S., Dall'osto, M., Harrison, R.M., 2009. Cluster analysis of rural, urban and curbside atmospheric particle size data. Environ. Sci. Technol. 43, 4694–4700.
Bergstra, A.D., Brunekreef, B., Burdorf, A., 2018. The effect of industry-related air pollution on lung function and respiratory symptoms in school children. Environ. Health 17 (30), 1–9.
Bhagat, A., Kshirsagar, N., Khodke, P., Dongre, K., Ali, S., 2016. Penalty parameter selection for hierarchical data stream clustering. Procedia Comput. Sci. 79, 24–31.
Borge, R., Lumbreras, J., Vardoulakis, S., Kassomenos, P., Rodríguez, E., 2007. Analysis of long-range transport influences on urban PM_{10} using two-stage atmospheric trajectory clusters. Atmos. Environ. 41 (21), 4434–4450.
Brankov, E., Rao, S.T., Porter, P.S., 1988. A trajectory-clustering-correlation methodology for examining the long-range transport of air pollutants. Atmos. Environ. 32 (9), 1525–1534.
Bycenkiene, S., Dudoitis, V., Ulevicius, V., 2014. The use of trajectory cluster analysis to evaluate the long-range transport of black carbon aerosol in the south-eastern Baltic region. Adv. Meteorol. Article ID 137694, 1–11.

Cakmak, S., Hebbern, C., Vanos, J., Crouse, D.L., Burnett, R., 2016. Ozone exposure and cardiovascular-related mortality in the Canadian Census Health and Environment Cohort (CANCHEC) by spatial synoptic classification zone. Environ. Pollut. 214, 598–599.
Cakmak, S., Hebbern, C., Pinault, L., Lavigne, E., Vanos, J., Crouse, D.L., Tjepkema, M., 2018. Associations between long-term $PM_{2.5}$ and ozone exposure and mortality in the Canadian Census Health and Environment Cohort (CANCHEC), by spatial synoptic classification zone. Environ. Int. 111, 200–211.
Cape, J.N., Methven, J., Hudson, L.E., 2000. The use of trajectory cluster analysis to interpret trace gas measurements at Mace Head, Ireland. Atmos. Environ. 34, 3651–3663.
Cheng, S., Ye, H., Kalkstein, L.S., 1992. An evaluation of pollution concentrations in Philadelphia using an automated synoptic approach. Middle States Geographer 25, 45–51.
Cheng, X., Wallace, J.M., 1993. Cluster analysis of the northern hemisphere wintertime 500-hPa height field: spatial patterns. J. Atmos. Sci. 50 (16), 2674–2696.
Cheng, S., Yang, L., Zhou, X., Wang, Z., Zhou, Y., Gao, X., Nie, W., Wang, X., Xua, P., Wang, W., 2011. J. Environ. Monit. 13, 1662–1671.
Cheng, S., Wang, F., Li, J., Chen, D., Li, M., Zhou1, Y., Ren, Z., 2013. Application of trajectory clustering and source apportionment methods for investigating trans-boundary atmospheric PM_{10} pollution. Aerosol Air Qual. Res. 13, 333–342.
Chu, H.J., Liau, C.J., Lin, C.H., Su, B.S., 2012. Integration of fuzzy cluster analysis and kernel density estimation for tracking typhoon trajectories in the Taiwan region. Expert Syst. Appl. 39, 9451–9457.
Comrie, A.C., 1996. An all-season synoptic climatology of air pollution in the U.S.-Mexico border region. Prof. Geogr. 48, 237–251.
Contini, D., Genga, A., Cesari, D., Siciliano, M., Donateo, A., Bove, M.C., Guascito, M.R., 2010. Characterisation and source apportionment of PM_{10} in an urban background site in Lecce. Atmos. Res. 95, 40–54.
Crecelius, E.A., Lepel, E.A., Laul, J.C., Rancitelli, L.A., McKeever, R.L., 1980. Background air particulate chemistry near Colstrip, Montana. Environ. Sci. Technol. 14 (4), 422–428.
Davis, R.E., Kalkstein, L.S., 1990. Using a spatial synoptic climatological classification to assess changes in atmospheric pollution concentrations. Phys. Geogr. 11 (4), 320–342.
Davis, R.E., Gay, D.A., 1993. A synoptic climatological analysis of air quality in the Grand Canyon National Park. Atmos. Environ. 27A, 713–727.
Davis, R.E., 1991. A synoptic climatological analysis of winter visibility trends in the mideastern United States. Atmos. Environ. B Urban Atmos. 25 (2), 165–175.
Davis, J.M., Eder, B.K., Nychka, D., Yang, Q., 1998. Modeling the effects of meteorology on ozone in Houston using cluster analysis and generalized additive models. Atmos. Environ. 32 (14), 2505–2520.
Davis, R.E., Walker, D.R., 1992. An upper-air synoptic climatology of the western United States. J. Clim. 5 (12), 1449–1467.
Davuliene, L., Sakalys, J., Dudoitisa, V., Reklaite, A., Ulevicius, V., 2019. Long-term black carbon variation in the south-eastern baltic region in 2008-2015. Atmos. Pollut. Res. 10, 123–133.
Ding, H., Liu, Y., Yu, Z., Cheung, C., Zhan, J., 2017. Spatial and temporal characteristics and main contributing regions of high $PM_{2.5}$ pollution in Hong Kong. Aerosol Air Qual. Res. 17, 2955–2965.
Dominick, D., Juahir, H., Talib, L.M., Zain, S.M., Aris, A.Z., 2012. Spatial assessment of air quality patterns in Malaysia using multivariate analysis. Atmos. Environ. 60, 172–181.
Donnelly, A.A., Broderick, B.M., Misstear, B.D., 2015. The effect of long-range air mass transport pathways on PM_{10} and NO_2 concentrations at urban and rural background sites in Ireland: quantification using clustering techniques. J. Environ. Sci. Health, Part A 50 (7), 647–658.
Dorling, S.R., Davies, T.D., Pierce, C.E., 1992. Cluster analysis: a technique for estimating the synoptic meteorological controls on air and precipitation chemistry – method and applications. Atmos. Environ. 26A (14), 2575–2581.
Dorling, S.R., Davies, T.D., 1995. Extending cluster analysis-synoptic meteorology links to characterize chemical climates at six northwest European monitoring stations. Atmos. Environ. 29 (2), 145–167.
Dubes, R., Jain, A.K., 1976. Clustering techniques: the user's dilemma. Pattern Recognit. 8, 247–260.
Eder, B.K., Davis, J.M., Bloomfield, P., 1994. An automated classification scheme designed to better elucidate the dependence of ozone on meteorology. J. Appl. Meteorol. 33, 1182–1199.
Everitt, B.S., Landau, S., Leese, M., Stahl, D., 2011. Cluster Analysis. Wiley, London.
Fang, X., Bi, X., Xu, H., Wu, J., Zhang, Y., Feng, Y., 2017. Source apportionment of ambient PM_{10} and $PM_{2.5}$ in Haikou, China. Atmos. Res. 190, 1–9.
Fernau, M.E., Samson, P.J., 1990a. Use of cluster analysis to define periods of similar meteorology and precipitation chemistry in eastern North America. Part I: transport patterns. J. Appl. Meteorol. 29, 735–750.
Fernau, M.E., Samson, P.J., 1990b. Use of cluster analysis to define periods of similar meteorology and precipitation chemistry in eastern North America. Part II: Precipitation patterns and pollutant deposition. J. Appl. Meteorol. 29, 751–761.
Flemming, J., Stern, R., Yamartino, R.J., 2005. A new air quality regime classification scheme for O_3, NO_2, SO_2 and PM_{10} observations sites. Atmos. Environ. 39, 6121–6129.
Fullerton, D.G., Brucec, N., Gordon, S.B., 2008. Indoor air pollution from biomass fuel smoke is a major health concern in the developing world. Trans. Royal Soc. Trop. Med. Hyg. 102, 843–885.
Gao, H., Chen, J., Wang, B., Tan, S.C., Lee, C.M., Yao, X., Yan, H., Shi, J., 2011. A study of air pollution of city clusters. Atmos. Environ. 45, 3069–3077.
Giri, D., Murthy, V.K., Adhikary, P.R., Khanal, S.N., 2007. Cluster analysis applied to atmospheric PM_{10} concentration data for determination of sources and spatial patterns in ambient air-quality of Kathmandu Valley. Curr. Sci. 93 (5), 684–688.
Gómez-Losada, A., Lozano-García, A., Pino-Mejías, R., Contreras-González, J., 2014. Finite mixture models to characterize and refine air quality monitoring networks. Sci. Total Environ. 485–486, 292–299.
Gómez-Losada, A., Pires, J.C.M., Pino-Mejías, R., 2018. Modelling background air pollution exposure in urban environments: implications for epidemiological research. Environ. Model. Softw 106, 13–21.
Gong, X., Richman, M.B., 1995. On the application of cluster analysis to growing season precipitation data in North America East of the Rockies. J. Clim. 8, 897–931.
Gorham, E., Martin, F.B., Litzau, J.T., 1984. Acid rain: ionic correlations in the eastern United States. Science 225, 407–409.
Gramsch, E., Cereceda-Balic, F., Oyola, P., von Baer, D., 2006. Examination of pollution trends in Santiago de Chile with cluster analysis of PM_{10} and ozone data. Atmos. Environ. 40, 5464–5475.
Greene, J.S., Kalkstein, L.S., Ye, H., Smoyer, K., 1999. Relationships between synoptic climatology and atmospheric pollution at 4 US cities. Theor. Appl. Climatol. 62, 163–174.
Grivas, G., Chaloulakou, A., Kassomenos, P., 2008. An overview of the PM_{10} pollution problem, in the metropolitan area of Athens, Greece. Assessment of controlling factors and potential impact of long range transport. Sci. Total Environ. 389, 165–177.
Hafner, W.D., Solorzano, N.N., Jaffe, D.A., 2007. Analysis of rainfall and fine aerosol data using clustered trajectory analysis for National Park sites in the Western US. Atmos. Environ. 41, 3071–3081.
Halkidi, M., Batistakis, Y., Vazirgiannis, M., 2001. On clustering validation techniques. J. Intell. Inf. Syst. 17, 107–145.
Harinath, S., Murthy, U.N., 2012. Effect of air pollution on human health in industrial areas- a case study. J. Ind. Pollut. Control 28 (1), 9–11.
Harris, J.M., Kahl, J.D., 1990. A descriptive atmospheric transport climatology for the Mauna Loa Observatory, using clustered trajectories. J. Geophys. Res. 95 (D9) 13,651-13,667A.
Han, J., Kamber, M., Pei, J., 2012. Data Mining Concepts and Techniques. Elsevier, USA.
Hartigan, J.A., Wong, M.A., 1979. Algorithm AS 136: a *k*-means clustering algorithm. Appl. Stat. 28, 100–108.
He, Z., Kim, Y.J., Ogunjobi, K.O., Hong, C.S., 2003. Characteristics of $PM_{2.5}$ species and long-range transport of air masses at Taean background station, South Korea. Atmos. Environ. 37, 219–230.
Hsu, C.H., Cheng, F.Y., 2016. Classification of weather patterns to study the influence of Meteorological characteristics on $PM_{2.5}$ concentrations in Yunlin County, Taiwan. Atmos. Environ. 144, 397–408.
Hsu, C.H., Cheng, F.Y., 2019. Synoptic weather patterns and associated air pollution in Taiwan. Aerosol Air Qual. Res. 19, 1139–1151.
Huang, P., Zhang, J., Tang, Y., Liu, L., 2015. Spatial and temporal distribution of $PM_{2.5}$ pollution in Xi'an City, China. Int. J. Environ. Res. Public Health 12, 6608–6625.
Hussein, T., Mølgaard, B., Hannuniemi, H., Martikainen, J., Järvi, L., Wegner, T., Ripamonti, G., Weber, S., Timo Vesala, T., Hämeri, K., 2014. Fingerprints of the urban particle number size distribution in Helsinki, Finland: local versus regional characteristics. Boreal Environ. Res. 19, 1–20.
Iizuka, A., Shirato, S., Mizukoshi, A., Noguchi, M., Yamasaki, A., Yanagisawa, Y., 2014. A cluster analysis of constant ambient air monitoring data from the Kanto region of Japan. Int. J. Environ. Res. Public Health 11, 6844–6855.
Jain, A.K., Murty, M.N., Flynn, P.J., 1999. Data clustering: a review. ACM Comput. Surv. 31 (3), 264–323.
Jain, A.K., 2010. Data clustering: 50 years beyond *k*-means. Pattern Recognit. Lett. 31, 651–666.
Jin, L., Harley, R.A., Brown, N.J., 2011. Ozone pollution regimes modeled for a summer season in California's San Joaquin Valley: a cluster analysis. Atmos. Environ. 45, 4707–4718.
Jolliffe, I.T., Philipp, A., 2010. Some recent developments in cluster analysis. Phys. Chem. Earth 35, 309–315.
Jorba, O., Perez, C., Rocadenbosch, F., Baldasano, J.M., 2004. Cluster analysis of 4-day back trajectories arriving in the Barcelona area, Spain, from 1997 to 2002. J. Appl. Meteorol. 43, 887–901.
Kahya, C., Balcik, F.B., Oztaner, Y.B., Ozcomak, D., Seker, D.Z., 2017. Spatio-temporal analysis of $PM_{2.5}$ over marmara region. Turkey. Fresen. Environ. Bull. 26 (1), 310–317.
Kalkstein, L.S., Corrigan, P., 1986. A synoptic climatological approach for geographical analysis: assessment of sulphur dioxide concentrations. Ann. Assoc. Am. Geogr. 76 (3), 381–395.
Kalkstein, L.S., Tan, G., Skindlov, J.A., 1987. An evaluation of three clustering procedures for use in synoptic climatological classification. J. Clim. Appl. Meteorol. 26, 717–730.
Kampa, M., Castanas, E., 2008. Human health effects of air pollution. Environ. Pollut. 151, 362–367.
Kaufman, L., Rousseeuw, P.J., 1990. Finding Groups in Data: an Introduction to Cluster Analysis. Wiley, New Jersey.
Kassomenos, P., Vardoulakis, S., Borge, R., Lumbreras, J., Papaloukas, C., Karakitsios, S., 2010. Comparison of statistical clustering techniques for the classification of modelled atmospheric trajectories. Theor. Appl. Climatol. 102, 1–12.
Kim, S.B., Temiyasathit, C., Chen, V.C.P., Park, S.K., Sattler, M., Russell, A.G., 2008. Characterization of spatially homogeneous regions based on temporal patterns of fine particulate matter in the continental United States. J. Air Waste Manag. Assoc. 58 (7), 965–975.
Kwon, Y., Lee, H., Lee, H., 2018. Implication of the cluster analysis using greenhouse gasemissions of Asian countries to climate change mitigation. Mitig. Adapt. Strategies Glob. Change 23, 1225–1249.
Latif, M.T., Dominick, D., Ahamad, F., Khan, M.F., Juneng, L., Hamzah, F.M., Nadzir,

M.S.M., 2014. Long term assessment of air quality from a background station on the Malaysian Peninsula. Sci. Total Environ. 482–483, 336–348.
Laumbach, R.J., Kipen, H.M., 2012. Respiratory health effects of air pollution: update on biomass smoke and traffic pollution. Allergy Clin. Immunol. 129 (1), 3–13.
Legras, B., Desponts, T., Piguet, B., 1988. Cluster analysis and weather regimes. Proc. of the Workshop on the Nature and Prediction of Extratropical Weather Systems 2. ECMWF, Reading, Shinfield Park, UK, pp. 123–149.
Li, M., Huang, X., Zhu, L., Li, J., Song, Y., Cai, X., Xie, S., 2012. Analysis of the transport pathways and potential sources of PM_{10} in Shanghai based on three methods. Sci. Total Environ. 414, 525–534.
Liao, T.W., 2005. Clustering of time series data-a survey. Pattern Recognit. 38, 1857–1874.
Liu, Y., Zhu, Q., Wang, R., Xiao, K., Cha, P., 2019. Distribution, source and transport of the aerosols over central Asia. Atmos. Environ. 210, 120–131.
Lu, H.C., Chang, C.L., Hsieh, J.C., 2006. Classification of PM_{10} distributions in taiwan. Atmos. Environ. 40, 1452–1463.
Lu, W.Z., He, H.D., Dong, L.Y., 2011. Performance assessment of air quality monitoring networks using principal component analysis and cluster analysis. Build. Environ. 46, 577–583.
Luo, M., Chen, C., 2015. Potential sources and transport pathways of $PM_{2.5}$ in Shanghai, China. In: Proceedings of the 2nd IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM), Fuzhou, China.
Lv, B., Liu, Y., Yu, P., Zhang, B., Bai, Y., 2015. Characterizations of PM2.5 pollution pathways and sources analysis in four large cities in China. Aerosol Air Qual. Res. 15, 1836–1843.
Lyapina, O., Schultz, M.G., Hense, A., 2016. Cluster analysis of European surface ozone observations for evaluation of MACC reanalysis data. Atmos. Chem. Phys. 16, 6863–6881.
MacQueen, J.B., 1967. Some Methods for classification and analysis of multivariate observations. Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1. University of California Press, Berkeley, pp. 281–297.
Madulatha, T.S., 2012. An overview on clustering methods. IOSR J. Eng. 2 (4), 719–725.
Mahura, A.G., Jaffe, D.A., Andres, R.J., Merrill, J.T., 1999. Atmospheric transport pathways from the Bilibino nuclear power plant to Alaska. Atmos. Environ. 33, 5115–5122.
Makra, L., Matyasovszky, I., Guba, Z., Karatzas, K., Anttila, P., 2011. Monitoring the long-range transport effects on urban PM_{10} levels using 3D clusters of backward trajectories. Atmos. Environ. 45, 2630–2641.
Markou, M.T., Kassomenos, P., 2010. Cluster analysis of five years of back trajectories arriving in Athens, Greece. Atmos. Res. 98, 438–457.
McGregor, G.R., Bamzelis, D., 1995. Synoptic typing and its application to the investigation of weather air pollution relationships, Birmingham, United Kingdom. Theor. Appl. Climatol. 51, 223–236.
Milligan, G.W., Cooper, M.C., 1985. An examination of procedures for determining the number of clusters in a data set. Psychometrika 50 (2), 159–179.
Mo, K., Ghil, M., 1988. Cluster analysis of multiple planetary flow regimes. J. Geophys. Res. 93, 10927–10952.
Moody, J.L., Galloway, J.N., 1988. Quantifying the relationship between atmospheric transport and the chemical composition of precipitation on Bermuda. Tellus B Chem. Phys. Meteorol. 40 (5), 463–479.
Moody, J.L., Samson, P.J., 1989. The influence of atmospheric transport on precipitation chemistry at two sites in the mid-western United States. Atmos. Environ. 23, 2117–2132.
Munir, S., Habeebullah, T.M., Mohammed, A.M.F., Morsy, E.A., 2015. An analysis into the temporal variations of ground level ozone in the arid climate of Makkah applying *k*-means algorithms. Environ. Asia 8 (1), 53–60.
Namratha, M., Prajwala, T.R., 2012. A comprehensive overview of clustering algorithms in pattern recognition. IOSR J. Comput. Eng. 4 (6), 23–30.
Omar, A.H., Won, J.G., Winker, D.M., Yoon, S.C., Dubovik, O., McCormick, M.P., 2005. Development of global aerosol models using cluster analysis of Aerosol Robotic Network (AERONET) measurements. J. Geophys. Res. 110 (D10S14), 1–14.
Omran, M.G.H., Engelbrecht, A.P., Salman, A., 2007. An overview of clustering methods. Intell. Data Anal. 11, 583–605.
Pandey, B., Agrawal, M., Singh, S., 2014. Assessment of air pollution around coal mining area: emphasizing on spatial distributions, seasonal variations and heavy metals, using cluster and principal component analysis. Atmos. Pollut. Res. 5, 79–86.
Pires, J.C.M., Sousa, S.I.V., Pereira, M.C., Alvim-Ferraz, M.C.M., Martins, F.G., 2008a. Management of air quality monitoring using principal component and cluster analysis-part I: SO_2 and PM_{10}. Atmos. Environ. Times 42, 1249–1260.
Pires, J.C.M., Sousa, S.I.V., Pereira, M.C., Alvim-Ferraz, M.C.M., Martins, F.G., 2008b. Management of air quality monitoring using principal component and cluster analysis-part II: CO, NO_2 and O_3. Atmos. Environ. 42, 1261–1274.
Pongkiatkul, P., Oanh, N.T.K., 2007. Assessment of potential long-range transport of particulate air pollution using trajectory modeling and monitoring data. Atmos. Res. 85, 3–17.
Qiao, X., Ying, Q., Li, X., Zhang, H., Hu, J., Tang, Y., Chen, X., 2018. Source apportionment of PM2.5 for 25 Chinese provincial capitals and municipalities using a source-oriented community multiscale air quality model. Sci. Total Environ. 612, 462–471.
Rozwadowska, A., Zielinski, T., Petelski, T., Sobolewski, P., 2010. Cluster analysis of the impact of air back-trajectories on aerosol optical properties at Hornsund, Spitsbergen. Atmos. Chem. Phys. 10, 877–893.
Saksena, S., Joshib, V., Patilc, R.S., 2003. Cluster analysis of Delhi's ambient air quality data. J. Environ. Monit. 5, 491–499.
Sanchez, M.L., Ramos, M.C., Pascual, D., Perez, I., 1990. Application of cluster analysis to identify sources of airborne particles. Atmos. Environ. 21, 1521–1527.
Sausy, D., Anderson, J.R., Buseck, P.R., 1987. Cluster analysis samples from the Norwegian Arctic. Atmos. Environ. 21, 1649–1657.
Shi, P., Xie, P.H., Qin, M., Si, F.Q., Dou, K., Du, K., 2014. Cluster analysis for daily patterns of SO_2 and NO_2 measured by the DOAS System in Xiamen. Aerosol Air Qual. Res. 14 (5), 1455–1465.
Skiles, M.J., Lai, A.M., Olson, M.R., Schauer, J.J., Foy, B.D., 2018. Source apportionment of PM2.5 organic carbon in the San Joaquin Valley using monthly and daily observations and meteorological clustering. Environ. Pollut. 237, 366–376.
Soares, J., Makar, P.A., Aklilu, Y., Akingunola, A., 2018. The use of hierarchical clustering for the design of optimized monitoring networks. Atmos. Chem. Phys. 18, 6543–6566.
Solazzo, E., Galamarini, S., 2015. Comparing apples with apples: using spatially distributed time series of monitoring data for model evaluation. Atmos. Environ. 112, 234–245.
Steinhaus, H., 1956. Sur la division des corp materiels en parties. Bull. Acad. Polon. Sci. IV (C1.III), 801–804.
Taubman, B.F., Hains, J.C., Thompson, A.M., Marufu, L.T., Doddridge, B.G., Stehr, J.W., Piety, C.A., Dickerson, R.R., 2006. Aircraft vertical profiles of trace gas and aerosol pollution over the mid-Atlantic United States: statistics and meteorological cluster analysis. J. Geophys. Res. 111, 1–14 D10S07.
Terrouche, A., Ali-Khodja, H., Kemmouche, A., Bouziane, M., Derradji, A., Charron, A., 2015. Identification of sources of atmospheric particulate matter and trace metals in Constantine, Algeria. Air Qual. Atmos. Health. 9 (1), 69–82.
Toledano, C., Cachorro, V.E., Frutos, M.E., Torres, B., Berjon, A., Sorribas, M., Stone, R.S., 2009. Airmass classification and analysis of aerosol types at El Arenosillo (Spain). J. Appl. Meteorol. Climatol. 48, 962–981.
Tshela, C., Djolov, G., 2018. Source profiling, source apportionment and cluster transport analysis to identify the sources of PM and the origin of air masses to an industrialised rural area in Limpopo. Clean Air J. 28 (2), 54–66.
Tuffery, S., 2011. Data Mining and Statistics for Decision Making. Wiley, Sussex.
Unal, Y.S., Toros, H., Deniz, A., Incecik, S., 2011. Influence of meteorological factors and emission sources on spatial and temporal variations of PM_{10} concentrations in Istanbul metropolitan area. Atmos. Environ. 45, 5504–5513.
Valenzuela, A., Olmo, F.J., Lyamani, H., Antón, M., Quirantes, A., Alados-Arboledas, L., 2012. Classification of aerosol radiative properties during African desert dust intrusions over southeastern Spain by sector origins and cluster analysis. J. Geophys. Res. 117, 1–18.
von Bismarck-Osten, C., Weber, S., 2014. A uniform classification of aerosol signature size distributions based on regression-guided and observational cluster analysis. Atmos. Environ. 89, 346–357.
Wang, Y.Q., Zhang, X.Y., Arimoto, R., Cao, J.J., Shen, Z.X., 2004. The transport pathways and sources of PM10 pollution in Beijing during spring 2001, 2002 and 2003. Geophys. Res. Lett. 31 (L14110), 1–4.
Wang, F., Chen, D.S., Cheng, S.Y., Li, J.B., Li, M.J., Ren, Z.H., 2010. Identification of regional atmospheric PM_{10} transport pathways using HYSPLIT, MM5-CMAQ and synoptic pressure pattern analysis. Environ. Model. Softw 25 (8), 927–934.
Wang, H.L., Qiao, L.P., Lou, S.R., Zhou, M., Ding, A.J., Huang, H.Y., Chen, J.M., Wang, Q., Tao, S.K., Chen, C.H., Li, L., Huang, C., 2016. Chemical composition of $PM_{2.5}$ and meteorological impact among three years in urban Shanghai, China. J. Clean. Prod. 112, 1302–1311.
Ward, J.H., 1963. Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. 58, 236–244.
Wegner, T., Hussein, T., Hämeri, K., Vesala, T., Kulmala, M., Weber, S., 2012. Properties of aerosol signature size distributions in the urban environment as derived by cluster analysis. Atmos. Environ. 61, 350–360.
Wong, M.A., Nichol, J.E., Lee, K.H., 2010. Remote Sensing of the atmosphere and clouds III. Proc. SPIE 7859, 78590E.
Wu, X., Ding, Y., Zhou, S., Tan, Y., 2018. Temporal characteristic and source analysis of $PM_{2.5}$ in the most polluted city agglomeration of China. Atmos. Pollut. Res. 9 (6), 1221–1230.
Yu, S., Zhang, Q., Yan, R., Wang, S., Li, P., Chen, B., Liu, W., Zhang, X., 2014. Origin of air pollution during a weekly heavy haze episode in Hangzhou, China. Environ. Chem. Lett. 12, 543–550.
Zhan, C.C., Xie, M., Fang, D.X., Wang, T.J., Wu, Z., Lu, H., Li, M.M., Chen, P.L., Zhuang, B.L., Li, S., Zhang, Z.Q., Gao, D., Reng, J.Y., Zhao, M., 2019. Synoptic weather patterns and their impacts on regional particle pollution in the city cluster of the Sichuan Basin, China, Atmos. Environ. Times 208, 34–47.
Zhang, J., Huang, X., Chen, Y., Luo, B., Luo, J., Zhang, W., Rao, Z., Yang, F., 2019. Characterization of lead-containing atmospheric particles in a typical basin city of China: seasonal variations, potential source areas, and responses to fireworks. Sci. Total Environ. 661, 354–363.
Zhang, J.J., Smith, K.R., 2007. Household air pollution from coal and biomass fuels in China: measurements, health impacts, and interventions. Environ. Health Perspect. 115 (6), 848–855.
Zhao, S., Yu, Y., Yin, D., He, J., Liu, N., Qu, J., Xiao, J., 2016. Annual and diurnal variations of gaseous and particulate pollutants in 31 provincial capital cities based on in situ air quality monitoring data from China National Environmental Monitoring Center. Environ. Int. 86, 92–106.
Zhao, S., Yu, Y., Qin, D., Yin, D., Dong, L., He, J., 2018. Analyses of regional pollution and transportation of $PM_{2.5}$ and ozone in the city clusters of Sichuan Basin, China. Atmos. Pollut. Res. 10 (2), 374–385.