

# A new metric space incorporating radon-222 for generation of back trajectory clusters in atmospheric pollution studies

Jagoda Crawford\*, Wlodek Zahorowski, David D. Cohen

Australian Nuclear Science and Technology Organisation, PMB 1 Menai, NSW 2234, Australia

## ARTICLE INFO

### Article history:

Received 22 May 2008

Received in revised form

3 September 2008

Accepted 4 September 2008

### Keywords:

Aerosol

Back trajectory cluster analysis

Radon

## ABSTRACT

A novel metric space for the clustering of back trajectories to be used in fine particle aerosol data analysis was proposed and evaluated. The metric is based on spatial and non-spatial variables incorporating great-circle distance, altitude and radon-222.

Its performance was examined using the intra-cluster variation of measured and fingerprint apportioned aerosol mass as the quantitative criterion. The new metric was demonstrated to perform better than those based on great-circle distance, or a great-circle distance and altitude alone. The same criterion was applied to investigate the clustering performance as a function of the length of its back trajectories. The optimum back trajectory length was found to be dependent on the pollution source being considered. Performance tests, as well as the application of the new metric space to re-analysis of previously published results, were based on a three year long dataset comprising co-located aerosol fine particles (PM<sub>2.5</sub>) collection and hourly measurements of radon-222 concentration.

The new metric space can easily be redefined to include other trace species.

Crown Copyright © 2008 Published by Elsevier Ltd. All rights reserved.

## 1. Introduction

Atmospheric back trajectories are widely used to quantify the influence of air transport on the pollution at a site (Stohl, 1998). There are two major ways to visualise air quality data (Owega et al., 2006). The first is a probability map which identifies areas around the receptor site that contribute to the pollution observed at the site (the Potential Source Contribution Function, PSCF; Hopke et al., 1995; Crawford et al., 2007). The second is cluster analysis where the data is split into a number of groups representing distinct fetch areas and atmospheric transport patterns. The application of clusters to back trajectory analysis was introduced by Moody (1986), who recognised its ability to simultaneously account for variations in both wind speed and direction (see also Harris and Kahl, 1990).

We propose a new, non-Euclidean metric space for generation of back trajectory clusters, which incorporates radon-222 (radon) concentrations measured at the receptor site. Radon is a naturally occurring radioactive tracer widely used for the identification of recent terrestrial influence on chemical composition of an air mass. It is the best naturally occurring gaseous tracer of entrainment of terrestrial aerosols on regional scales. The radon source function does not change significantly on such scales (Conen and Robertson, 2002), its half life is of the same order of magnitude as the life time

of aerosols, and it is chemically inert. Radon concentration in air can be measured with high accuracy and precision for a large range of concentrations.

Metric spaces used previously for cluster definition are based on spatial variability in two or three dimensions. Typically, they include either sums of the squared distance (on the Euclidean plane) between end points (Moody and Galloway, 1988) or sums of the great-circle distance (on a Euclidean sphere) between trajectory end points (e.g. Lin et al., 2001 and Jorba et al., 2004). Other metrics in this category include Manhattan and Pearson Correlation distance functions (Kaufman and Rousseeuw, 2005). Hafner et al. (2007) and Cape et al. (2000) reported that including the altitude of the trajectory seems to have little impact on cluster formation. However, Hafner et al. (2007), following Harris et al. (2005), point out that mean altitudes correlate with trajectory length, as mean wind speeds tend to be greater at higher altitudes. Consequently, trajectories arriving at a receptor site from a similar direction could form separate clusters for higher and lower altitudes if altitude is considered.

It should be noted that a metric space and hence various degrees of similarity between trajectories can, but does not rely directly on spatial variables only and its definition depends on the problem at hand. To our knowledge, this is the first attempt to include a non-spatial measure in the definition of a metric space for cluster formation. While radon is used in this study, the metric space can include any species of interest, for example CO (Wang et al., 2003) could present a short range tracer option. The choice of tracer

\* Corresponding author. Tel.: +61 2 9717 2885; fax: +61 2 9717 9260.

E-mail address: [Jagoda.Crawford@ansto.gov.au](mailto:Jagoda.Crawford@ansto.gov.au) (J. Crawford).

depends on the timescale of interest and the assumed mix of the pollutant under study.

We compare performance of the proposed metric with other metrics. The criterion we adopt is based on an assumption that the better the cluster identifies uniform fetch area and transport patterns, the smaller the resulting intra-cluster variation of measured aerosol mass attributable to the clusters. This is a criterion which directly relates cluster performance to the experimental quantity (measured aerosol mass) which can be determined with adequate precision. We apply the criterion to clusters of different back trajectory lengths, with the range limited by the residence time of aerosols in the atmosphere (Seinfeld and Padnis, 1998) and the accuracy of back trajectories, which decreases with distance from the receptor (Stohl, 1998). Further, we apply the criterion using aerosol mass of individual fingerprints. Finally, we briefly discuss the results of the application of the new metric for 12 clusters and 8 source fingerprints.

## 2. Method

### 2.1. Study site and sampling

The Hok Tsui site is located on Cape d'Aguilar at the south-eastern end of Hong Kong Island at 22.22°N, 114.25°E atop a 60 m cliff facing the South China Sea. The population density of Cape d'Aguilar is relatively low and the nearest major urban/industrial town of Chai Wan is 10 km away, thus the site is generally considered a rural background site (Louie et al., 2005b).

The experimental data underpinning the above investigation comes from a 3-year sampling program of PM<sub>2.5</sub> aerosols and hourly radon observations at Hok Tsui from 2001 till 2003. The discussed set of source fingerprints was previously derived from the same data (Crawford et al., 2007) using factor analysis proposed by Paatero and Tapper (1994). The meteorology of the study site and the aerosol and radon sampling has been previously described in Crawford et al. (2007).

### 2.2. Trajectory calculation

Air mass trajectories give an approximation of the path of polluted air parcels over a period of time. Hence, a trajectory model can be used to identify the potential source regions for pollutants measured at the receptor site (e.g. Brankov et al., 1997). Various methods to compute trajectories, based on different assumptions, have been developed (Stohl, 1998). The accuracy of an individual trajectory is limited by the temporal and spatial resolutions of meteorological observations, measurement and analysis errors, and any simplifying assumptions used in the trajectory model (Stohl, 1998; Brankov et al., 1997). An assumption that is often made is that the errors are random, and instead of relying on the analysis of any individual trajectory, a large number of trajectories are used (Harris and Kahl, 1990; Brankov et al., 1997; Hopke et al., 1995; Man and Shih, 2001), which results in the cancelling-out of errors from individual trajectories.

The PC version of HYSPLIT v4.0 (HYbrid Single-Particle Lagrangian Integrated Trajectory; Draxler and Rolph, 2003) was used to generate a database of 10-day back trajectories for every hour of years 2001–2003, which was subsequently analysed based on aerosol sampling days. A starting height of 300 m was chosen to reduce topographic effects on the simulations, and to be within the atmospheric boundary layer year-round. In comparing HYSPLIT calculated trajectories with tracer gas releases, Draxler (1991) estimated that the model error is between 20% and 30% of the travel distance.

Small-scale features, such as the impact of local terrain, cannot be resolved by the data assimilation system that produces the

global-gridded wind fields from which trajectories are calculated. The trajectories therefore reflect the large-scale atmospheric transport characteristics of the air (Harris and Kahl, 1990) and in this study the meteorological data used is of 1° by 1° resolution.

There is still some disagreement over the most suitable length of the back trajectories used for aerosol analysis. For instance, Hafner et al. (2007) studied 1, 5 and 10-day back trajectories (to analyse precipitation) and PM<sub>2.5</sub> data from three National Park sites in the Western US, and found that for these sites, 1-day clusters are a better predictor of precipitation and PM<sub>2.5</sub> concentrations, followed by 5-day clusters. Man and Shih (2001) used 10-day back trajectories. Taking into consideration that the residence time of aerosols in the atmosphere can be anywhere between 3 and 20 days (Seinfeld and Padnis, 1998) and that the accuracy of back trajectories reduces with time, we considered 1, 3, 5 and 7 days back trajectories.

### 2.3. Definition of metric space and cluster analysis

A frequently used clustering method for trajectories is that of Dorling and Davies (1992), which is referred to as the “*k*-means” method. The analysis starts with a set of trajectories, which are then grouped into *k* clusters using a specific distance measure. This measure, called metric (or distance function), defines a metric space in which similarities or differences between trajectories can be quantified. In the *k*-means approach, a cluster centre is defined as the mean trajectory for the cluster, in reference to which the distance to other trajectories is calculated in a specific metric space. Trajectories are then arranged to minimise the variance of the distance between trajectories in the same cluster, and maximise the variance for trajectories belonging to different clusters. A variant of the *k*-means method is the *k*-modes (or medoids) method, where a specific trajectory is selected to represent the “centroid” as opposed to an average trajectory in the *k*-means method.

We have investigated four metric spaces for cluster formation. The first two metrics (Metrics 1 and 2 in equations (1) and (2)) are defined by horizontal spatial variability of trajectories. They facilitate comparison of our results with those previously published and, in the context of this paper, provide a direct comparison with Metrics 3 and 4 (equations 3 and 4). Metric 3 includes an explicit trajectory altitude term. Metric 4, in a departure from previously published metrics, contains, apart from horizontal and vertical spatial terms, a non-spatial term derived from radon concentrations measured in air parcels arriving at the receptor site. Inclusion of the non-spatial term is expected to improve the overall performance of the resulting clusters applied to aerosol analysis. This is because (a) a significant trajectory density over an area is necessary but not a sufficient condition for entrainment of terrestrial pollution from that area, and (b) the radon term provides a measure directly related to entrainment velocity of terrestrial aerosols into the analysed air parcels. Also, radon vertical distributions in the lower atmosphere are directly related to turbulence and mixing processes which are not very well resolved by back trajectories. It should be noted that a strong correlation between radon and the measured aerosol mass is not a necessary condition for improved performance. In addition, such a strong correlation is not a norm. To expect otherwise would be equivalent to assuming that aerosol mass flux is constant in the study domain.

We use great-circle distance in all four metrics as it offers a better approximation of horizontal spatial distances for longer trajectories where Earth's curvature might become important.

In Metric 1, the distance between two trajectories is defined as a sum of great-circle distances between the corresponding end points of trajectories (equation (1)). Metrics 2, 3, and 4 have their distance terms normalised (e.g. Kaufman and Rousseeuw, 2005) to reduce the impact that regions further from the receptor site have

on the metrics. Metrics 3 and 4 require the altitude and radon terms to be normalised as their horizontal and vertical variability differs by more than one order of magnitude and the radon is a non-spatial variable. The normalised terms are expressed as a fraction of their respective ranges.

Equations (1)–(4) define the four metrics:

$$D_{i,j} = \frac{1}{n} \sum_{k=1}^n d_{i,j}(k) \quad (\text{Metric 1}) \quad (1)$$

$$D_{i,j} = \frac{1}{n} \sum_{k=1}^n \left( \frac{d_{i,j}(k)}{R_{k,x}} \right) \quad (\text{Metric 2}) \quad (2)$$

$$D_{i,j} = \frac{1}{n} \sum_{k=1}^n \left( \frac{d_{i,j}(k)}{R_{k,x}} + \frac{|h_i(k) - h_j(k)|}{R_{k,h}} \right) \quad (\text{Metric 3}) \quad (3)$$

$$D_{i,j} = \frac{1}{n} \sum_{k=1}^n \left( \frac{d_{i,j}(k)}{R_{k,x}} + \frac{|h_i(k) - h_j(k)|}{R_{k,h}} + \frac{|r_i - r_j|}{R_r} \right) \quad (\text{Metric 4}) \quad (4)$$

$$R_{k,x} = \max(d_{i,j}(k)) \quad \text{for all } i, j \text{ in the dataset}$$

$$R_{k,h} = \max(|h_i(k) - h_j(k)|)$$

$$R_r = \max(r_i - r_j)$$

where  $d_{i,j}(k)$  is the great-circle distance (m) between the two end points of trajectory  $i$  and  $j$  at time  $k$  (calculated using the haversian distance equation, Sinnott, 1984),  $h_i(k)$  and  $h_j(k)$  is the height (m) of trajectory  $i$  and  $j$ , respectively, at the  $k$ th end point,  $r_i$  and  $r_j$  are the radon concentrations ( $\text{mBq/m}^3$ ) corresponding to trajectory  $i$  and  $j$ , respectively, and  $n$  is the number of hours considered for the back trajectory.

Normalisation factors  $R$  represent the largest distance or radon concentration change amongst all pairs of trajectories in the dataset. The metrics in equations (1)–(4) fulfil the usual conditions of positive definiteness, symmetry, sub-additivity and translation invariability.

For cluster formation we use the PAM (Partitioning Around Medoids; Kaufman and Rousseeuw, 2005) program. Selecting an optimum number of clusters is done in the PAM program by calculating and comparing a quantity called “the average silhouette width” (see Section 3.1 below) Our choice of the particular implementation of the clustering method was driven by convenience alone, as the definition of metrics is part of the input of that implementation rather than being rendered in the source code. Similar studies can be carried out using other clustering methods, including the  $k$ -means method.

#### 2.4. Elemental source fingerprint identification

Crawford et al. (2007) used Positive Matrix Factorisation (PMF; Paatero and Tapper, 1994) to analyse the aerosol data considered in this paper and apportion them to possible pollutant sources. Eight possible sources were identified, based on the elemental composition of the source fingerprints, as *Sea1* (sea spray), *Sea2* (old sea air), *Soil* (dust), *Auto* (vehicle exhaust), *2ndryS* (secondary sulfate formed from the reaction of  $\text{SO}_2$  in the atmosphere), *Smoke* (biomass burning), *Oil/Diesel* (burning of oil), *Org/Coal* (coal combustion). For

convenience, the monthly average, of the three year composite data, of each source fingerprint is reproduced in Fig. 1.

The percentage contribution to the total mass from each factor being *Sea 1* 1%, *Sea 2* 10%, *Soil* 8%, *Oil/Diesel* 5%, *Organic/Coal* 25%, *2ndryS* 26%, *Smoke* 14% and *Auto* 9%.

### 3. Results and discussions

#### 3.1. Cluster generation

We defined four sets of clusters in the four metric spaces (Equations (1)–(4)) using 1, 3, 5 and 7 days long back trajectories. The clustering program, PAM, was run for each of the 16 cases, generating between 3 and 15 clusters for each case.

We applied two criteria to the identification of an optimum range of clusters for the set. The first criterion is based on the idea of average silhouette width (Kaufman and Rousseeuw, 2005). The quantity is defined as a dataset average of  $s_i$  numbers calculated for each trajectory  $i$  where  $s_i = 1 - a_i/b_i$  and  $a_i$  is the average distance of trajectory  $i$  to all other trajectories in the cluster to which trajectory  $i$  belongs, and  $b_i$  is the minimum of average distances between trajectory  $i$  and trajectories belonging to other clusters. The definition implies that a good classification corresponds to the number of clusters for which the average silhouette width, estimated for the entire dataset, is as high as possible (Kaufman and Rousseeuw, 2005).

Fig. 2a shows the average silhouette width for the dataset, calculated for a range of clusters of 7 days trajectories. The common feature of the four curves, each representing one of the four metrics, is a steep decrease of the average silhouette width for classifications using between 3 and 8 clusters. By comparison, the average silhouette width varies significantly less for each metric with higher cluster numbers, pointing to little change in the classification quality for these numbers of clusters. It is obvious that we cannot simply use the classification with the highest average silhouette width. Nor do the curves indicate any particular number of clusters for that purpose. It is more likely that classifications corresponding to higher cluster numbers where the average silhouette width varies as little as possible indicate an optimum range of clusters.

The second criterion for identifying an optimum range of clusters relies on comparing the variation of the measured aerosol mass for different classifications. The total aerosol mass is experimentally determined for each measurement and hence its variation is probably the most reliable indicator of cluster classification. It is influenced, if only implicitly, by the most important characteristics of the system including seasonal variation in air parcel paths, source function and its temporal and spatial variation for different aerosol types, and their resident time in the atmosphere. Thus it overcomes the weakness of the first criterion which, at least for the first three metrics, is nothing more than a measure of average relative position of trajectories within and between their respective parent clusters.

Fig. 2b shows the variation of the measured aerosol mass, calculated for the entire dataset classified by a range of cluster sets between 3 and 15, calculated for four metrics and 7 days trajectories. It is clear that the criterion is more conclusive than silhouette width. Classifications based on lower number of clusters are characterised by higher variability of total aerosol mass. The quality of the classification is optimal for cluster numbers between 8 and 12 and it does not deteriorate within this range.

Based on the two criteria a tentative range of optimal number of clusters is between 8 and 12. Visual inspection of sets of clusters within this range demonstrated that 12 clusters offered the highest degree of detail on separation of air transport patterns. Hence, 12 clusters were chosen for further analysis. It is

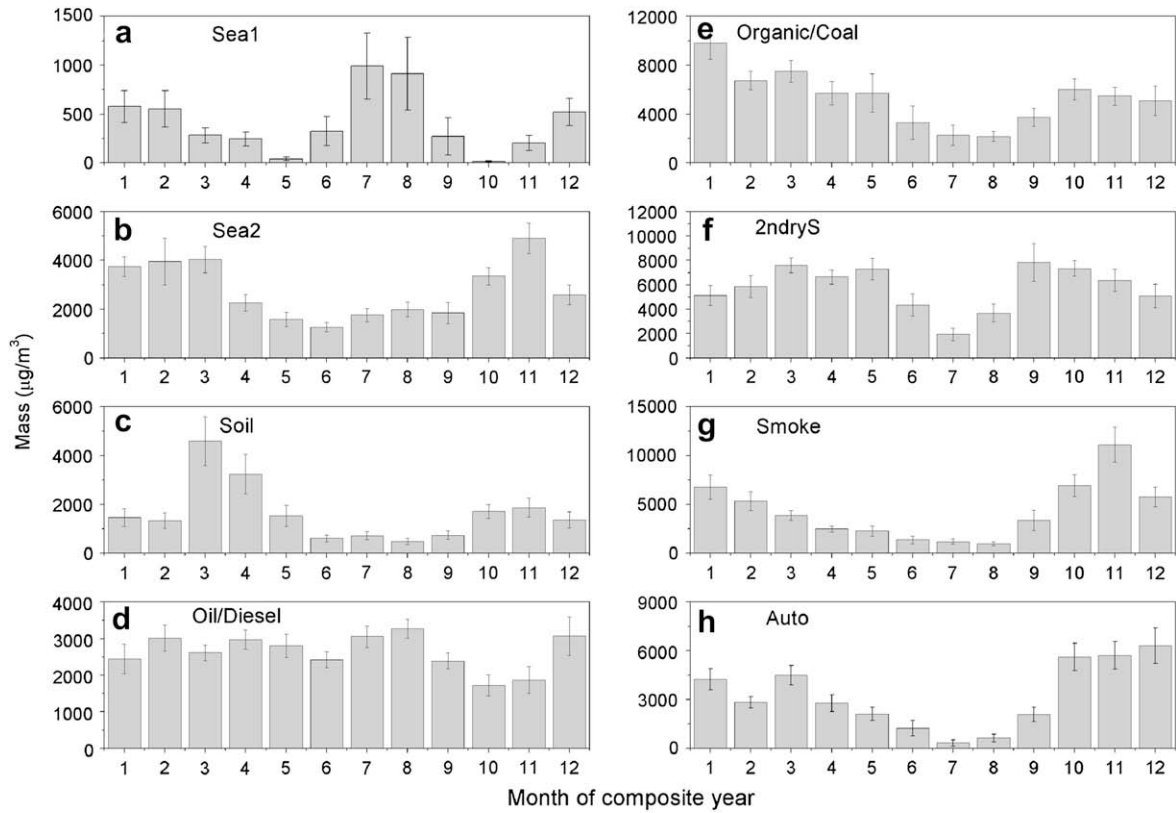


Fig. 1. Monthly composite mass concentration for each source fingerprint with standard error, averaged over the 3 years of data (from Crawford et al., 2007).

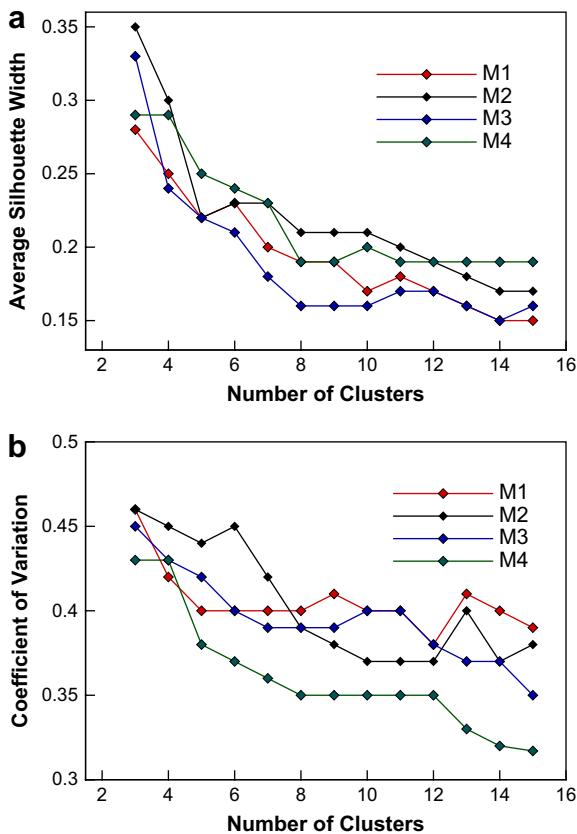


Fig. 2. (a) Average silhouette and (b) average variation of measured aerosol mass; for each of the four metric spaces when 7 days back trajectories are used.

important to stress, however, that there is no unique solution to determine the optimum number of clusters for a given dataset. This is the case even for relatively simple datasets where considerations based on geometrical properties of clusters might be adequate. Hence, the final choice is necessarily, if only partially, a qualitative one.

### 3.2. Impact of metrics on cluster performance

We use variation of the measured aerosol mass to quantify the impact of four metrics on cluster performance. The measure is suitable for evaluation of metric/cluster performance for the same reasons as it is for the determination of classification quality. By contrast, the mass of each source type is estimated using the PMF

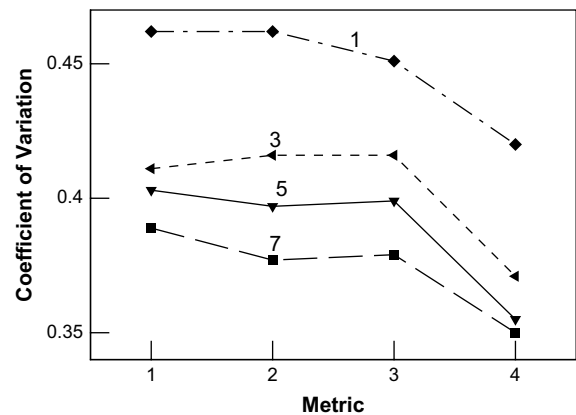


Fig. 3. Average coefficient of variation of measured aerosol mass for clusters defined using four metrics and 1, 3, 5 and 7 day back trajectories (numbered).

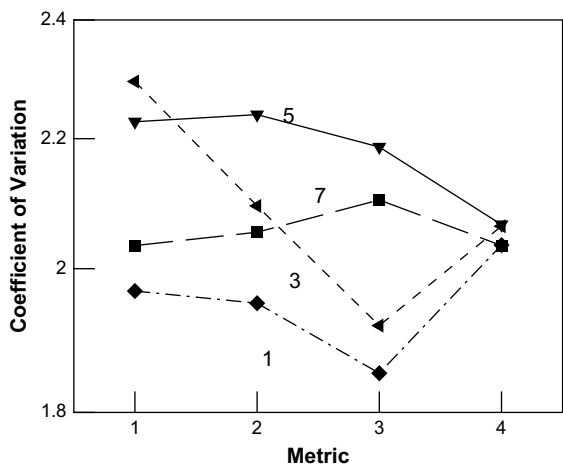


Fig. 4. Coefficient of variation of *Sea 1* aerosol mass calculated in the four metric spaces and 1, 3, 5 and 7 days trajectory lengths (numbered).

analysis and thus its accuracy depends on the accuracy of the elemental analysis of all samples, and, probably more importantly, on the accuracy of the PMF analysis itself. These qualifications notwithstanding, we include an evaluation of cluster performance based on variation of the mass of each source as defined by the PMF analysis. We characterise the variation of mass within a cluster by the coefficient of variation defined by the standard deviation as a fraction of the mean.

Each aerosol sample is collected for 24 h. Hence, hourly trajectories calculated for each collection day were allocated 1/24th of the corresponding aerosol mass. Such an allocation is based on the assumption that, on a diurnal time scale, trajectories within a cluster point to an area of approximately uniform aerosol

emissions. To strengthen the validity of the assumption, we calculated the mean and standard deviations of mass for clusters only for those days in which 13 or more trajectories were classified in the same cluster. The average coefficient of variation of measured aerosol mass for all sets of the 12 clusters, defined by using the four metric spaces and four trajectory lengths in each metric space, is shown in Fig. 3. It is clear that application of Metric 4 (comprising normalised spatial and non-spatial terms) results in a significant decrease in the coefficient of variation. This is also evident from Fig. 2b discussed in Section 3.1 above. Increasing the length of back trajectories also results in less intra-cluster variation. However, normalising the distance between trajectories at each end point and adding trajectory altitude as in Metric 2 and Metric 3, respectively, has little impact on the variation of the measured mass within a cluster. It is possible that using horizontal distance (Metrics 1 and 2), which accounts for horizontal separation of the trajectories and, implicitly, for speed (Moody, 1986), does also account for some of the altitude differences, as higher altitude trajectories tend to move faster than lower altitude trajectories.

Variation of the source fingerprint mass, defined as the average coefficient of variation for the 12 clusters, determined for each source fingerprint using the four metric spaces and the four trajectory lengths is shown in Fig. 4 (for *Sea 1* only) and in Fig. 5 for all other source fingerprints.

Variation of aerosol mass for *Sea1* is the highest of all sources, with the coefficient of variation in the range of 1.9–2.3 for all trajectories considered. This is probably due to it being the lowest source mass (1% of total mass); the second lowest source mass (*Oil/Diesel*) is 5%. Although Hok Tsui is located in coastal areas, Ho et al. (2003) found that the contribution of sea salt to the fine particulate (PM<sub>2.5</sub>) was small, while at the same time it was a major contributor to PM<sub>10</sub>. We found (Fig. 4) that 1-day clustering offered the best results for fresh sea salt when horizontal distance and altitude defined the metric space used. This is consistent with residence

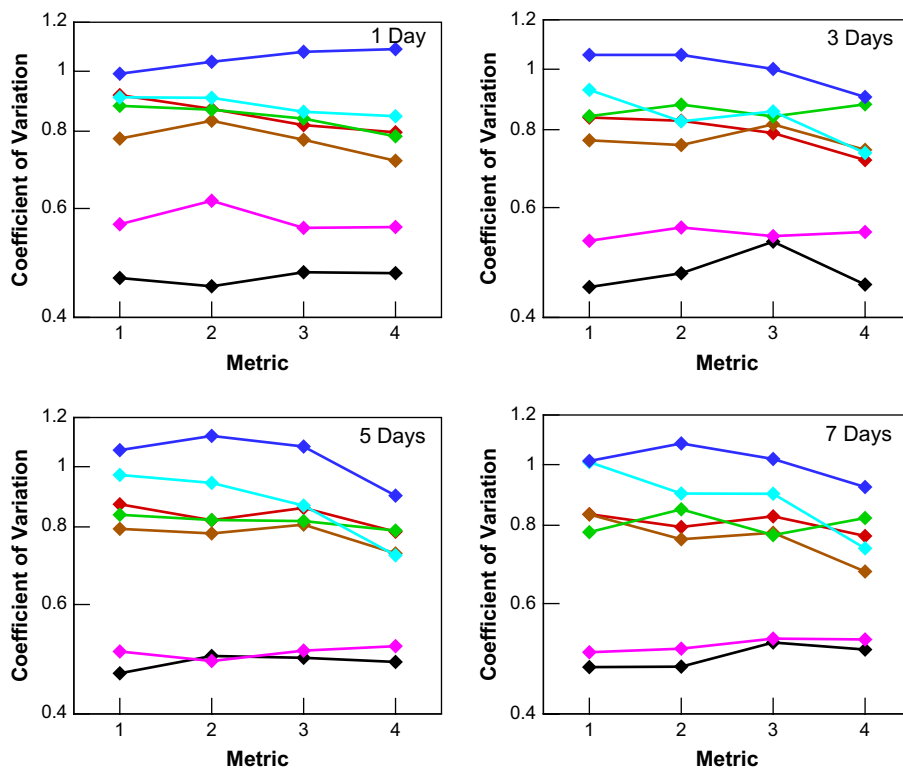


Fig. 5. Coefficient of variation of each source fingerprint mass calculated in four metric spaces and for the four trajectory lengths considered. The colour code is *Oil/Diesel* (black), *Org/Coal* (red), *Smoke* (brown), *Sea 2* (green), *Auto* (light blue), *Soil* (dark blue), *2ndryS* (magenta). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

time of sea-salt aerosols, found to range from 30 min ( $rd = 4\text{--}8\ \mu\text{m}$ ) to 60 h ( $rd = 0.13\text{--}0.25\ \mu\text{m}$ ) in the first atmospheric layer (0–166 m) depending on the size of the particles and local meteorological conditions (Gong and Barrie, 1997). It seems that the variation of aerosol mass for *Sea1* is determined by the distribution of the *Sea1* source function around Hok Tsui, which is in turn strongly affected by local meteorology.

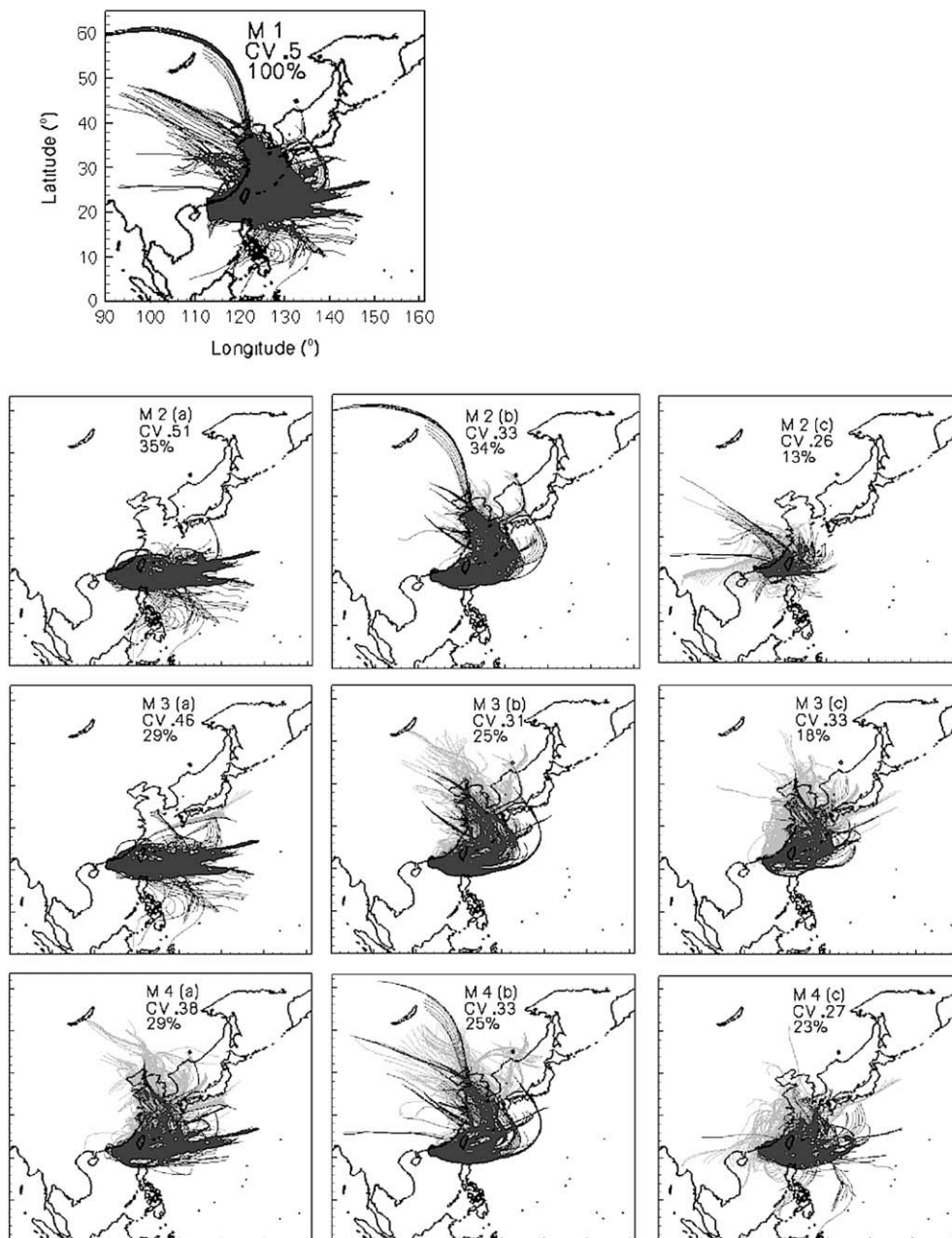
A well separated group (Fig. 5) of sources comprising *Sea 2*, *Soil*, *Organic/Coal*, *Smoke* and *Auto* (accounting for approximately 68% of the measured aerosol mass) shows the variation of their mass as about half of that for *Sea 1* and ranging from about 0.7 to 1.1 with *Soil* the strongest and *Smoke* the weakest. Of the five sources, only the *Sea2* source does not seem to change with the metric. The

variation of mass of each of the other four sources in the group depends on the metric space used, with Metric 4 producing the minimum for 15 out of 16 cases.

Variation of the mass of *Oil/Diesel* and *2ndryS* is consistently the lowest (Fig. 5). This is probably due to a homogeneous distribution of these two sources around the receptor site.

The nature of the local and distant sources is reflected when considering the variation of aerosol mass for each of the source factors for different combinations of the metric space and back trajectory length used.

For *Oil/Diesel*, it appears that the combination of Metric 2 and 1-day back trajectories minimises the variation. Louie et al. (2005a,b) reported very little urban–rural contrast for *Oil/Diesel* as it results



**Fig. 6.** Seven day back trajectories classified in a specific cluster when Metric 1 is used (1st Row), followed by the redistribution of these trajectories (shown in black) in separate clusters when Metric 2 (2nd Row), Metric 3 (3rd Row) and Metric 4 (4th Row) is used. The indicated percentage is the percentage of the original trajectories that have been classified in the top cluster. Note that the sum of percentages does not add up to 100, as only the 3 most populated (majority) clusters for each of the Metrics 2–4 are included in the figure.

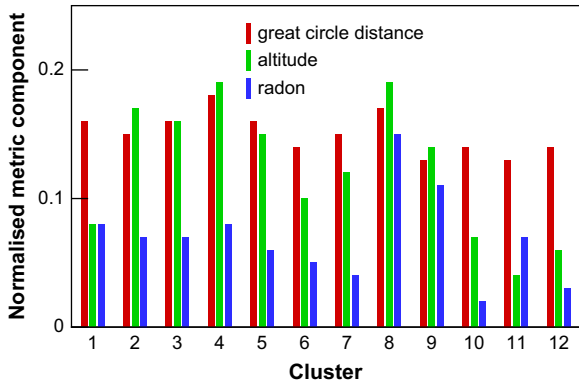


Fig. 7. Average contribution to the distance measure from Metric 4 terms (great-circle distance, altitude and radon) for each of the 12 clusters.

from residual oil-fired combustion and the residual from oil and diesel fuel used in marine vessels. The variation of mass of *2ndryS* (which accounts for 26% of the total mass) shows the best result for Metric 2 and 5 days back trajectories. These results reflect the local and regional nature of *Oil/Diesel* and *2ndryS* sources, respectively (Qin et al., 1997 and Louie et al., 2005a,b).

Metric 4 provided the smallest coefficient of variation for *Org/Coal*, *Smoke* and *Auto* for all trajectory lengths. Metric 4 also showed the smallest variation for *Soil* for 3 days and longer. This is not surprising as these aerosols are of continental nature and radon is a tracer for recent continental fetch.

*Sea 2* shows little variation in the distance measures used. Cl depleted sea salt is formed as a result of NaCl reaction with pollutants such as Nitric and Sulphuric acids, which require the

mixing of marine air with polluted mainland air and could be the reason for the better results observed when trajectories are classified on the synoptic scale.

Overall, for back trajectory lengths of 3 days or more, Metric 4 results in the lowest variation of source mass. *Oil/Diesel* is an exception, but its contribution to total mass is small (5%) and it has both land and sea origins, hence Metric 1 performs better for this source.

### 3.3. Impact of metrics on cluster membership

Different metric spaces for cluster generation define different membership of the clusters. Hence, trajectories belonging to a cluster defined in one metric space will be redistributed to a set of clusters defined in another metric space. It is therefore important to examine the influence of the four metric spaces defined in equations (1)–(4) on cluster membership and, in particular, whether the metrics result in any artefacts which can be detected in resulting clusters on the horizontal plane. The latter is important as the best performing metric space (M4) does include, apart from altitude, a non-spatial component (radon).

We have followed the redistribution process by tracking trajectories originally classified in one cluster formed in Metric 1 space. The cluster (M1) we chose for illustrating the process is shown in the top row in Fig. 6. The graphs in the second row show the redistribution of M1 trajectories among clusters defined in Metric 2 space (Fig. 6 M2 (a)–M2(c)). The third and fourth rows show the same for Metrics 3 and 4. All redistributed trajectories are shown in black whereas the other trajectories belonging to M2–M4 clusters are in grey. The coefficient of variation as well as the percentage of the original trajectories redistributed to M2–M4 clusters is shown in each graph. For clarity, the figure includes only the three M2–M4 which contain the majority of the redistributed trajectories.

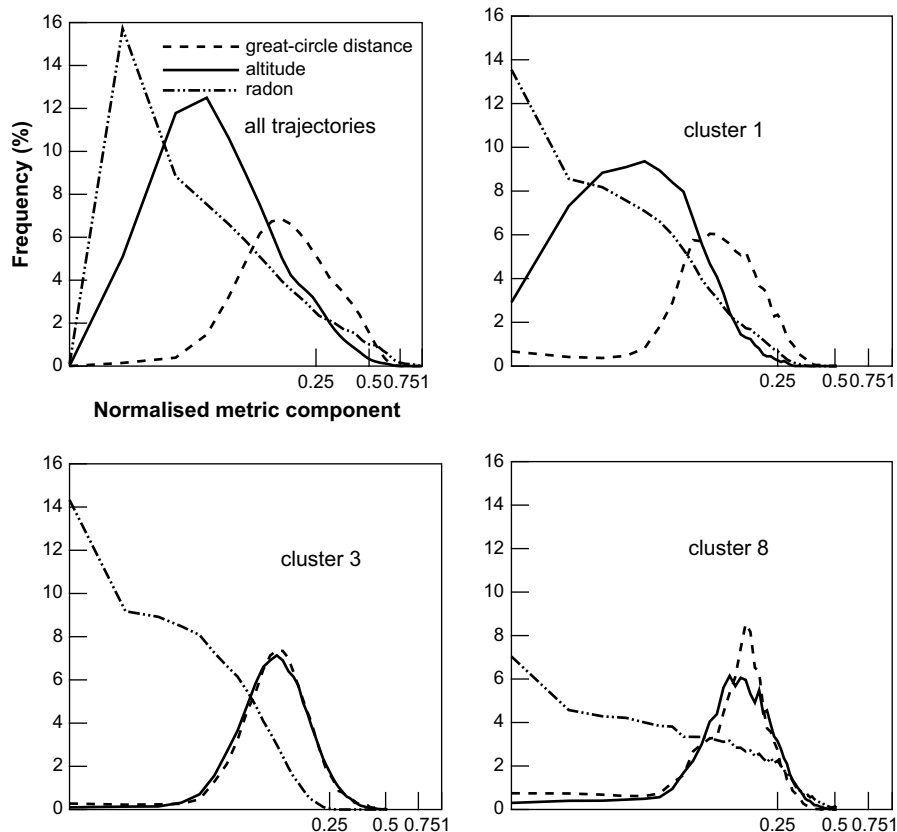


Fig. 8. Example of distributions of the three terms (great-circle distance, altitude and radon) in Metric 4.

Examination of the results in Fig. 6 indicates that the redistribution does not introduce any artefacts and that the geographical coverage of these clusters is consistent with the original. About 75% of the original M1 trajectories are contained in three clusters formed using any of the M2–M4 metrics. These observations apply to other redistributions, i.e. where trajectories were traced from 11 other M1 clusters.

A common feature for all redistributions is that when the aerosol mass variation in the original M1 cluster is high (i.e. the coefficient of variation is higher than 0.4), the variation is substantially reduced in the corresponding majority M2–M4 clusters. This is illustrated in Fig. 6. When the variation in the original cluster is about 0.3 or lower then the redistribution does not lead to any substantial reductions or increases in the variation coefficient. It is interesting to see that the trajectories classified in a cluster which appears similar in M2 (a) and M3 (a) have been redistributed based on radon under the M4 metric resulting in smaller mass variation. This has better resolved the terrestrial influence on the air masses.

### 3.4. Analysis of Metric 4 results

We have shown that clusters formed in the Metric 4 space offer best performance in terms of the variation of the measured aerosol mass. Overall, the same applies to fingerprint source mass. Hence, in this section we focus on results obtained using Metric 4.

#### 3.4.1. Distribution of distance metric

We examine the contribution of three components: great-circle distance (horizontal), altitude, and radon in Metric 4. Fig. 7 shows the three average (within cluster) contributions for each of the 12 clusters based on 7 day back trajectories. The horizontal term dominates in 8 clusters, thus justifying the attention given to it in previously used metrics. The metric term based on altitude is similar to or higher than the horizontal term in 5 clusters, signifying the importance of this component, contrary to previous suggestions (Hafner et al., 2007). The radon term is similar to the other two in 2 clusters (8 and 9). It is also significant (above 30%) in 7 other clusters.

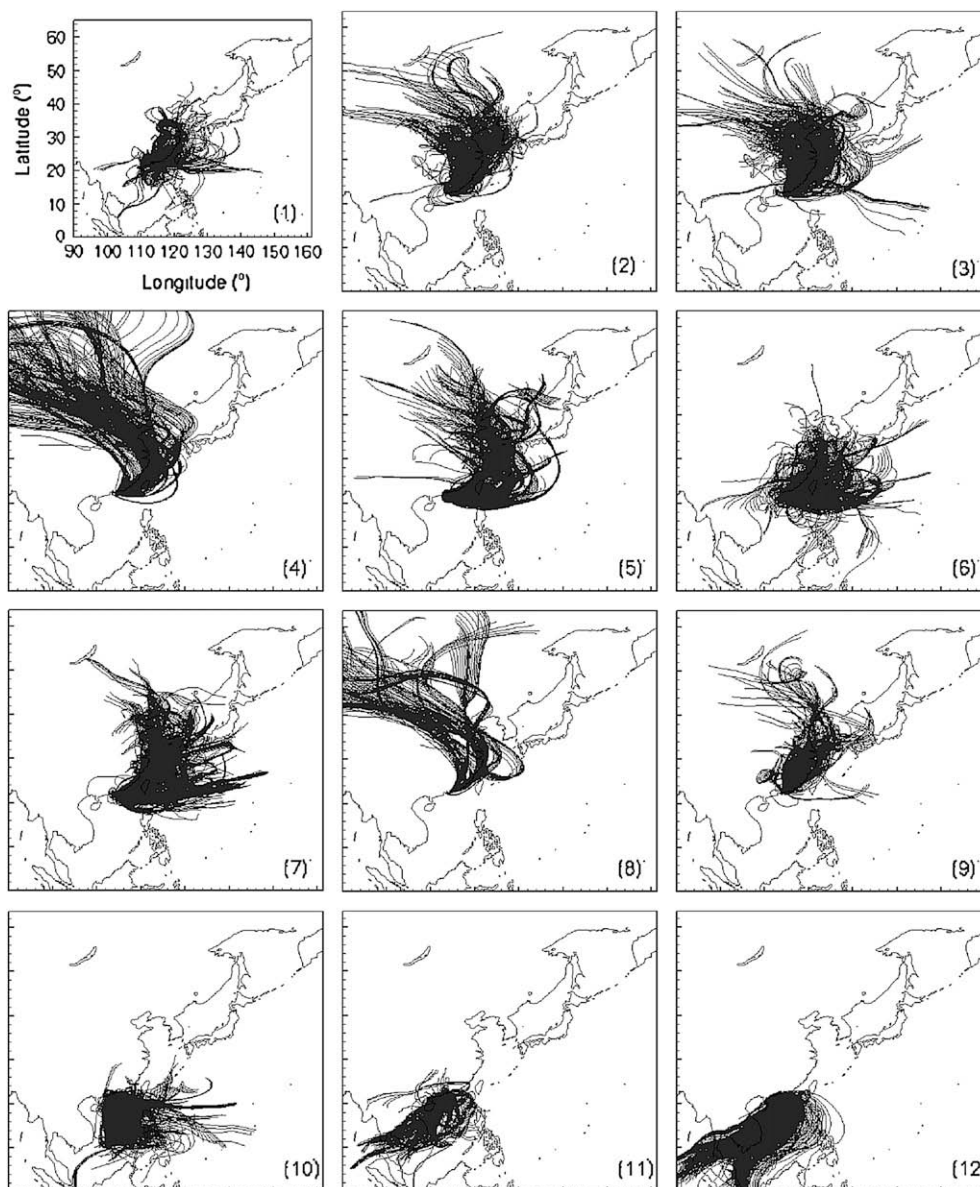


Fig. 9. The 12 clusters for 7 days back trajectories and Metric 4. For clarity, only 5 days of back trajectories are displayed.

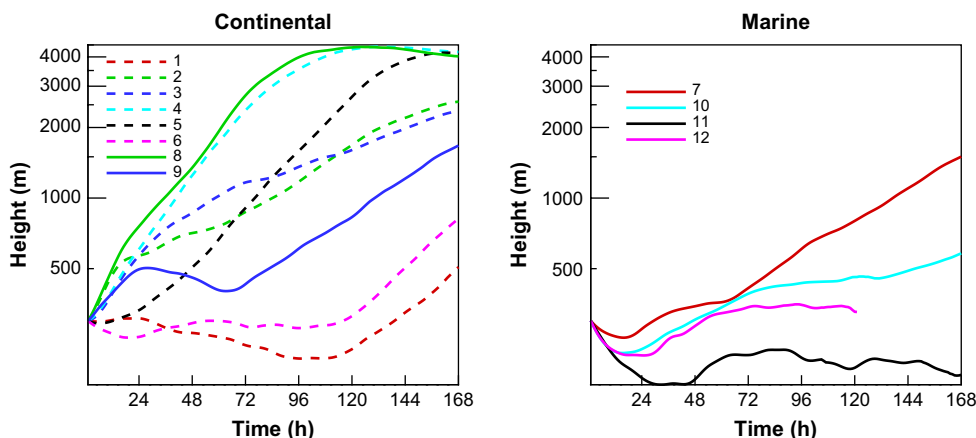


Fig. 10. Average altitude of trajectories within each cluster (for 7 days back trajectories and Metric 4 distance measure) at each end point back in time from the receptor.

The distribution of the three terms for all trajectories is shown in Fig. 8. The terms, calculated between pairs of trajectories within three selected clusters, are shown as well to illustrate typical variations in the distribution for cluster groups where the horizontal term dominates (cluster 1), both horizontal and vertical (altitude) terms dominate (cluster 3), and where all three terms make a similar contribution (cluster 8).

3.4.2. Aerosol analysis within clusters

The 12 clusters generated in Metric 4 space are presented in Fig. 9. The average altitude within each cluster as a function of time before arrival at the receptor site is shown in Fig. 10. Table 1 quantifies seasonal contributions to the clusters. A summary of essential cluster characteristics is given in Table 2. The table includes the average radon and average total mass for each cluster as well as the average mass for each fingerprint within the clusters. Contributions to the average mass are calculated for days where 13 or more hourly trajectories have been classified in the same cluster (see Section 3.2).

All analysed air masses can be divided into four groups of clusters with different average altitude and, to a lesser degree, speed. Division by altitude clearly follows from Fig. 10. Of the four groups, three are, to a various degree, continental air masses. The fourth group comprises air masses of marine origin. We discuss the main characteristics of the groups below, starting from the continental group with the highest average altitude and ending with the marine group where clusters are closest to the surface and mostly within the marine boundary layer.

Group 1 is comprised of clusters 4 and 8 and represents 10% of the analysed air masses (Table 1). The clusters correspond to air

masses originating further inland. The average altitude is the highest (Fig. 10) with the longest land fetch (Fig. 9). These events occur mostly in Winter and Autumn (Table 1). Differences in the two clusters are mainly due to the way the air masses approach the receptor site, with Cluster 8 taking a land route along the northern coast, whereas Cluster 4 approaches from the East China Sea to Hong Kong (Fig. 9). Consequently, average radon is about four times higher for Cluster 8 indicating a strong land influence. The average total aerosol mass for Cluster 8 is amongst the highest of all clusters. The cluster has also the highest *Smoke* and *Auto*, indicating that these two sources have their origins along the coast north of Hong Kong (Crawford et al., 2007). The aerosol mass for Cluster 4 is in the middle of the range. The second highest *Soil* composition and the highest old sea air (*Sea 2*) indicates that air masses with significant pollution species are entrained over land and stay long enough over the sea to result in depletion of Cl from fresh sea-salt aerosols (Song and Carmichael, 1999).

Group 2 (Clusters 2, 3, and 5) represents 22% of all air masses and is characterised by the next highest altitude and speed. Air masses originate from land but differ in the distance they travel out to sea before turning to the south west or west to reach the receptor site. The sea fetch increases with the clusters' number along with the rising number of Spring events. This trend results in lower average radon and *Auto* and *Smoke*. At the same time *2ndryS* increases, indicating regional transport (Qin et al., 1997 and Louie et al., 2005a,b).

Group 3 (Clusters 1, 6, and 9) represents 24% of the analysed air masses, with the lowest average altitude among the three continental groups. Cluster 9 represents the coastal air masses; it has the

Table 1

The number of trajectories classified in each cluster by season (in column Count) and the seasonal % of trajectories classified in each cluster. The cluster % column gives the % of the total trajectories classified in each cluster. The data are Metric 4 clusters and 7 days back trajectories.

Cluster	Winter		Spring		Summer		Autumn		Cluster %
	Count	%	Count	%	Count	%	Count	%	
1	74	5.7	185	9.8	79	5.2	75	5.4	6.8
2	169	13.0	89	4.7	15	1.0	218	15.7	8.0
3	141	10.9	210	11.1	14	0.9	180	12.9	8.9
4	186	14.4	75	4.0	4	0.3	169	12.1	7.1
5	143	11.0	246	13.0	30	2.0	113	8.1	8.7
6	177	13.7	283	14.9	90	5.9	100	7.2	10.6
7	121	9.3	323	17.0	98	6.4	226	16.2	12.6
8	102	7.9	22	1.2	0	0.0	79	5.7	3.3
9	167	12.9	85	4.5	10	0.7	103	7.4	6.0
10	11	0.8	269	14.2	198	12.9	44	3.2	8.5
11	5	0.4	79	4.2	216	14.1	46	3.3	5.7
12	0	0.0	30	1.6	779	50.8	39	2.8	13.9
Total	1296	100	1896	100	1533	100	1392	100	100

**Table 2**  
The average radon (mBq/m<sup>3</sup>), total aerosol mass (µg/m<sup>3</sup>) and source fingerprint mass (µg/m<sup>3</sup>) for each cluster. The data are Metric 4 clusters and 7 days back trajectories.

Cluster	Radon	Total Mass	Oil/Diesel	Org/Coal	Smoke	Sea 2	Auto	Soil	2ndryS	Sea 1
1	6906	1445	74	416	233	41	164	89	386	5
2	9326	1446	40	287	383	84	226	130	258	4
3	4356	1610	44	350	228	127	228	259	346	3
4	2718	1425	43	224	210	265	144	189	233	11
5	2069	1320	61	291	113	175	124	176	367	4
6	2896	1229	53	272	127	108	115	111	426	6
7	1116	830	55	128	60	105	56	43	345	3
8	10,362	1814	23	294	456	185	451	162	237	9
9	14,492	1352	74	237	317	73	313	59	201	27
10	554	460	60	52	52	41	9	23	228	2
11	3559	566	70	105	63	51	36	68	192	9
12	1010	368	47	45	26	81	4	37	95	29

highest land fetch (occurring mostly in Autumn and Winter, with some in Spring) indicated by the highest average radon concentration. It also shows high *Auto*, *Smoke* and *Oil/Diesel*. Cluster 1 (occurring more often in Spring), with mid range radon concentrations, shows stagnant air circulating around Hong Kong and indicates both land and sea fetch. The highest *Org/Coal* and *Oil/Diesel* indicate the predominance of local sources. Cluster 6 with some continental fetch shows the highest *2ndryS*. Air masses contribute to Clusters 1 and 6 throughout the year with Spring events being most frequent.

Group 4 (Clusters 7, 10–12) is the largest group including 42% of all analysed air masses. Air masses of marine origin dominate. While air masses in Cluster 7 have originated from the East China Sea mainly in Spring and Autumn, those corresponding to Clusters 10–12 are increasingly dominated by Summer events. The group is characterised by low aerosol mass and radon concentrations. The exceptions are *2ndry* and aged sea air *Sea2*, which indicate the availability of some pollutants to react over the ocean. These sources are formed due to reactions in the atmosphere. The low radon suggests the existence of a large proportion of old polluted air with a significant time over the sea. Cluster 10 represents air masses predominantly occurring in Spring and Summer. Its land fetch is the smallest, as is the corresponding radon. *Oil/Diesel* is above average indicating possible sources from shipping. Cluster 12 represents perturbed marine air masses coming in Summer from the south. The presence of *Oil/Diesel* indicates influence of shipping. The highest fresh sea salt probably results from the average low altitude of air masses (Lewis and Schwartz, 2004; Katoshevski et al., 1999).

Multi-correlation analysis for the four groups revealed a strong correlation between radon and *Smoke*, and radon and *Auto*, with correlation coefficients 0.85 and 0.84, respectively. Another important inter-group observation is that the *Soil* factor is the highest for Cluster 3 and is significantly low for the local Cluster 9, indicating that distant sources are more important than local ones. It should be noted that the radon correlation with the total aerosol mass is relatively low and equal to 0.49 which points to a heterogeneous aerosol source function.

#### 4. Conclusion

A novel metric space based on spatial and non-spatial variables has been proposed, tested, and applied.

The overall performance of the new metric space is better than those previously used and based on spatial variables in two or three dimensions alone.

The performance criterion based on the measured or fingerprint apportioned source aerosol mass delivered consistent results.

Application of the new metrics to a re-analysis of the previously published data delivered an easily interpretable set of clusters for identification of aerosol source regions.

The results show that considerable thought is required when using cluster analysis, particularly in the choice of the metric space which, for optimal performance, should also be defined, apart from spatial variables, by suitable tracer(s). The new metric can easily be redefined for this purpose.

The choice of the length of the back trajectory to be used for cluster generation was found to be important and should be based on the predominant source of pollution under investigation. It was shown that for a complex pollution source, an optimisation of the trajectory length is required.

#### Acknowledgments

The authors would like to acknowledge the generous invitation by Tao Wang of the Hong Kong Polytechnic University to conduct the aerosol sampling and radon measurements at the Hok Tsui site. Day-to-day technical assistance was provided by Steven Poon of the same university. The NOAA Air Resources Laboratory (ARL) made available the HYSPLIT transport and dispersion model and the relevant input files for generation of back trajectories used in this paper.

#### References

- Brankov, E., Rao, S.T., Porter, P.S., 1997. A trajectory-clustering-correlation methodology for examining the long-range transport of air pollutants. *Atmospheric Environment* 32 (9), 1525–1534.
- Cape, J.N., Methven, J., Hundson, L.E., 2000. The use of trajectory cluster analysis to interpret trace gas measurements at Mace Head, Ireland. *Atmospheric Environment* 34, 3651–3663.
- Conen, F., Robertson, L.B., 2002. Latitudinal distribution of radon-222 flux from continents. *Tellus* 54B, 127–133.
- Crawford, J., Chambers, S., Cohen, D.D., Dyer, L., Wang, T., Zaborowski, W., 2007. Receptor modelling using positive matrix factorisation, back trajectories and Radon-222. *Atmospheric Environment* 41, 6823–6837.
- Dorling, S., Davies, T., 1992. Cluster analysis: a technique for estimating the synoptic meteorological controls on air and precipitation chemistry – method and application. *Atmospheric Environment* 26A (14), 2575–2581.
- Draxler, R.R., 1991. The accuracy of trajectories during ANATEX calculated using dynamic model analysis versus rawinsonde observations. *Journal of Applied Meteorology* 30, 1466–1467.
- Draxler, R.R., Rolph, G.D., 2003. Hybrid Single-Particle Lagrangian Integrated Trajectory (HYSPLIT) Model. <http://www.arl.noaa.gov/ready/hysplit4.html>.
- Gong, S.L., Barrie, L.A., 1997. Modeling sea-salt aerosols in the atmosphere. *Journal of Geophysical Research* 102 (D3), 3805–3818.
- 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. *Atmospheric Environment* 41, 3071–3081.
- Harris, J.M., Kahl, J., 1990. A descriptive atmospheric transport climatology for the Manua Loa Observatory, using clustered trajectories. *Journal of Geophysical Research* 95 (D9), 13651–13667.
- Harris, J.M., Draxler, R.R., Oltmans, S.J., 2005. Trajectory model sensitivity to differences in input data and vertical transport method. *Journal of Geophysical Research* 110, D14109.
- Ho, K.F., Lee, S.C., Chan, C.K., Yu, J.C., Chow, J.C., Yai, X.H., 2003. Characterization of chemical species in PM<sub>2.2</sub> and PM<sub>10</sub> aerosols in Hong Kong. *Atmospheric Environment* 37, 31–39.
- Hopke, P.K., Li, C.L., Ciszek, W., Landsberger, S., 1995. The use of bootstrapping to estimate conditional probability fields for source locations of airborne pollutants. *Chemometrics and Intelligent Laboratory Systems* 30, 69–79.

- 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. *Journal of Applied Meteorology* 43, 887–901.
- Katoshevski, D., Nenes, A., Seinfeld, J.H., 1999. A study of processes that govern the maintenance of aerosols in the marine boundary layer. *Journal of Aerosol Science* 30, 503–532.
- Kaufman, L., Rousseeuw, P.J., 2005. *Finding Groups in Data, an Introduction to Cluster Analysis*. John Wiley and Sons, Inc, New York, ISBN 0-471-87876-6.
- Lewis, E.R., Schwartz, S.E., 2004. *Sea Salt Aerosol Production: Mechanics, Methods, Measurements and Models – A Critical Review*. American Geophysical Union, Washington DC, ISBN 087590-417-3.
- Lin, C.L., Cheng, M.D., Schroeder, W.H., 2001. Transport patterns and potential sources of total gaseous mercury measured in Canadian high Arctic in 1995. *Atmospheric Environment* 35, 1141–1154.
- Louie, P.K.K., Watson, J.G., Chow, J.C., Chen, A., Sin, D.W.M., Lau, A.K.H., 2005a. Seasonal characteristics and regional transport of PM<sub>2.5</sub> in Hong Kong. *Atmospheric Environment* 39, 1695–1710.
- Louie, P.K.K., Watson, J.G., Chow, J.C., Chen, L.W.A., Sin, D.W.M., Lau, A.K.H., 2005b. PM<sub>2.5</sub> chemical composition in Hong Kong: urban and regional variations. *Science of the Total Environment* 338, 267–281.
- Man, C.K., Shih, M.Y., 2001. Identification of sources of PM<sub>10</sub> aerosols in Hong Kong by wind trajectory analysis. *Journal of Aerosol Science* 32, 1213–1223.
- Moody, J.L., 1986. The influence of meteorology on precipitation chemistry at selected sites in the Eastern United States. Ph.D. Thesis, Univ. of Mich., Ann Arbor, 176 pp.
- Moody, J.L., Galloway, J.N., 1988. Quantifying the relationship between atmospheric transport and the chemical composition of precipitation on Bermuda. *Tellus* 40B, 463–479.
- Owega, S., Khan, B.-U.-Z., Evans, G., Jervis, R.E., Fila, M., 2006. Identification of long-range aerosol transport patterns to Toronto via classification of back trajectories by cluster analysis and neural network techniques. *Chemometrics and Intelligent Laboratory Systems* 83 (1), 26–33.
- Paatero, P., Tapper, U., 1994. Positive Matrix Factorisation: a non-negative factor model with optimal utilisation of error estimates of data values. *Environmetrics* Vol 5, 111–126.
- Qin, Y., Chan, C.K., Chan, L.Y., 1997. Characteristics of chemical compositions of atmospheric aerosols in Hong Kong: spatial and seasonal distributions. *The Science of the Total Environment* 206, 25–37.
- Seinfeld, J.H., Padnis, S.N., 1998. *Atmospheric Chemistry and Physics, From Air Pollution to Climate Change*. John Wiley and Sons, Inc, New York.
- Sinnott, R.W., 1984. Virtues of the haversine. *Sky Telescope* 68, 159.
- Song, C.h., Carmichael, G.R., 1999. The aging process of naturally emitted aerosol (sea-salt and mineral aerosol) during long range transport. *Atmospheric Environment* 33, 2203–2218.
- Stohl, A., 1998. Computation, accuracy and applications of trajectories – a review and bibliography. *Atmospheric Environment* 32 (6), 947–966.
- Wang, T., Ding, A.J., Blake, D.R., Zahorowski, W., Poon, C.N., Li, Y.S., 2003. Chemical characterization of the boundary layer outflow of air pollution to Hong Kong during February–April 2001. *Journal of Geophysical Research* 108 (D20), 8787.