Assessment and Application of Clustering Techniques to Atmospheric Particle Number Size Distribution for the Purpose of Source Apportionment

F. Salimi, Z. Ristovski, M. Mazaheri, R. Laiman, L. R. Crilley*, C. He, S. Clifford and L. Morawska

International Laboratory for Air Quality and Health, Queensland University of Technology, GPO Box 2434, Brisbane QLD, 4001, Australia

[*]{now at: School of Geography, Earth and Environmental Science, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK}

Correspondence to: L. Morawska (l.morawska@qut.edu.au)

Abstract

Long-term measurements of particle number size distribution (PNSD) produce a very large number of observations and their analysis requires an efficient approach in order to produce results in the least possible time and with maximum accuracy. Clustering techniques are a family of sophisticated methods which have been recently employed to analyse PNSD data, however, very little information is available comparing the performance of different clustering techniques on PNSD data. This study aims to apply several clustering techniques (i.e. K-means, PAM, CLARA and SOM) to PNSD data, in order to identify and apply the optimum technique to PNSD data measured at 25 sites across Brisbane, Australia. A new method, based on the Generalised Additive Model (GAM) with a basis of penalised B-splines, was proposed to parameterise the PNSD data and the temporal weight of each cluster was also estimated using the GAM. In addition, each cluster was associated with its possible source based on the results of this parameterisation, together with the characteristics of each cluster. The performances of four clustering techniques were compared using the Dunn index and Silhouette width validation values and the K-means technique was found to have the highest performance, with five clusters being the optimum. Therefore, five clusters were found within the data using the K-means technique. The diurnal occurrence of each cluster was used together with other air quality parameters, temporal trends and the physical properties of each cluster, in order to attribute each cluster to its source and origin. The five clusters were attributed to three major sources and origins, including regional background particles, photochemically induced nucleated particles and vehicle generated particles. Overall, clustering was found to be an effective technique for attributing each particle size spectra to its source and the GAM was suitable to parameterise the PNSD data. These two techniques can help researchers immensely in analysing PNSD data for characterisation and source apportionment purposes.

1 Introduction

Atmospheric aerosols affect climate, air quality and subsequently human health (Stevens and Feingold, 2009;Pope and Dockery, 2006;Lohmann and Feichter, 2005). Despite their small contribution to particle volume and mass, ultrafine particles (particles with diameter <100nm) make a significant contribution to particle number concentration (PNC) (Morawska et al., 1998;Harrison and Yin, 2000) and toxicological studies show evidence of their adverse effects on human health (WHO, 2006). Therefore, measurements of the chemical and physical properties of aerosol particles are crucial in order to understand their effects on climate and human health. One of the most important properties of particles is their size distribution, which helps in understanding aerosol dynamics, as well as determining their sources (Charron et al., 2008;Harrison et al., 2011). Long-term particle number size distribution (PNSD) measurements have been conducted in a number of different environments and the measured size range can extend from less than 10nm up to more than 10µm. In addition, these long-term measurements generally result in a large number of observations and analysing such a massive data set often requires sophisticated techniques. The clustering technique has recently been used to divide particle size data into groups with similar characteristics and then relate each group to its sources and/or to investigate aerosol particle formation and evolution (Beddows et al., 2009;Dall'Osto et al., 2012;Wegner et al., 2012;Tunved et al., 2004;Costabile et al., 2009;Charron et al., 2007).

Several clustering algorithms currently exist, which makes the selection of an appropriate clustering technique a daunting task. Determining the most appropriate number of clusters can be an additional challenge for researchers. Clusters should ideally be compact, well-separated and scientifically relevant. Beddows et al (Beddows et al., 2009) assessed the performance of four clustering techniques (Fuzzy, K-means, K-median and model based clustering) on PNSD data using different validation indices, particularly the Dunn index, and found the K-means technique capable of finding clusters with smallest size, furthest separation and highest degree of inner cluster similarity compared to others. Throughout their work, four techniques were evaluated while several other methods (e.g. Partitioning around Medoids (PAM), Clustering of Large Applications (CLARA), and Self Organizing Map (SOM), and Affinity Propagation (AP)) are available which their performance on PNSD data has not been assessed so far. Therefore, these techniques were selected to be compared with K-means using two validation measures.

K-means is an iterative algorithm minimising the within cluster sum of squares to find a given number of clusters (Hartigan and Wong, 1979). PAM is an iterative algorithm similar to K-means which constructs clusters around a set of representative objects by assigning each data to the nearest representative object using sum of pair wise dissimilarities (Kaufman and Rousseeuw, 2009). CLARA performs PAM on a number of subgroups of data, allowing faster performance for a large number of observations (Kaufman and Rousseeuw, 2009). SOM is a neural networks based method, with the ability to map high dimensional data to two dimensions and has been widely used in data mining researches (Kohonen, 2001). The AP algorithm is a relatively new clustering technique which has been employed in different fields since its introduction in 2007. AP considers all data as potential exemplars and finds the best set of exemplars and corresponding clusters by exchanging messages between the data points (Frey and Dueck, 2007).

In a recent study, three years of PNSD data were clustered using the K-means technique to produce seven clusters. Those clusters were found to form three main groups, anthropogenic (69%), maritime (29%) and nucleation (2%), which characterised the whole data set (Wegner et al., 2012). In another study, Dall’Osto et al. found nine clusters within the PNSD data collected over a one year period in an urban area and found four typical PNSD groups using diurnal variation, directional and pollution association (Dall'Osto et al., 2012). The authors called those groups traffic, dilution, summer background and regional pollution, which included 69%, 15%, 4% and 12% of the total data respectively. In the above and all other previous studies, PNSD data were averaged to decrease the number of data and consequently, reduce computational cost and complexity. However, averaging can encumber the transient characterisation of PNSD data and the larger the averaging interval, the more transient characteristics will be lost.

Parameterisation of PNSD data in terms of a mixture of few log-Normal components is common and beneficial, particularly for data averaged over a longer interval. Multi-log-Normal function with predefined number of peaks where the means of each log-Normal distribution are constrained to vary around some initial estimates in the nucleation, Aitken, accumulation and coarse modes have been used in literature (Hussein et al., 2004;Hussein et al., 2005;Heintzenberg et al., 2011;Shen et al., 2011). However, not all of the measured particle size data are able to be expressed in this way and this imposing a predefined number of separated peaks may not accurately represent all of the variation in the collected data. In addition, this method can result in losing the transient trends if applied to each single particle size spectra.

This study aimed to identify the optimum clustering technique and number of clusters by comparing the performance of three clustering techniques (i.e. PAM, CLARA, and SOM) with the K-means technique for several numbers of clusters and to associate each cluster with its possible sources using the cluster characteristics, PNSD parameterisation results, diurnal variation, temporal variation and several air quality parameters.

2 Materials and Methods

2.1 Background

This study was performed within the framework of the Ultrafine Particles from Traffic Emissions and Children's Health (UPTECH) project, which aimed to determine the effects of exposure to traffic related ultrafine particles (UFPs) on the health of primary school-aged children. Air quality measurements were conducted for two consecutive weeks at each of the 25 randomly selected state primary schools across the Brisbane Metropolitan Area, in Australia, during the period October 2010 to August 2012. Further details regarding the UPTECH project can be found in (Salimi et al., 2013) and the study design is available online (UPTECH).

2.2 Instrumentation, quality assurance, and data processing

PNSD within the size range 9 - 414 nm was measured every 5 minutes using a TSI Scanning Mobility Particle Sizer (SMPS). The SMPS system included a TSI 3071 Differential Mobility Analyser (DMA) connected to a TSI 3782 water-based Condensation Particle Counter (CPC). A combination of a diaphragm pump and a critical orifice was used to supply a sheathe flow of 6.4 lpm. A zero particle filter and silica gel dryer were used to supply a dry, particle free air stream. PNC was measured using a TSI 3781 water-based CPC, particle mass concentration (PM2.5, and PM10) measurements were conducted using a TSI DustTrak, solar radiation and other meteorological parameters were measured by a Monitor Sensors weather station, and EcoTech gas analysers measured gaseous emission (i.e. CO, NOx and SO2) concentrations. These data were averaged according to five minute intervals prior to data analysis.

The sheathe and aerosol flow rate of the SMPS system was checked three times a week using a bubble flow meter. A zero check of the system was done at the start of the measurements at each school using a high efficiency particle (HEPA) filter connected to the inlet of the system. Size accuracy of the SMPS was calibrated using monodisperse polystyrene latex (PSL) particles with a nominal diameter of 100 nm. Size accuracy calibrations were conducted five times throughout the whole measurement campaign and all instruments passed the test with a maximum error of 3.5% from the nominal diameter, as recommended by Wiedensohler et al (Wiedensohler et al., 2012). Particle losses inside the tube were corrected using the formula derived for the laminar flow regime (Hinds, 1999). Equivalent tube length was used to correct for particle loss inside the bipolar charger and DMA (Karlsson and Martinsson, 2003;Covert et al., 1997). On-site calibration checks (span and zero) of the gas analysers were conducted on the second day of the two week measurement campaigns at each school when the analysers reached the stable running conditions. The DustTrak zero check was performed every second day and calibrated if it was not showing zero within the uncertainty of the instrument. Information regarding the quality assurance and data processing procedures for the CPC can be found in (Salimi et al., 2013).

2.3 Clustering particle number size distribution data

Principal Component Analysis (PCA) was performed on the whole data set, in order to reduce its high dimensionality and remove the correlation between features, so as to increase the performance of the clustering. Ideally, clusters should minimise intra-cluster variation (compactness) and maximise the distance between clusters (separation) (Handl et al., 2005;Brock et al., 2008), resulting in small, homogenous clusters which are clearly separated from each other. Validation measures reflecting the compactness and separation of the clusters were used to find the optimal method and number of clusters using the “clValid” package in R (R Development Core Team, 2010;Brock et al., 2008). As the number of clusters increases, improving compactness, the separation decreases due to multiple clusters being created which could be described by a single cluster (consider the extreme case of every observation belonging to its own cluster). Combining compactness and separation into a single measure is an effective way to address this issue. The Dunn index (Dunn, 1974) and Silhouette width (Rousseeuw, 1987) are scores resulting from nonlinear combination of compactness and separation. Therefore, those scores were chosen in order to compare the performance of different clustering techniques and to find the optimum number of clusters.

The maximum vector length allowed by R is “231-1”, therefore, the Dunn index can only be calculated for a maximum of 46,340 observations. To address this issue, half of the observations were randomly selected and their cluster validity values calculated by applying PAM, CLARA, SOM and K-means techniques using 2 to 20 clusters. Then validity values for the other half of observations were calculated. This procedure allowed us to evaluate the whole set of observations considering the vector size limitation.

The AP clustering technique was initially selected as a candidate, in addition to the aforementioned techniques. The AP technique was implemented using “APcluster” R package (Bodenhofer et al., 2011) and was computationally expensive but ultimately unsuccessful at clustering our huge dataset effectively. Based on our experience with this technique and the recommendations of its developers, AP is not recommended for finding a small number of clusters in a large dataset, but it is more appropriate for finding a large number of relatively small clusters.

2.4 Non-parametric estimation of particle number size distribution temporal trends

In this paper, a new approach was developed to parameterise the PNSD data by finding the local peaks and the normalised concentration at each peak. In order to find the real local peaks, the noisy trend of PNSD data should be firstly smoothed. To achieve this, we used the Generalised Additive Model (GAM), with a basis of penalised B-splines (Wood, 2003;Eilers and Marx, 1996), which allows for the flexible estimation of non-linear effects without assuming, a priori, the functional form of the non-linearity (Wood, 2003). In contrast with the multi lognormal fitting method, this approach keeps all the local peaks while smoothing the noisy data. The fitted function was then used to find the local peaks, which were defined to be the local maxima in a neighbourhood of five PNSD bins on each side (Figure 1). The bi-variate kernel density estimate, using the Gaussian kernel, was computed to visualise the distribution of the peaks of the PNSD data for each cluster (Wand, 1994).

Understanding the temporal trend of clusters provides further information about the nature and source of each cluster. GAMs allows for a flexible approach to investigating these temporal trends. Therefore, the GAM, with a basis of B-spline, was employed to calculate the temporal trend of each cluster. The resulting fitted smooth functions, and their 95% confidence intervals, indicate the temporal variation of each cluster.

3 Results and discussion

3.1 Particle number size distribution

Around 82,000 SMPS measurements with 5 min intervals were conducted during the whole UPTECH project, which comprised about 285 days of measurements. All of the previous long-term studies averaged the data to reduce the complexity and calculation costs, however, averaging can conceal the transient peaks and troughs in PNSD data (Beddows et al., 2009). Therefore, we decided not to average the measured PNSD data and to use the high performance computers for handling such a large size data instead.

3.2 Optimum clustering technique and number of clusters

Figure 2 illustrates the performance of each clustering technique for two randomly selected data. Horizontal and vertical axis show the number of clusters and validation value (Dunn index and Silhouette value) respectively. K-means produced the highest Dunn index and Silhouette width values for all of the number of clusters, for both sets of randomly selected data, while the rest of techniques showed the same level of performance (Figure 2). This indicates that the K-means technique was able to produce more compact clusters that were well separated from each other. The other techniques resulted in almost the same Dunn index value, however SOM resulted in better Silhouette width values, followed by PAM and then

CLARA. The performance of the K-means technique was significantly higher than the rest of the techniques and was therefore selected as the preferred technique.

Cluster schemes with 2 - 8 and 10 clusters had the highest Silhouette width values in the first and second set of randomly selected data, respectively. Five clusters had the highest Dunn index value for the first set of data and it resulted in a high value in the second set as well. However, 9 - 10 clusters resulted in a higher Dunn index value in the second set of data. As much as possible, clustering with the optimum number of clusters should have the highest validation value, while still having an appropriate number of clusters for distinguishing between different sources and processes (Wegner et al., 2012;Beddows et al., 2009). Therefore, five was selected as the optimum number of clusters, as it had a high validation value, as well as scientifically relevant number of clusters to relate each cluster to a source and to find the processes related to each cluster.

3.3 Clustering PNSD data

The K-means clustering technique was applied to all PNSD data in order to find five clusters. Figure 3 shows the normalised number spectra and 95% confidence interval associated with each cluster, together with the diurnal variation of their occurrence and their association with solar radiation intensity, particle mass concentration, PNC and gaseous pollutants concentrations.

Local peaks of each single PNSD spectra were found using the technique described in section 2.4. Plotting the normalized concentration of the peaks versus their diameter is useful in order to determine the frequency of peaks and their diameter for each cluster. However, with too many data, points are over-plotted which makes it impossible to distinguish the underlying trends and relationships. Therefore, Bi-kernel density estimation, which is a very effective technique to address this issue, was used to visualise the distribution of peaks in PNSD for each cluster (Figure 4). Characteristics of each cluster and the associated sources are summarised in Table 1 and explained in the following sections, based on Figures 3, 4, and5.

Cluster1: This cluster included 4.5% of the total measured particle size spectra with a mode at the SMPS lowest size detection limit (9 nm). The diurnal pattern of occurrence showed nocturnal minima with a peak at midday. Cluster1 was associated with the highest solar intensity and lowest PM2.5 among all clusters, and it was also associated with high PNC and low CO and NOx. Local peaks predominantly occurred at diameters lower than 30nm, with the highest density and normalised concentration at 9nm. This clearly shows the dominance of newly formed particles which have grown in size and reached the instrument's size detection limit. The strength of the local peaks was the highest among all clusters, indicating the dominance of the smallest particles.

The abovementioned observations indicate that this cluster was attributed to photochemically-induced nucleated particles in the ambient air. New particles mainly formed during the middle of the day and grew to reach the instrument's size detection limit. The same type of particles were observed in Barcelona for 4% of the observations (Dall'Osto et al., 2012). These types of particles were not observed in London, which could be due to long time averaging of the data and/or different climatological conditions of the monitored environment in that study (Beddows et al., 2009).

Cluster 2: This cluster included 14.1% of the total PNSD data and showed a mode for particles with a diameter less than 20 nm. The diurnal pattern of occurrence for Cluster 2 peaked strongly at two hours after midday (14:00 LT), with a nocturnal minimum, which was in agreement with its association with high solar radiation intensity. Its diurnal pattern of occurrence and association with a low concentration of traffic generated primary pollutants (NOx and CO) indicated that it had non-traffic related sources. This cluster was also associated with high PNC. Local peaks were mainly present at diameters less than 20 nm, with the highest density occurring at a normalized concentration around 0.03. Local peaks present at diameters larger than 30 nm were lower and corresponded to normalized concentrations of less than 0.01. This cluster was attributed mainly to aged, photochemically-induced nucleated particles in the ambient air. Minor peaks at diameters larger than 30 nm revealed the contribution of vehicle generated particles to this cluster (Morawska et al., 2008). However, their contribution to the total PNC was minimal.

Cluster 1 and 2 were both attributed to the same source of particles, but they were distinguished as relatively fresh and aged, respectively. Nucleated particles initially grew and reached the instrument detection size limit in Cluster 1, before growing further and gradually shifting to Cluster 2 after about two hours. Nucleated particles grew in size by condensation and coagulation, and as they aged their number concentration decreased and their size increased. This illustrates why Cluster 1 had higher PNC but lower PM2.5 compared to Cluster 2.

The fraction of PNSD data related to nucleated particles in this study was higher than in another study in the same environment, which used the classic approach to find the banana shaped new particle formation events (Cheung et al., 2011). This shows that the actual occurrence of new particle formation events is much higher than what is generally found by classic approaches, because in most particle formation events, the newly formed particles would be scavenged by the pre-existing particles as a result of coagulation, before they grew by means of condensation. However, this does not mean that the nucleation events would not occur and that the particles resulting from them would not be present in the air.

Cluster 3: This cluster included 31.6% of the total measured data, with a mode at 20 nm and a minor peak at the smallest instrument size limit (≈ 9nm), showing a strong association with morning and afternoon rush hour traffic (Salimi et al., 2013). This cluster was associated with low solar radiation and high vehicle generated primary pollutants, including CO and NOx, and the particle number size modes were in the range of vehicle generated primary particles and nucleated particles during the exhaust emissions dilution (Casati et al., 2007;Ntziachristos et al., 2007;Janhäll et al., 2004). Local peaks were mainly present at diameters less than 30 nm, particularly at the lowest detection size limit of the instrument. The normalized concentration corresponding to peaks were highest at the smallest particle size. These observations suggest that this cluster was attributed to vehicle generated particles, including primary particles (having a diameter of around 40 nm) and particularly secondary particles formed in the vehicle exhaust (having a diameter of around 9nm). Similar types of particles were observed in the literature (Dall'Osto et al., 2012;Beddows et al., 2009).

Cluster 4: This cluster included 22.6% of the total data, with a mode at 60 nm. The diurnal pattern of occurrence showed nocturnal maxima, with a minimum during the midday and early afternoon hours, and it was associated with the highest PM2.5 among all clusters as well as with high NOx and CO. Local peaks were present at diameters smaller than 20 nm and between 50-70 nm. The normalised concentration corresponding to either of these diameter ranges had a similar magnitude. These findings suggest that cluster 4 was attributed mainly to regional background aerosols. However, a source of small particles and gaseous emissions is also present in this cluster.

Cluster 5: This cluster included 27.2% of the total PNSD data, with a mode at 40 nm. The diurnal pattern of occurrence followed almost the same trend as Cluster 4, with a minimum during the early afternoon and a peak during the night-time. The local peaks’ density was almost evenly distributed through the whole range of diameters, with smaller particles having more peaks particularly for diameters less than 20 nm. However, normalized concentrations corresponding to each local peak were highest at diameters between 30-60 nm, showing the dominance of particles at this size range. The observations mentioned above suggest that, like Cluster 4, this cluster was attributed to regional background aerosols.

3.4 Temporal trend of clusters

A non-parametric regression model was fitted to the daily occurrence fraction of each cluster, in order to determine the temporal trend of each cluster. The regression model quantified the presence of each cluster as the sum of a smooth function for each month and a trend which included the day, month and year. It should be noted that no data were collected during the first two months of the year.

The smooth monthly function for Cluster 1 did not show any significant variation before November, at which point it increased moderately and peaked during December. Solar radiation intensity is also known to increase during the last two months of the year in the southern hemisphere, which indicates that this may be the driving force of more atmospheric nucleation events. Cluster 2 showed similar trends of increase in December in conjunction with the expected increase in solar radiation intensity. The prevalence of Cluster 4 increased and peaked around July-August and then decreased again. Cluster 4 showed the strongest monthly variation among all clusters. As mentioned earlier, this cluster was associated with regional background aerosols and its monthly variation showed a positive correlation with biomass burning within the studied region, which mostly took place during July to September. Biomass burning associated PNSDs peak at 100-200 nm (Friend et al., 2012) , however the PNSD of Cluster 4 peaked at around 60 nm, which was 20 nm larger than the Cluster 5, which was also associated with background aerosols. This implies that Cluster 4 included the mixture of background and biomass burning aerosols, which made the average PNSD peak shift by 20 nm. In addition, peaks at diameters larger than 100nm were present with higher normalised concentration compared to Cluster 5 indicating the effect of biomass burning aerosols (Figure 4).

The prevalence of Cluster 3 decreased, with a trough during August, before increasing again, which is likely to be the result of the biomass burning which occurred during August and consequently decreased the prevalence of particles belonging to Cluster 3. For Cluster 5, a peak was observed during June, which decreased to show a trough during August, before peaking again during November. This trend is the opposite to what was observed for Cluster 4 and given that the peaks for Cluster 5 were observed when there were less biomass burning events, it was most likely associated with regional background aerosols, without the influence of biomass burning events. The PNSDs which were attributed to regional background aerosols in this study had different modes compared to the one observed in London (Beddows et al., 2009), but similar ones to the observations in Barcelona (Dall'Osto et al., 2012).

3.5 Conclusions

In summary, the K-means clustering technique was found to be the preferred technique when compared to SOM, PAM and CLARA. The K-means clustering technique categorised the PNSD data into five clusters and each cluster was attributed to its source and origin. Five clusters were attributed to three major sources and origins, as follows: 1) Regional background particles: Clusters 4 and 5, which included 49.8% of the total data, were attributed to regional background aerosols with clear modes at 60nm and 40nm, respectively; 2) Photochemically induced nucleated particles: Clusters 1 and 2, which included 18.6% of the total data, were attributed to photochemically-induced nucleated particles; and 3) Vehicle generated particles: Cluster 3, which included 31.6% of the data, was attributed to vehicle generated particles. A new method was proposed for the parameterisation of particle size spectra, based on the GAM, which was found to be an effective tool and is recommended to be used for particle size data. K-means clustering successfully attributed each particle size spectra to its source and/or origin. However, while this technique could attribute each particle size spectra to its major contributing source, several sources with different levels of contribution are often responsible for the pattern of each particle size spectra. The structure of the contribution of sources can be further investigated using other techniques, such as Bayesian infinite mixture modelling (Wraith et al., 2011;Kulis and Jordan, 2011), Bayesian K-means, Bayesian Beta-process clustering (Broderick et al., 2012) and positive matrix factorisation (Harrison et al., 2011).

Acknowledgements

This work was supported by the Australian Research Council (ARC), QLD Department of Transport and Main Roads (DTMR) and QLD Department of Education, Training and Employment (DETE) through Linkage Grant LP0990134. Our particular thanks go to R.

Fletcher (DTMR) and B. Robertson (DETE) for their vision regarding the importance of this work. We would also like to thank Prof G. Marks, Dr P. Robinson, Prof K. Mengersen, Prof G. Ayoko, Dr G. Johnson, Dr R. Jayaratne, Dr S. Low Choy, Prof G. Williams, W. Ezz, S. Clifford, M. Mokhtar, N. Mishra, R. Laiman, L. Guo, Prof C. Duchaine, Dr H. Salonen, Dr X. Ling, Dr J. Davies, Dr L. Leontjew Toms, F. Fuoco, Dr A. Cortes, Dr B. Toelle, A. Quinones, P. Kidd and E. Belousova, Dr M. Falk, Dr F. Fatokun, Dr J. Mejia, Dr D. Keogh, Prof T. Salthammer, R. Appleby and C. Labbe for their contribution to the UPTECH project.

References

Beddows, D. C. S., Dall'Osto, M., and Harrison, R. M.: Cluster Analysis of Rural, Urban, and Curbside Atmospheric Particle Size Data, Environmental Science & Technology, 43, 4694-4700, 10.1021/es803121t, 2009.

Bodenhofer, U., Kothmeier, A., and Hochreiter, S.: APCluster: an R package for affinity propagation clustering, Bioinformatics, 27, 2463-2464, 2011.

Brock, G., Pihur, V., Datta, S., and Datta, S.: ClValid: An R package for cluster validation, Journal of Statistical Software, 25, 1-22, 2008.

Broderick, T., Kulis, B., and Jordan, M. I.: MAD-Bayes: MAP-based asymptotic derivations from Bayes, arXiv preprint arXiv:1212.2126, 2012.

Casati, R., Scheer, V., Vogt, R., and Benter, T.: Measurement of nucleation and soot mode particle emission from a diesel passenger car in real world and laboratory in situ dilution, Atmospheric Environment, 41, 2125-2135, 10.1016/j.atmosenv.2006.10.078, 2007.

Charron, A., Birmili, W., and Harrison, R. M.: Factors influencing new particle formation at the rural site, Harwell, United Kingdom, Journal of Geophysical Research: Atmospheres, 112, D14210, 10.1029/2007JD008425, 2007.

Charron, A., Birmili, W., and Harrison, R. M.: Fingerprinting particle origins according to their size distribution at a UK rural site, J. Geophys. Res., 113, D07202, 10.1029/2007jd008562, 2008.

Cheung, H. C., Morawska, L., and Ristovski, Z. D.: Observation of new particle formation in subtropical urban environment, Atmos. Chem. Phys., 11, 3823-3833, 10.5194/acp-11-3823-2011, 2011.

Costabile, F., Birmili, W., Klose, S., Tuch, T., Wehner, B., Wiedensohler, A., Franck, U., König, K., and Sonntag, A.: Spatio-temporal variability and principal components of the particle number size distribution in an urban atmosphere, Atmos. Chem. Phys, 9, 3163-3195, 2009.

Covert, D., Wiedensohler, A., and Russell, L.: Particle Charging and Transmission Efficiencies of Aerosol Charge Neutralizes, Aerosol Science and Technology, 27, 206-214, 1997.

Dall'Osto, M., Beddows, D. C. S., Pey, J., Rodriguez, S., Alastuey, A., Harrison, R. M., and Querol, X.: Urban aerosol size distributions over the Mediterranean city of Barcelona, NE Spain, Atmos. Chem. Phys., 12, 10693-10707, 10.5194/acp-12-10693-2012, 2012.

Dunn, J. C.: Well-separated clusters and optimal fuzzy partitions, Journal of cybernetics, 4, 95-104, 1974.

Eilers, P. H., and Marx, B. D.: Flexible smoothing with B-splines and penalties, Statistical science, 89-102, 1996.

Frey, B. J., and Dueck, D.: Clustering by Passing Messages Between Data Points, Science, 315, 972-976, 10.1126/science.1136800, 2007.

Friend, A. J., Ayoko, G. A., Jayaratne, E. R., Jamriska, M., Hopke, P. K., and Morawska, L.: Source apportionment of ultrafine and fine particle concentrations in Brisbane, Australia, Environmental Science and Pollution Research, 19, 2942-2950, 2012.

Handl, J., Knowles, J., and Kell, D. B.: Computational cluster validation in post-genomic data analysis, Bioinformatics, 21, 3201-3212, 2005.

Harrison, R. M., and Yin, J.: Particulate matter in the atmosphere: which particle properties are important for its effects on health?, Science of The Total Environment, 249, 85-101, 10.1016/s0048-9697(99)00513-6, 2000.

Harrison, R. M., Beddows, D. C. S., and Dall'Osto, M.: PMF analysis of wide-range particle size spectra collected on a major highway, Environmental Science & Technology, 45, 5522-5528, 2011.

Hartigan, J. A., and Wong, M. A.: Algorithm AS 136: A K-Means Clustering Algorithm, Journal of the Royal Statistical Society. Series C (Applied Statistics), 28, 100-108, 10.2307/2346830, 1979.

Heintzenberg, J., Birmili, W., Otto, R., Andreae, M., Mayer, J., Chi, X., and Panov, A.: Aerosol particle number size distributions and particulate light absorption at the ZOTTO tall tower (Siberia), 2006–2009, Atmos. Chem. Phys, 11, 8703-8719, 2011.

Hinds, W. C.: Aerosol technology: properties, behavior, and measurement of airborne particles, Book, Whole, Wiley, New York, 1999.

Hussein, T., Puustinen, A., Aalto, P., Mäkelä, J., Hämeri, K., and Kulmala, M.: Urban aerosol number size distributions, Atmos. Chem. Phys., 4, 391-411, 2004.

Hussein, T., Dal Maso, M., PETÄJÄ, T., KOPONEN, I. K., PAATERO, P., AALTO, P. P., HÄMERI, K., and KULMALA, M.: Evaluation of an automatic algorithm for fitting the particle number size distributions, Boreal Environment Research, 10, 337-355, 2005.

Janhäll, S., M. Jonsson, Å., Molnár, P., A. Svensson, E., and Hallquist, M.: Size resolved traffic emission factors of submicrometer particles, Atmospheric Environment, 38, 4331-4340, 10.1016/j.atmosenv.2004.04.018, 2004.

Karlsson, M. N. A., and Martinsson, B. G.: Methods to measure and predict the transfer function size dependence of individual DMAs, Journal of Aerosol Science, 34, 603-625, 2003.

Kaufman, L., and Rousseeuw, P. J.: Finding Groups in Data: An Introduction to Cluster Analysis, Book, Whole, John Wiley & Sons, Inc, New York, 2009.

Kohonen, T.: Self-organizing maps, Book, Whole, Springer, New York, 2001.

Kulis, B., and Jordan, M. I.: Revisiting k-means: New algorithms via Bayesian nonparametrics, arXiv preprint arXiv:1111.0352, 2011.

Lohmann, U., and Feichter, J.: Global indirect aerosol effects: a review, Atmos. Chem. Phys., 5, 715-737, 2005.

Morawska, L., Bofinger, N. D., Kocis, L., and Nwankwoala, A.: Submicrometer and Supermicrometer Particles from Diesel Vehicle Emissions, Environmental Science & Technology, 32, 2033-2042, 10.1021/es970826+, 1998.

Morawska, L., Ristovski, Z., Jayaratne, E. R., Keogh, D. U., and Ling, X.: Ambient nano and ultrafine particles from motor vehicle emissions: Characteristics, ambient processing and implications on human exposure, Atmospheric Environment, 42, 8113-8138, 10.1016/j.atmosenv.2008.07.050, 2008.

Ntziachristos, L., Ning, Z., Geller, M. D., and Sioutas, C.: Particle Concentration and Characteristics near a Major Freeway with Heavy-Duty Diesel Traffic, Environmental Science & Technology, 41, 2223-2230, 10.1021/es062590s, 2007.

Pope, C. A., and Dockery, D. W.: Health Effects of Fine Particulate Air Pollution: Lines that Connect, Journal of the Air & Waste Management Association, 56, 709-742, 10.1080/10473289.2006.10464485, 2006.

Rousseeuw, P. J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Journal of computational and applied mathematics, 20, 53-65, 1987.

Salimi, F., Mazaheri, M., Clifford, S., Crilley, L. R., Laiman, R., and Morawska, L.: Spatial Variation of Particle Number Concentration in School Microscale Environments and Its Impact on Exposure Assessment, Environmental Science & Technology, 47, 5251-5258, 10.1021/es400041r, 2013.

Shen, X., Sun, J., Zhang, Y., Wehner, B., Nowak, A., Tuch, T., Zhang, X., Wang, T., Zhou, H., and Zhang, X.: First long-term study of particle number size distributions and new particle formation events of regional aerosol in the North China Plain, Atmos. Chem. Phys, 11, 1565-1580, 2011.

Stevens, B., and Feingold, G.: Untangling aerosol effects on clouds and precipitation in a buffered system, Nature, 461, 607-613, 2009.

Tunved, P., Ström, J., and Hansson, H. C.: An investigation of processes controlling the evolution of the boundary layer aerosol size distribution properties at the Swedish background station Aspvreten, Atmos. Chem. Phys., 4, 2581-2592, 10.5194/acp-4-2581-2004, 2004.

http://www.ilaqh.qut.edu.au/Misc/UPTECH%20Home.htm.

Wand, M.: Fast computation of multivariate kernel estimators, Journal of Computational and Graphical Statistics, 3, 433-445, 1994.

Wegner, T., Hussein, T., Hämeri, K., Vesala, T., Kulmala, M., and Weber, S.: Properties of aerosol signature size distributions in the urban environment as derived by cluster analysis, Atmospheric Environment, 61, 350-360, 10.1016/j.atmosenv.2012.07.048, 2012.

WHO: Air quality guidelines: global update 2005: particulate matter, ozone, nitrogen dioxide, and sulfur dioxide: global update 2005 : summary of risk assessment, Geneva : World Health Organization, 2006.

Wiedensohler, A., Birmili, W., Nowak, A., Sonntag, A., Weinhold, K., Merkel, M., Wehner, B., Tuch, T., Pfeifer, S., and Fiebig, M.: Mobility particle size spectrometers: harmonization of technical standards and data structure to facilitate high quality long-term observations of atmospheric particle number size distributions, Atmos, Meas. Tech, 5, 657-685, 2012.

Wood, S. N.: Thin plate regression splines, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65, 95-114, 2003.

Wraith, D., Alston, C., Mengersen, K., and Hussein, T.: Bayesian mixture model estimation of aerosol particle size distributions, Environmetrics, 22, 23-34, 2011.

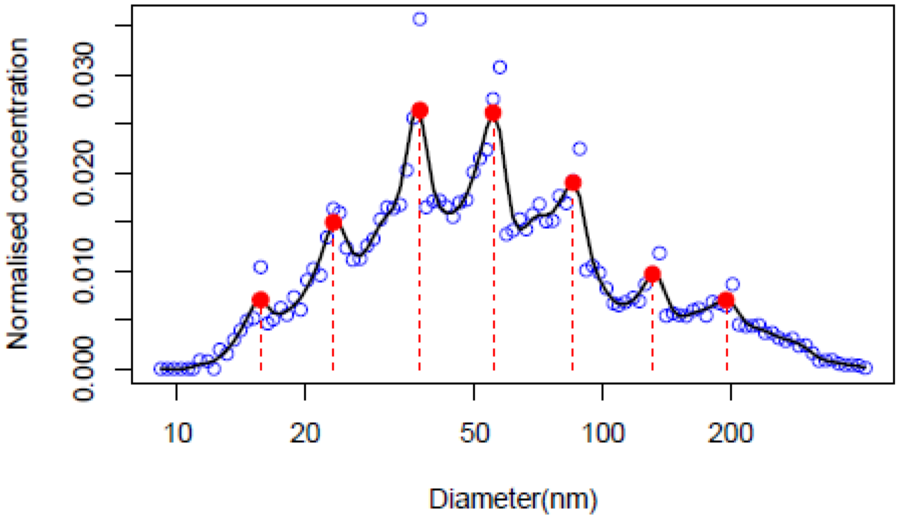

Figure 1: An example of a fitted smooth function (solid line) on PNSD data (circles) and the identified local peaks (solid circles).

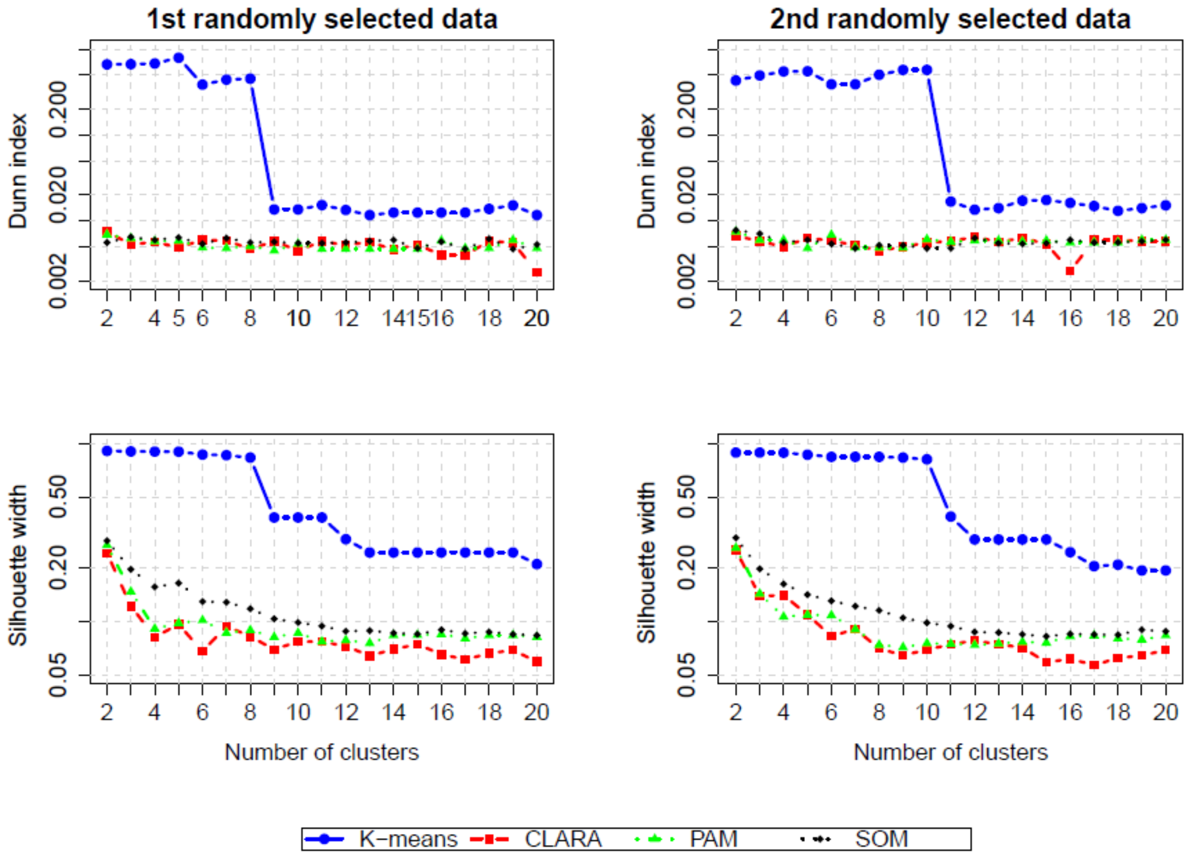

Figure 2: Dunn index and Silhouette width for Kmeans, SOM, PAM, and CLARA clustering techniques for different cluster numbers.

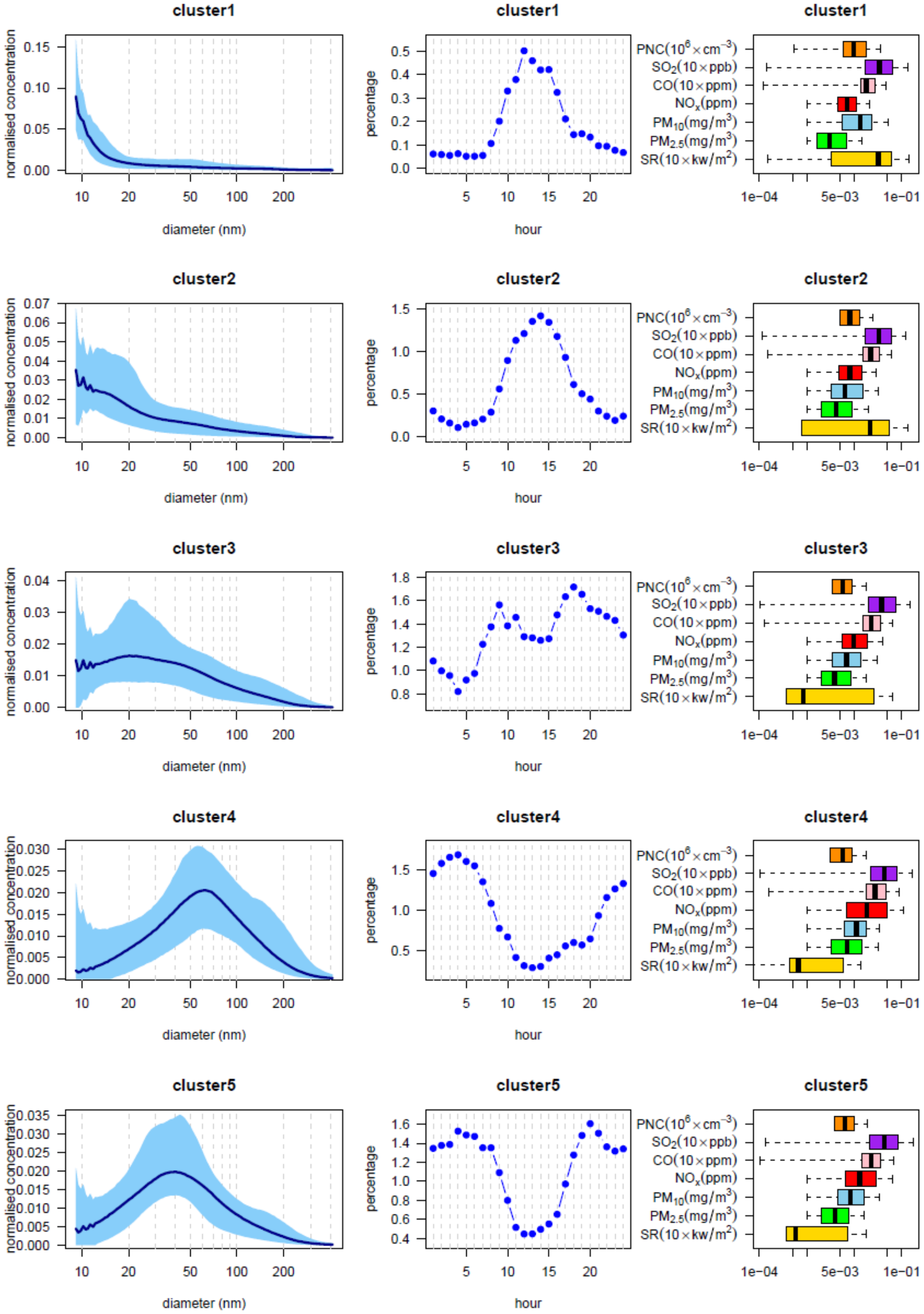
cluster1
cluster2
cluster3
cluster4
cluster5
normalised concentration
diameter (nm)
10
20
50
100
200
0.00
0.05
0.10
0.15
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
percentage
hour
5
10
15
20
0.0
0.1
0.2
0.3
0.4
0.5
1.0
1.5
0.4
0.6
0.8
1.2
1.4
1.6
1.8
PNC($10^6 \times cm^{-3}$)
SO$_2$($10 \times ppb$)
CO($10 \times ppm$)
NO$_x$(ppm)
PM$_{10}$(mg/m^3)
PM$_{2.5}$(mg/m^3)
SR($10 \times kw/m^2$)
1e−04
5e−03
1e−01

Figure 3: Clustering results using SMPS data. Graphs on the left show the particle size distributions with 95% confidence intervals, the associated graphs on the middle show the diurnal cycle of the hourly percentage of occurrence for each cluster, and the graphs on right show the associated solar radiation (SR), PM2.5,, PM10, NOx, CO, SO2, total PNC.

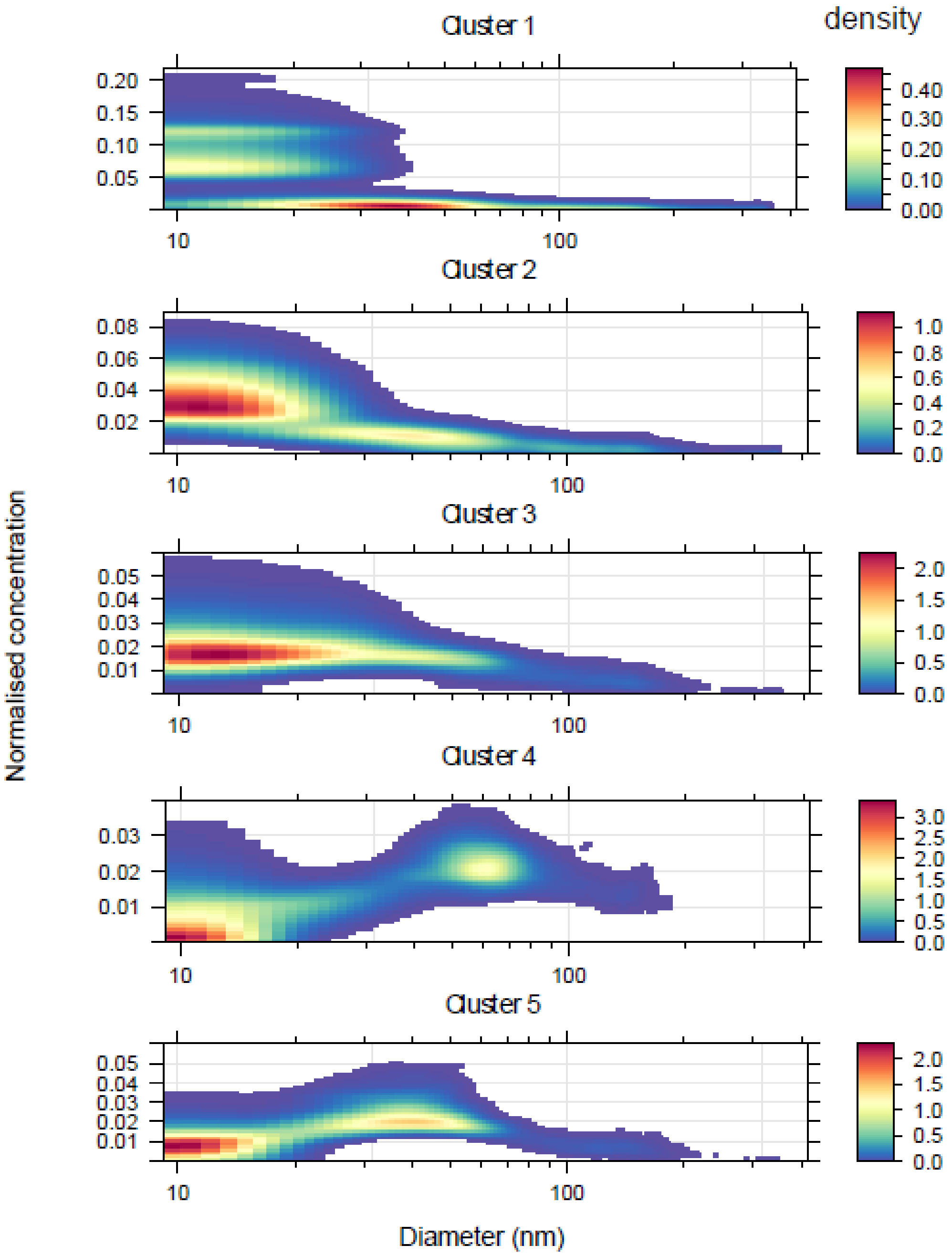

Figure 4: Density of peaks in particle number size data at each cluster.

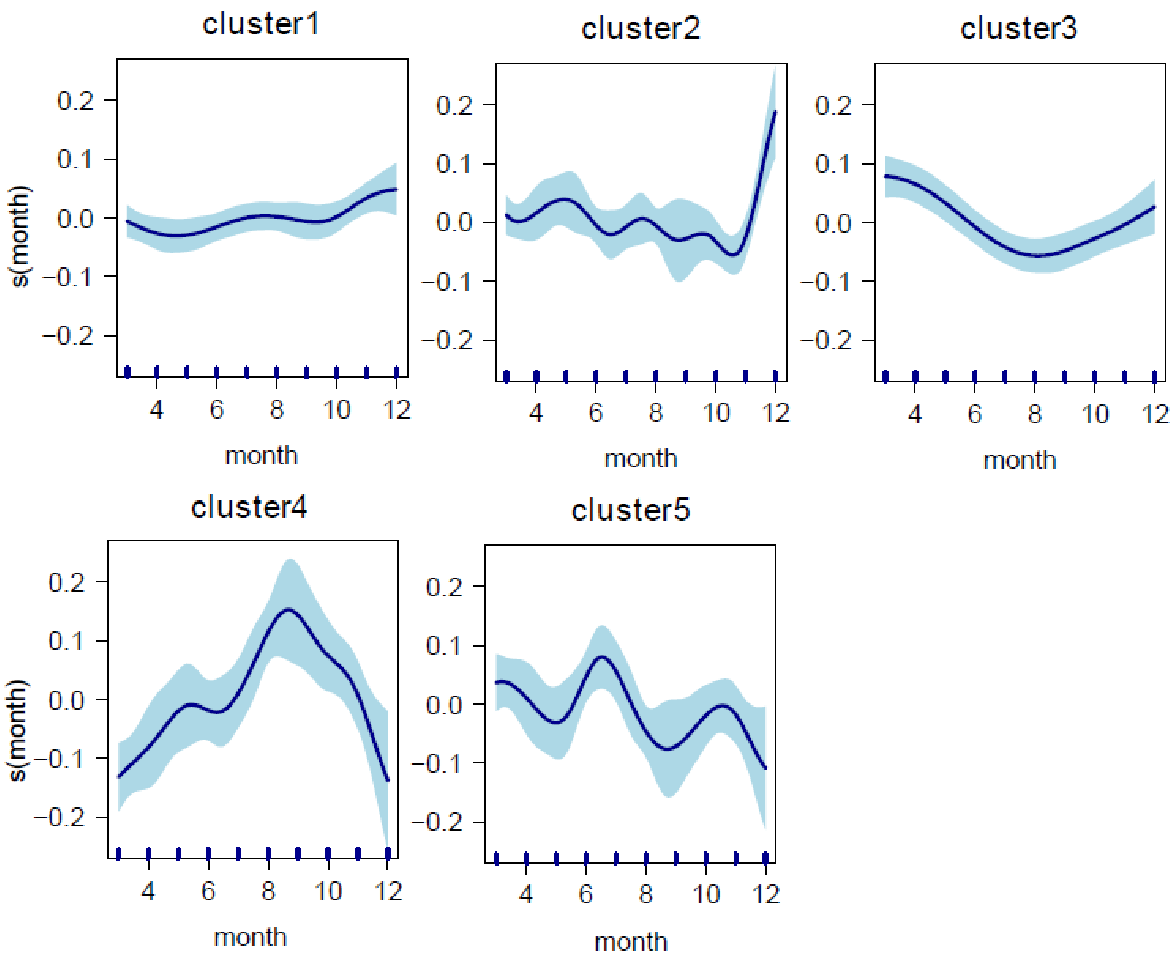

Figure 5: Temporal trend of occurrence of each cluster, with their 95% confidence intervals.

Table 1: Characteristics of Clusters

Cluster #	Occurrence percentage	Source/Origin
1	4.5%	Photochemically-induced nucleated particles (fresh)
2	14.1%	Photochemically-induced nucleated particles (relatively aged)
3	31.6%	Vehicle generated particles (primary and secondary)
4	22.6%	Regional background aerosols + biomass burning
5	27.2%	Regional background aerosols