Science of the Total Environment 744 (2020) 140923

Contents lists available at ScienceDirect

Science of the Total Environment

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

Efficient data preprocessing, episode classification, and source apportionment of particle number concentrations

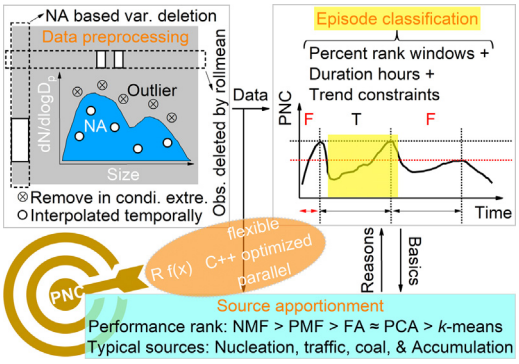
Chun-Sheng Liang [a,b,1], Hao Wu [d,g,1], Hai-Yan Li [a,f], Qiang Zhang [c], Zhanqing Li [e,*], Ke-Bin He [a,b,**]

[a] State Key Joint Laboratory of Environment Simulation and Pollution Control, School of Environment, Tsinghua University, Beijing 100084, China
[b] State Environmental Protection Key Laboratory of Sources and Control of Air Pollution Complex, Beijing 100084, China
[c] Ministry of Education Key Laboratory for Earth System Modeling, Department of Earth System Science, Tsinghua University, Beijing 100084, China
[d] College of Global Change and Earth System Science, Beijing Normal University, Beijing 100875, China
[e] Department of Atmospheric and Oceanic Science, University of Maryland, College Park, MD 20742, USA
[f] Institute for Atmospheric and Earth System Research/Physics, Faculty of Science, University of Helsinki, Helsinki 00014, Finland
[g] China Global Atmosphere Watch Baseline Observatory (WMO/GAW Station), Xining 810001, China

HIGHLIGHTS

- Auto-identification of consecutive NA via moving averages benefits interpolation.
- Point-by-point weighed outlier removal by conditional extremum saves non-outliers.
- Auto-division of episodes via threshold windows, durations, and trend constraints
- Performance rank of source apportionment models is NMF > PMF > FA ≈ PCA > *k*-means.
- Traffic kept dominant while coal heating decreased by 40%–86% over recent 5 years.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 31 March 2020
Received in revised form 7 July 2020
Accepted 10 July 2020
Available online 18 July 2020

Editor: Pavlos Kassomenos

Keywords:
Data preprocessing
Episode classification
Source apportionment
Particle pollution

ABSTRACT

Number concentration is an important index to measure atmospheric particle pollution. However, tailored methods for data preprocessing and characteristic and source analyses of particle number concentrations (PNC) are rare and interpreting the data is time-consuming and inefficient. In this method-oriented study, we develop and investigate some techniques via flexible conditions, C++ optimized algorithms, and parallel computing in R (an open source software for statistics and graphics) to tackle these challenges. The data preprocessing methods include deletions of variables and observations, outlier removal, and interpolation for missing values (NA). They do better in cleaning data and keeping samples and generate no new outliers after interpolation, compared with previous methods. Besides, automatic division of PNC pollution events based on relative values suites PNC properties and highlights the pollution characteristics related to sources and mechanisms. Additionally, basic functions of *k*-means clustering, Principal Component Analysis (PCA), Factor Analysis (FA), Positive Matrix Factorization (PMF), and a newly-introduced model NMF (Non-negative Matrix Factorization) were tested and compared in analyzing PNC sources. Only PMF and NMF can identify coal heating and produce more explicable

* Corresponding author.
** Correspondence to: K.-B. He, State Key Joint Laboratory of Environment Simulation and Pollution Control, School of Environment, Tsinghua University, Beijing 100084, China.
E-mail addresses: zli@atmos.umd.edu (Z. Li), hekb@tsinghua.edu.cn (K.-B. He).
[1]These authors contributed equally.

https://doi.org/10.1016/j.scitotenv.2020.140923

Number concentration

results, meanwhile NMF apportions more distinctly and runs 11–28 times faster than PMF. Traffic is interannually stable in non-heating periods and always dominant. Coal heating's contribution has decreased by 40%–86% in recent 5 heating periods, reflecting the effectiveness of coal burning control.

1. Introduction

Particle number concentrations (PNC) can be used in describing atmospheric particle number size distributions (PNSD) (Brines et al., 2015; Buseck and Adachi, 2008; Seinfeld and Pandis, 2006; Whitby, 1978), which have 4 modes: Nucleation (Dp ≤ 30 nm), Aitken (30 nm < Dp ≤ 100 nm), Accumulation (100 nm < Dp ≤ 1 μm) and Coarse (Dp > 1 μm) (Harrison et al., 2000; Hussein et al., 2005; Kulmala et al., 2004; von Bismarck-Osten et al., 2013; Vu et al., 2015). Nucleation and Aitken modes, namely ultrafine particles (UFP) whose mass is harder to measure than their number (Baldauf et al., 2016), dominate the total number and Accumulation and Coarse modes dominate the total surface and volume or mass (Seinfeld and Pandis, 2006). Unfortunately, the current mass-based regulations, standards, and monitoring systems fail in characterizing most particles in number (Brines et al., 2015; Liang et al., 2016a; Liang et al., 2016b).

PNSD are important to fully explain and assess particle pollution's impacts on human health, climate, and visibility. PNC-dominant small particles are more relevant to health effects due to their larger specific surface area per mass and stronger biological activity (Baldauf et al., 2016; Beaudrie et al., 2016; Meng et al., 2013; Oberdorster et al., 2005; Stanier et al., 2004; Wichmann and Peters, 2000), demonstrating the importance of PNC for health impact assessments (Buonanno et al., 2008; Kulkarni et al., 2011; Tan et al., 2014). Moreover, people's exposure to UFP has increased dramatically since the 20th century (Oberdorster et al., 2005), especially in Asia (Kumar et al., 2014). The United States Environmental Protection Agency (US EPA) (Baldauf et al., 2016), the Health Impact Institute (HEI) (Frampton et al., 2013), and the European Union (Bartczak and Goenaga-Infante, 2016; EU, 2012) have attached great importance to research and legislation on UFP and PNC. PNSD are crucial in cloud condensation nuclei (CCN) estimation (Dusek et al., 2006). Visibility is greatly affected by particle size and mainly related to big particles above 300 nm (Baumer et al., 2008; Chen et al., 2012; Kittelson, 1998; Sloane et al., 1991).

PNSD provide plentiful information on the sources and atmospheric processes of particles (Vu et al., 2015). Particles of different modes come from different sources or chemicophysical processes (Seinfeld and Pandis, 2006; Vu et al., 2015). In turn, the specific shapes and modal structures of PNSD contain valuable clues to sector sources, geographic origins, and particle formation mechanisms (Charron et al., 2008; Tunved et al., 2004; Vu et al., 2015). Some scientific and realistic fundamentals of source apportionment based on PNSD include: 1) PNSD have tempo-spatial quasi-stability (Kim et al., 2004; Mcelroy et al., 1982; Ogulei et al., 2006; Zhou et al., 2005; Zhou et al., 2004); 2) advanced measurements can capture short-term dramatic changes in PNSD (Hameri et al., 2004; Morawska et al., 1999; Weber et al., 2006; Wegner et al., 2012); 3) the particle size spans multiple orders of magnitude (Beddows et al., 2015) so the PNSD contain rich manifest (Charron et al., 2008; Hussein et al., 2014) and latent (Masiol et al., 2017a; Masiol et al., 2016) fingerprints (Morawska and Zhang, 2002) for distinguishing sources; 4) PNC have large spatial heterogeneity and are very sensitive to emission sources (Lianou et al., 2007; Price et al., 2014; Rodriguez et al., 2007; Weber et al., 2013); 5) source apportionments of PNC are indispensable to know the source contributions of PNC (Ogulei et al., 2007); 6) number concentration supplements mass concentration (Pey et al., 2010); 7) source apportionments of PNC are cost-effective (Thimmaiah et al., 2009; Yue et al., 2008).

Despite the extreme significance of PNSD in the above-mentioned impact assessment and source apportionment, knowledge on PNC pollution episodes and the source analysis based on PNC are still infant because of lacking standards (Chen et al., 2017; Chen et al., 2018; Mertens et al., 2020; Trojanowski and Fthenakis, 2019), insufficient studies, and complex data. For example, preprocessing of PNC data is much trickier than that of mass-based parameters due to the complexity of often measured PNC variables of around 100 (16–158) size bins (Friend et al., 2012; Vu et al., 2015), which are many more than often measured chemical species (Liang et al., 2016a). Researchers preprocessed their raw data by simply deleting abnormal (wrong) observations (Wiedensohler et al., 2012) and replacing missing data by mean values (Gu et al., 2011) or temporal linear interpolation (Liu et al., 2014). Masiol et al. (2017a) and Masiol et al. (2016) further stated using the top 0.5% percentile to remove outliers and nearest size bins for NA interpolation. Except new particle formation (NPF) events (Carnerero et al., 2019; Dal Maso et al., 2005; Kulmala et al., 2004; Sun et al., 2016; Wu et al., 2007), there are few descriptions of PNC pollution events (Rodriguez et al., 2007) and their classification methods. NPF event classifications are manual (Carnerero et al., 2019; Hussein et al., 2020; Sun et al., 2016), semi-automatic (Gross et al., 2018; Heintzenberg et al., 2007), or computation-consuming (Joutsensaari et al., 2018). PNC source analysis techniques mainly include receptor models Factor Analysis (FA) (Wåhlin et al., 2001), Positive Matrix Factorization (PMF) (Ogulei et al., 2006; Zhou et al., 2004), Principal Components Analysis (PCA) (Khan et al., 2015; Liang et al., 2013; Pey et al., 2009), *k*-means clustering (Beddows et al., 2009), and source models Potential Source Contribution Function (PSCF) and Concentration Weighted Trajectory (CWT) (Bycenkiene et al., 2014; Bycenkiene et al., 2013). In practical use, the receptor models based on extracted fingerprints are often combined with associated pollutants, time patterns, meteorology, and source models (Beddows et al., 2015; Harrison et al., 2011; Masiol et al., 2017a).

Yet, it is unclear that if the traditional percentile-based method of outlier removal needs improvement and, if yes, how to improve, which is more reliable for interpolation: in adjacent size bins or time series, and what else needs to be cared in data preprocessing for PNC. Furthermore, it is unknown how to define and automatically divide PNC pollution episodes and how to efficiently calculate geometric mean diameter (GMD) and count median diameter (CMD). Besides, it remains unresolved which source apportionment model is the most suitable and if there is any new and more suitable model.

We try to solve these problems by using R (The R Core Team, 2019), based on PNC observations in Beijing during 2015–2019. First we develop a series of data preprocessing methods to prepare cleaned and reliable data for research on characteristics, sources, and mechanisms of PNC pollution. Then an automatic episode classification of PNC and simple algorithms of GMD and CMD are designed and introduced. Moreover, we establish source apportionment method by comparing five different techniques *k*-means clustering (Hartigan and Wong, 1979), PCA (Pearson, 1901), FA (Spearman, 1904), PMF (Paatero and Tapper, 1994), and NMF (Non-negative Matrix Factorization) (Lee and Seung, 1999). Afterwards we apply the established method to different cases and test its ability in identifying typical sources. In addition, the features of the typical sources are examined and the actual impact of coal heating is assessed. In data quality assurance, pollution characteristics analysis, and source analysis of PNC, the series of methods exhibit more robust, smarter, more distinct, and faster performance. These efficient analysis methods for PNC pollution characteristics and sources can promote the establishment and actual applications of PNC source apportionment and provide multi-faceted support for the research and control of PNC pollution.

2. Materials and methods

2.1. Sampling of particles

Three sampling sites are located at Tsinghua University (THU, lat = 40.003, lon = 116.32) (H.Y. Li et al., 2019; Li et al., 2016), Institute of Atmospheric Physics, Chinese Academy of Sciences (IAP, lat = 39.974, lon = 116.372) (Du et al., 2017; Wang et al., 2019), and Nanjiao Meteorological Bureau (NJ, lat = 39.807, lon = 116.471, namely Meteorological Observatory in the southern suburb of Beijing) respectively (Fig. S1). There are more mountains in the northwest and more main roads in the southeast. All these sites are within central Beijing of 5 rings and have no major industrial sources nearby. THU and IAP housed in Haidian District are similar ordinary sites with moderate traffic volumes and the distance between them is only 5 km. However, NJ housed in Daxing District which is an important transportation node for Beijing-Tianjin-Hebei (Jing-jin-Ji) integration is surrounded by dense traffic. Many large trucks are only allowed to pass nearby during 23:00–6:00 at nights. Moreover, NJ is right beside the main road South Fifth Ring Road.

Scanning mobility particle sizers (SMPS, TSI, 3938) were deployed, coupled with condensation particle counters (CPC, TSI, 3772, except 3787 from Sep., 2018 and 3787 and 3756 from Jan. 2019 in NJ) and differential mobility analyzers (DMA, TSI, 3081, except 3081 and 3085 from Jan. 2019 in NJ), for observing PNSD at ground levels, successively in the three sites (Table S1). The channel and time resolutions are all 64 per decade of particle size and 5-min respectively. A total of 259,604 observations of 102 size bins in the 14.1–532.8 nm range involving 33 months from 2015 to 2019 were synthesized as raw data.

2.2. Auxiliary data

The auxiliary data of PNSD here mainly consist of criteria pollutants (CP), meteorological parameters (Met), and backward trajectories (Traj) (Table S2). The distances between sites of PNSD and their nearest sites of CP are around 2 km. There is around 25 km from the Met site to each of the PNSD sites THU, IAP, and NJ.

2.3. Data handling tools

The R and its self-contained functions (The R Core Team, 2019), multiple packages provided by other developers, and EPA PMF 5.0 were used to handle the data (Norris et al., 2014; Paatero, 1997; Paatero and Tapper, 1994). We selected and used about 30 high-performance R packages (in user library only, excluding those in system library of R) to develop the methods in this work (Table S3) (Bache and Wickham, 2014; Carslaw, 2019; Carslaw and Ropkins, 2019; Carslaw and Ropkins, 2012; Cheng et al., 2019; Demin, 2019; Dowle and Srinivasan, 2019; Fellows and Stotz, 2019; Gaujoux and Seoighe, 2010; Gaujoux and Seoighe, 2018; Grolemond and Wickman, 2011; Iannone, 2018; Müller and Wickham, 2019; Revelle, 2020; Robinson and Hayes, 2019; Sarkar, 2008; Sarkar, 2018; Slowikowski, 2019; Spinu et al., 2018; Ushey, 2018; Wickham, 2007; Wickham, 2016; Wickham, 2017; Wickham, 2019a; Wickham, 2019b; Wickham et al., 2019a; Wickham et al., 2019b; Wickham and Henry, 2019; Wickham and Seidel, 2019; Wilke, 2019; Yu, 2019; Zeileis and Grothendieck, 2005; Zeileis et al., 2019).

2.4. Definition and classification of PNC pollution episodes

This work defines PNC pollution episodes as pollution events with high PNC in a certain mode. For example, 3 kinds of episodes: Nucleation (Dp ≤ 30 nm) episodes, Aitken (30 nm < Dp ≤ 100 nm) episodes, and Accumulation (100 nm < Dp ≤ 1 μm) episodes are included here. Percentile rank windows, minimum duration hours, and constraints of trends were used as thresholds and conditions to classify the episodes. In order to stably classify pollution episodes in different periods and sites, we temporarily calculate ('borrow') 3-hour moving averages with tidyr::complete (Wickham and Henry, 2019) and C++ optimized RcppRoll::roll_mean (Ushey, 2018), normalize the moving averages by percentile rank for convenient comparison among different sites, and set three requirements (conditions) to detect and grab rising trends. The first requirement for a pollution episode is the percentile rank normalized moving averages (PRNMA) change from below 20 to above 80 and last for over 4, 8, and 16 h for Nucleation, Aitken, and Accumulation respectively. The second major requirement is there are PRNMA <20 in the first half and >80 in the second half. The third major requirement includes: in time-series order, the PRNMA in the second half, middle half, and 4th percentile interval are 50% greater than those in the first half, 1st percentile interval, and middle half respectively.

2.5. Principles and evaluation criteria of source apportionment receptor models

Four previously reported receptor models *k*-means clustering (Beddows et al., 2009), PCA (Khan et al., 2015; Liang et al., 2013; Pey et al., 2009), FA (Wåhlin et al., 2001), and PMF (Ogulei et al., 2006; Zhou et al., 2004) and one newly introduced receptor model NMF (Gaujoux and Seoighe, 2010; Gaujoux and Seoighe, 2018; Lee and Seung, 1999) are used for source apportionment of PNC in this study.

2.5.1. k-means clustering

$$ss(k) = \sum_{i=1}^{n} \sum_{j=0}^{p} \left(x_{ij} - \overline{x}_{kj}\right)^2 \tag{1}$$

where x_{ij} is the value of the j-th variable in the i-th observation, $\overline{x}_{kj}$ is the mean value of the j-th variable in the k-th cluster, n is the number of observations, p is the number of variables, and ss is the sum of squares (Forgy, 1965; Hartigan and Wong, 1979; Kabacoff, 2015; Lloyd, 1982; MacQueen, 1967; Morissette and Chartier, 2013).

2.5.2. PCA

$$PC_i = a_{i1}V_1 + a_{i2}V_2 + +a_{in}V_n \tag{2}$$

where PC_i is the i-th principal component, V_j is the j-th observation variable, and a_{ij} is the load, that is, the linear correlation coefficient between PC_i and V_j (Hotelling, 1933; Pearson, 1901; Rao, 1964; Statheropoulos et al., 1998). Principal components are linear combinations of observed variables and are obtained by maximizing the variance explained by each principal component (Kabacoff, 2015).

2.5.3. FA

$$X_i = a_1 F_1 + a_2 F_2 + \cdots a_p F_p + U_i \tag{3}$$

where X_i is the ith observable variable ($i = 1 \ldots k$), F_j is the common factor ($j = 1 \ldots p$), and $p < k$. Ui is a unique part of the Xi variable (cannot be explained by common factors) (Bartholomew, 1995; Kabacoff, 2015; Spearman, 1904). a_i can be considered as the contribution value of each factor to the composite observable variable.

2.5.4. PMF

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{m} \left[\frac{E}{U}\right]^2 \tag{4}$$

or

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{m} \left[\frac{x_{ij} - \sum_{k=1}^{p} g_{ik} f_{kj}}{u_{ij}}\right]^2 \tag{5}$$

where U (u_{ij}) is the uncertainty of the PNC of particle size bin j in sample i (Kim et al., 2004), E is the residual error (Krecl et al., 2015), n is the total number of samples, and m is the total number of particle size bins (Krecl et al., 2008). For a given number of factors, PMF (Paatero and Tapper, 1994) adjusts the values of f_{kj} and g_{ik} by the method of least squares (with the constraint that f_{kj} and g_{ik} values are non-negative) until the minimum Q value is found (Krecl et al., 2008; Norris et al., 2014).

2.5.5. NMF

$$X \approx WH \tag{6}$$

where X is the M-variable and N-observation matrix (Lee and Seung, 1999). The goal of NMF is to find the non-negative M × L base matrix W and the L × N coefficient matrix H so that X ≈ WH (Gaujoux and Seoighe, 2010; Gaujoux and Seoighe, 2018). Because EPA PMF lacks alternative matrix factorization algorithms and uses serial computing that is slow, NMF that has multiple factorization algorithms and uses parallel computing is introduced.

The performance of the basic functions of these models will be tested and compared in this work. The evaluation criteria include: 1) Discrimination of concentrations; 2) Discrimination of peak particle sizes; 3) Ratios of species; 4) Contribution moment.

To more conveniently and efficiently compare the impact of parameters such as wind and trajectory, contribution moment (CM) is put forward. CM is the value of a normalized parameter with an average value of 1 multiplied by a normalized contribution with an average value of 1.

$$\mathrm{CM_{Parm}} = \mathrm{Parm_1^{Mean}} \times \mathrm{C_1^{Mean}} \tag{7}$$

where Parm is the parameter, $\mathrm{Parm_1^{Mean}}$ is the normalized parameter with an average value of 1, and $\mathrm{C_1^{Mean}}$ is the normalized contribution with an average value of 1. It is expressed as CM followed by the subscript or lowercase letters of the parameter, such as the wind speed (WS) contribution moment $\mathrm{CM_{WS}}$, the trajectory pressure contribution moment CMtp, and the trajectory height contribution moment CMth.

2.6. Source profiles

The general law of the sources of PNC is summarized in Table 1 (details in Table S4) based on the 30 results (29 publications): 24 of PMF (Al-Dabbous and Kumar, 2015; Beddows et al., 2015; Dall'Osto et al., 2012; Friend et al., 2013; Friend et al., 2012; Gu et al., 2011; Harrison et al., 2011; Kasumba et al., 2009; Krecl et al., 2008; Krecl et al., 2015; Z. Liu et al., 2017; Liu et al., 2014; Liu et al., 2016; Masiol et al., 2017a; Masiol et al., 2016; Ogulei et al., 2007; Sowlat et al., 2016; Squizzato et al., 2019; Thimmaiah et al., 2009; Wang et al., 2013; Yue et al., 2008; Zhou et al., 2005; Zhou et al., 2004; Zong et al., 2019); 4 of k-means clustering (Charron et al., 2008; Dall'Osto et al., 2012; Hussein et al., 2014; Wegner et al., 2012); 2 of PCA (Cusack et al., 2013; Pey et al., 2009).

3. Results and discussion

3.1. Data preprocessing

3.1.1. Dispersedness and NA of raw data

The exemplary original data of THU in January and July of 2015 are very dispersed but outliers are not too many (Fig. 1a, b, and d). The hidden distribution of the low-value area with ordinary coordinates (Fig. 1a) can be seen in logarithmic scales (Fig. 1b). Inspired by the previous work (Masiol et al., 2017a; Masiol et al., 2016) which excluded the highest 0.5% of values and because there are outliers in the bottom (Fig. 1c and d), we deleted the 0.5% of values both at the head and tail (Fig. 1c). However, such an outlier removal method simply based on the percentiles would delete many values that are not outliers (Fig. 1d).

The basic descriptive statistics (Fig. S2) show that NA are mainly distributed at both ends of size bins, especially small size bins.

3.1.2. Variable deletion

To avoid information distortion (Fig. S2), unqualified size bin variables with too many NA were removed (Fig. S3), which are the first 4 small bins 14.1, 14.6, 15.1, and 15.7 nm. These NA originate from instruments (SMPS). If these 4 variables are not removed first and the observations with too many consecutive NA in time series are directly deleted, there will be only 141,617 observations left. However, if we delete these 4 variables and then delete the observations with too many NA in time series, we can retain 19,640 (= 161,257–141,617) more observations.

3.1.3. Observation deletion

Observations with at least one variable containing too many consecutive NA in time series are not reliable (Fig. S3), which cannot be properly interpolated temporally because of producing new outliers, need to be deleted. The observations were deleted based on consecutive NA.

To delete observations based on consecutive NA, it is necessary to remove observations that have incomplete half-hour averages in all size bins, namely observations with any bin(s) containing consecutive NA for more than half an hour. We achieve it by ingeniously 'borrowing' hourly moving averages around a center of two neighboring half-hour sides to capture the cases of continuous NA for more than half an hour (Figs. S4 and S5 and KC S1) (KC: Key codes). The key points are: temporarily (recovery afterwards) expand the complete time period with tidyr::complete (Wickham and Henry, 2019) and calculate hourly moving averages with C++ optimized RcppRoll::roll_mean (Ushey, 2018) (much faster than ordinary functions) to detect consecutive NA. It is called 'borrowing' because instead of calculating the moving averages on the columns of the original data, it calculates the moving averages of the original data temporarily in columns other than the original data. Based on the completeness of the calculated moving averages, whether the original data is missing consecutively for more than half an hour (no matter one size bin or multiple size bins) is logically judged. The centered 59-min moving average is not used to judge one side but both sides that are missing continuously for more than half an hour, namely whether there is any effective value within half an hour either before or after a point for interpolation. This method continuously applies to the deletion of outliers that may cause such observations, ensuring the rationality of NA interpolation in time series. Grouping by year and month and 'borrowing' hourly moving averages correct and improve previous time series interpolation that did not consider time span and easily brought new outliers.

3.1.4. Outlier removal

As can be seen from Figs. S4 and S5, after deleting the observations with incomplete half-hour averages in all size bins, there are still outliers that need to be removed. Outlier removal consists of traditional percentile removal (Masiol et al., 2017a; Masiol et al., 2016) (Figs. S6 and S7) and conditional extremum removal (Figs. S8 and S9 and KC S2). The percentile of the top is 0.5% which is the same as the reports (Masiol et al., 2017a; Masiol et al., 2016), while the bottom is 0.25% because it is not so dispersed nor influential as the top. The difference in concentrations among different months can be relatively large. If the outliers are removed together without being grouped by month, there will be overdone removals and unclean removals in different months, so the removals are grouped by month. During the removal, observations with incomplete half-hour averages in all size bins were deleted.

Advantages of conditional extremum removal (Figs. S8–S10) over traditional percentile removal (Figs. S6 and S7) include: 1) Reduced number of removed values that are not necessary to be removed: 210819–159,812 = 51,007; 2) Increased number of retained observations that should be kept: 159732–159,485 = 247; 3) Cleaner data after removal, especially at the tops in January and July 2015 and January

Table 1
General PNSD source profiles.

Sources	Peaks (nm)	Associated pollutants	Diurnal patterns
Nucleation	<20	O_3	Daytime
Fresh traffic	<50	NO_2	Morning/evening rush hour
Aged traffic	30–100	NO_2	Morning/evening rush hour
Coal heating	100–200 (or 300)	SO_2	Nighttime
Regional accumulation/secondary	>200 (or 300)	$PM_{2.5}$	Nighttime

2017 (Figs. S7, S9, and S10). These differences are mainly because the percentile based method is "one size fits all" and does not judge whether the values may be deleted excessively or insufficiently; while the conditional extremum based method considers point by point and judges each potentially to-be-deleted value whether it is really an outlier that should be deleted. The proportion of identified outliers from both the top and the bottom is very limited in PNC data, namely (159812–126,060) / (159,732 * 98) ≈ 0.2% (0.2156%), about 56.9% (= (0.5–0.2156) / 0.5) less than the previously proposed top-only 0.5% (Masiol et al., 2017a; Masiol et al., 2016). The conditional extremum based method may also fit well with other huge datasets by adjusting constraints and times (KC S2) in batch processing for studies like monthly, seasonal, (non-) heating-period, or yearly PNC source apportionments. Outliers should be removed with more caution or even not be removed for individual-episode studies or short-term observations.

3.1.5. NA interpolation

The reasons why the nearest bin interpolation (Masiol et al., 2016) should be discarded include: 1) Simple nearest bin interpolation produces many new outliers (Fig. S11); 2) Many new outliers are still generated after improving nearest bin interpolation: neighboring size bin based NA interpolation (Fig. S12) and middle (as starting point) neighboring size bin based NA interpolation (Fig. S13); 3) The number of observations is also reduced by hundreds (462 = 159,732–159,270, Figs. S10 and S11) or even more (4189 = 159,732–155,543 and 1322 = 159,732–158,410, Figs. S10, S12, and S13).

Different from the nearest bin interpolation that generates new outliers and retains fewer observations (Masiol et al., 2016), temporal interpolation is used here (Fig. 2): 1) No new outliers are generated. This is mainly due to the key steps in the aforementioned process of deleting observations and outliers: observation deletion based on hourly moving averages; and the use of time period grouping during interpolation: dividing the time between missing observations up to half an hour into periods, and group interpolation by the time periods. This makes the cleaned data frame without any variable that is missing for half an hour or more and confines the interpolation to a relatively continuous period of time, ensuring the robustness and rationality of time series interpolation. It's different from simple time series interpolation that does not consider consecutive NA, time span, and period grouping. 2) Maintain number of observations as the original cleaned data contains.

3.2. Episode classification

3.2.1. PNSD and temporal variation

The 5-min concentration distribution and cumulative mean of sorted number concentrations in each bin and temporal (hourly) variation of size distribution can be seen from Figs. S14 and S15. A common feature of the three different sites THU, IAP, and NJ within the Beijing

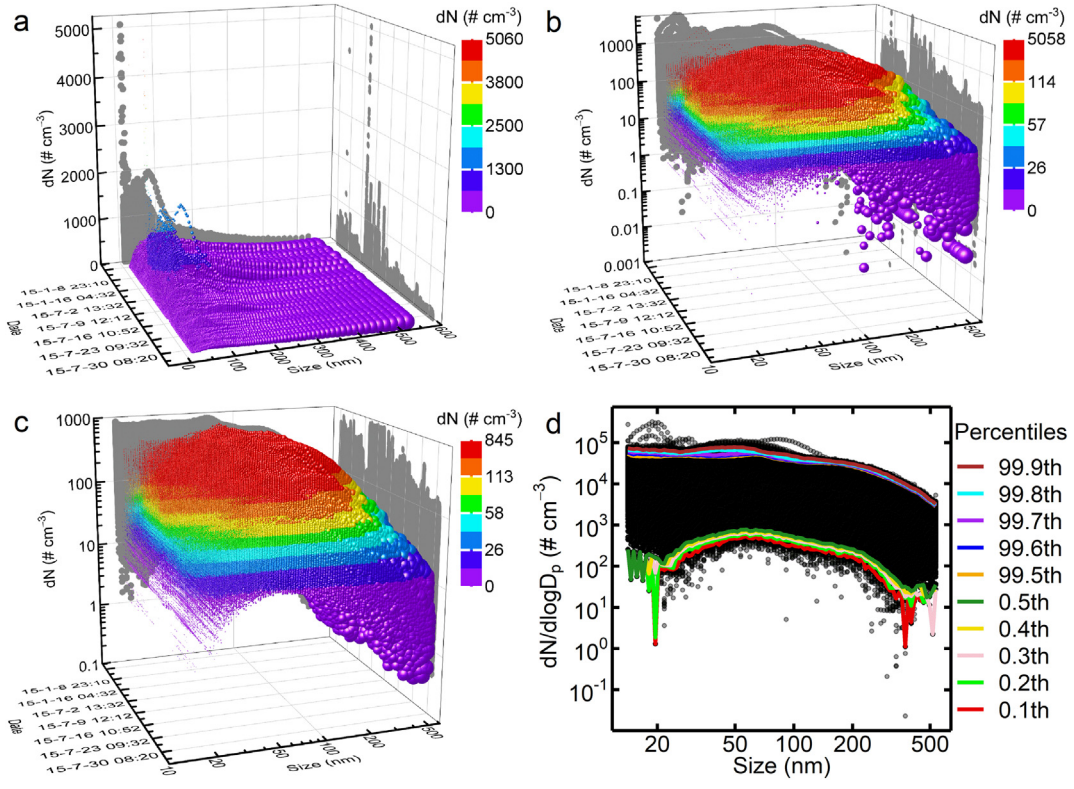


Fig. 1. 3D and logarithm normalized PNSD of THU in January and July of 2015.

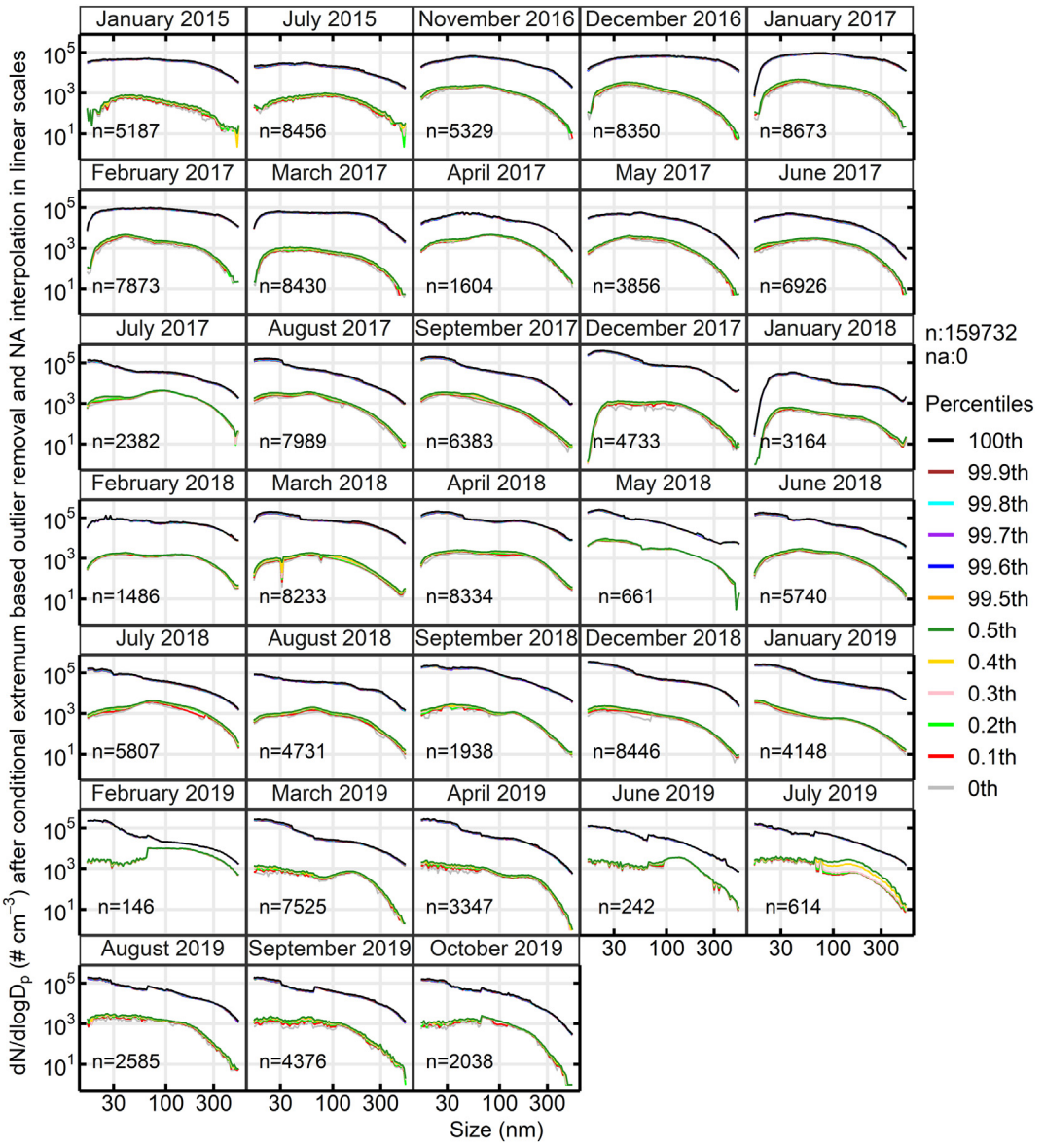


Fig. 2. Top and bottom percentiles after conditional extremum based outlier removal and NA interpolation in time series in linear y-axis scales.

Fifth-Ring Road is that the concentrations of particles above 100 nm do not exceed 10*10^4. However, the concentrations of particles below 100 nm (UFP) are very different between ordinary sites (THU and IAP) and transportation sites (NJ).

Fig. 3 also shows the difference of UFP between ordinary sites (THU and IAP) and transportation site (NJ). The proportion of UFP in NJ is 13%–15% higher than those in THU and IAP (Fig. 3a). This is mainly caused by the distinction of particles below 50 nm. Besides the ordinary traffic peaks at around 7:00 and 18:00, there is a more obvious high-value platform in NJ during 23:00–6:00 (Fig. 3b) when a large number of big trucks are allowed to pass nearby. The calculation methods of GMD and CMD (KC S3) are very efficient and useful especially after data preprocessing the original GMD and CMD from the instruments need to be recalculated. These methods may also find their applications in NPF research besides the study of ordinary pollution characteristics here.

3.2.2. Pollution episodes

Details of automatically classifying PNC pollution episodes are attached in KC S4. Figs. S16–S18 show the temporal variations of the pollution episodes. Applications can be extended by adjusting the percentile rank window and minimum duration hours. For example, more episodes are expected to be classified when widening the window from 20/80 to 30/70 or lowering the minimum duration.

Nucleation episodes didn't occur in THU and occurred only once in IAP, while almost all this kind of episodes were observed only in NJ (Fig. S16). The PNC of Nucleation mode are very high in the transportation site (NJ), hence NJ owned almost all the Nucleation episodes. Peaks of Nucleation episodes appeared at both afternoons and midnights. Afternoon Nucleation episode peaks indicate the impact of photochemical Nucleation process (Tan et al., 2018) that is regarded as a source, while midnight Nucleation episode peaks indicate the impact of traffic source.

Peaks of Aitken episodes appeared at nights and afternoons (Fig. S17), indicating the impact of particle growth (Liu et al., 2008; Meier et al., 2009). This kind of episodes occurred only in IAP and NJ but didn't occur in THU. It is mainly because among the three sites, the number of observations is the most limited and the distance from the main road is farthest in THU.

Peaks of Accumulation episodes appeared at nights and mornings (Fig. S18). Accumulation episodes occurred in all sites, indicating

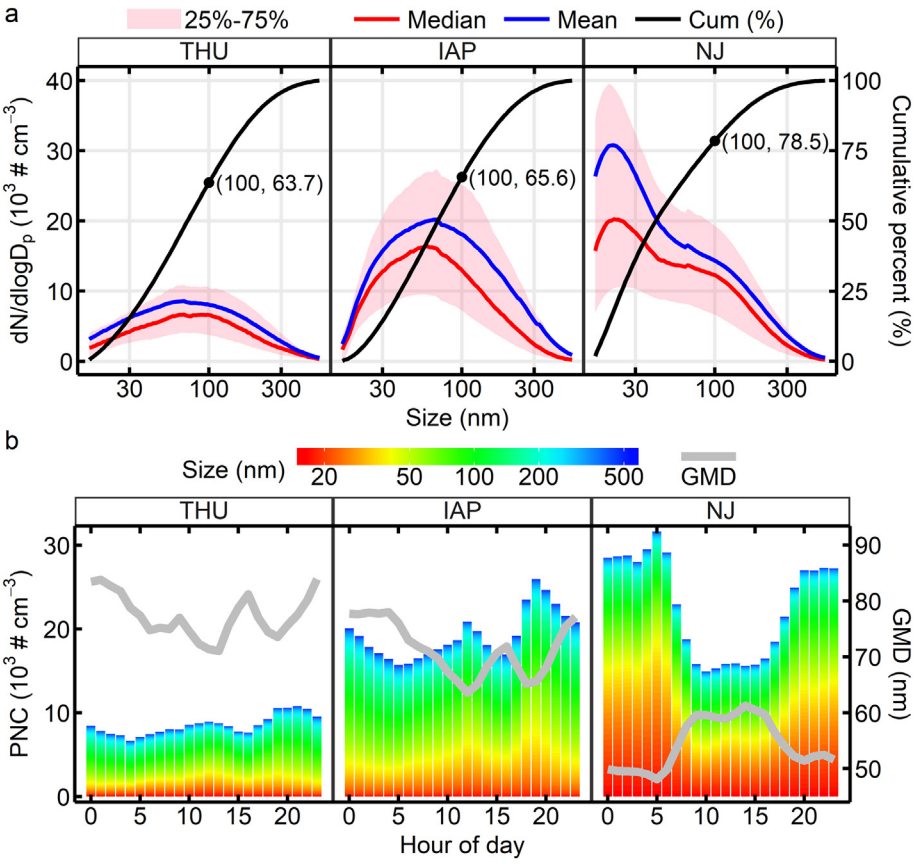


Fig. 3. Averaged and cumulative hourly PNSD and diurnal variation of size-segregated PNC and GMD. Calculation methods of GMD and CMD are achieved in R and can be found in KC S3.

regional impact. Night and morning Accumulation episode peaks indicate the impact of RH (Zhao et al., 2020). Accumulation episodes occurred once in THU (with least observations among the three sites) and occurred more in IAP (with medium number of observations) than in NJ (with most observations), therefore recent cleaner atmosphere corresponded to less Accumulation episodes than previous dirtier atmosphere did. The automatic episode classification makes pollution characteristics clearer and offers basics to infer sources.

There were 25 Nucleation episodes, 23 Aitken episodes, and 11 Accumulation episodes (Fig. S19a). And these episodes lasted for around 11, 13, and 24 h respectively. Since the thresholds and constraints are set strictly here, the numbers of episodes are small. Numbers of episodes can be changed by adjusting the percent rank windows, the minimum duration hours, and trend constraints (KC S4). For example, if the window is expanded from 20/80 to 30/70 and the minimum duration hours are reduced from 4, 8, and 16 h to 3, 6, and 12 h, more (46 Nucleation, 51 Aitken, and 23 Accumulation) episodes will be classified, almost doubled (Fig. S19b).

Fig. 4 summarizes the characteristics of episodes in combination with criteria pollutants and meteorology that would be selected as a driving variable by their respective percentile rank maxima within a set. The colors only represent the PNC levels of a certain mode and should be compared inside individual modes. We cut hours into intervals that are open on the left and closed on the right, specifically 5:00–10:00 (namely (5:00, 10:00]) and 15:00–20:00 are morning and evening rush hours, 10:00–15:00 are midday, and 23:00–5:00 are nighttime. The Nucleation episode mainly occurred at night at the traffic site (NJ). The highest PNC level corresponds to NO_2, which indicates that there is a synergistic relationship with NO_2. This type of pollution event is mainly caused by fresh traffic (Z. Liu et al., 2017; Liu et al., 2016; Rivas et al., 2020) at short distance (a few line sources, Fig. S1) at night. The Aitken episode is similar to the nucleation mode in terms of main geographical location (NJ) and synergistic pollutant (NO_2), but the highest PNC level appeared mainly at rush hours. It's related to urban commuting, namely a large number of morning and evening traffic line sources constitute an approximate area source. Compared with a small number of line sources (fresher particles) that allow large trucks to pass near NJ during nights, the area source is farther on average, so the impact is mainly Aitken (older than Nucleation) mode. The Accumulation episode is different from both the Nucleation and Aitken episodes, and mainly occurred at ordinary sites (IAP). Its high PNC levels only appeared at night, indicating its dependence on RH (Figs. 4 and S18).

The robustness in data preprocessing and automation in episode classification are achieved by designing flexible and fastidious conditions (constraints) such as period grouping and extending, moving average borrowing, and value normalization and comparison (one by one in conditional extremum outlier removal and interval by interval in episode classification) and by using efficient calculating methods such as C++ optimized algorithms.

3.3. Establishment of source apportionment method

3.3.1. Case division

We divided a certain annual period into a heating period (HP) and a non-heating period (NHP) for source apportionment (Table S5). HP is from November 15 of the previous year to March 15 of the current year. NHP is from August of the previous year to July of the current year excluding HP. The observations of THU, IAP, and NJ involve years from 2015 to 2019 (12,966 h after the 13,555-hour PNSD merge with criteria pollutants and meteorology) that are divided into 8 cases (12,033 h): 4 HPs and 4 NHPs. There were 933 h that were not classified into any cases in NJ due to inconsistent sites and time range issues.

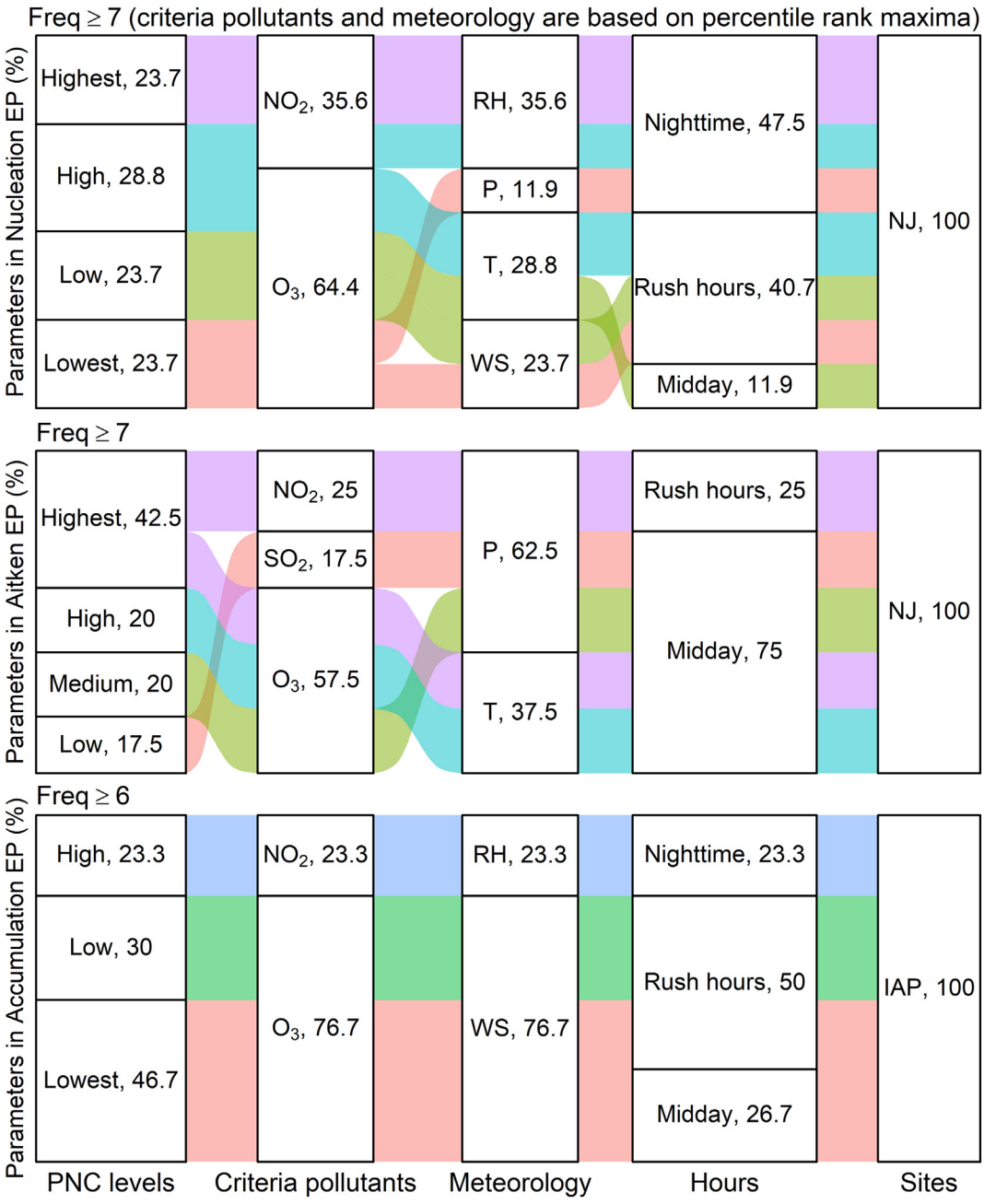


Fig. 4. Episodic parallel sets plot of modes, criteria pollutants, and meteorology.

3.3.2. Number of factors and declarations

To test and compare the performance of the basic functions of different models, considering the actual situation (three previous studies included 8 factors for each case in Beijing, but there are repeated or duplicate sources in the 8 factors) (Z. Liu et al., 2017; Liu et al., 2014; Liu et al., 2016), and according to many other practices that 5–6 clusters or factors could well cover most major sources (Table S4) (Brines et al., 2015; Masiol et al., 2017a; Masiol et al., 2016; Vu et al., 2015), we uniformly set the number of clusters, components, or factors for every case as 6. It's rational to cover most major sources and very convenient to test and compare the basic functions of different models. Although every model has advanced functions or/and can be optimized, basic functions (such as base run in PMF) are core for each model and thus are valuable and necessary to be compared. This is one of the goals in the present study. In deep investigation of PNC sources with a single receptor model, it might be better to further use optimization tools to find the "best" number of cluster, components, or factors. Appropriate optimization tools should fully consider both the nature of PNC data and the specific emissions around sites to reduce the gap between mathematical results and real situations. Factor analysis itself is a combination of art (user's understanding (Norris et al., 2014)) and science (Teetor, 2011), the results of aerosol source apportionment with factor analysis can be used as a limited but useful reference rather than being completely trusted because of the gap just mentioned.

It should first be pointed out that the source apportionment results obtained by *k*-means clustering and factor analysis models (PCA, FA, PMF, and NMF) are mainly the artificially named sources. Such a source of them is determined according to the characteristics of PNSD peaks, time patterns, correlations, and CWT. It is generally the main source of the factor or type, not the only actual source.

Fingerprints of PNSD, coupled with associated pollutants, time patterns, etc., can give information on sector sources and chemicophysical processes (Harrison et al., 2011; Hussein et al., 2014; Masiol et al., 2017b). The association of fingerprints with certain pollutants would reveal specific sector sources. By combining fingerprints with local winds and regional backward trajectories, the relationship between each fingerprint and atmospheric conditions and geographical origins can be determined. To distinguish the effects of photochemistry and liquid chemistry, meteorological parameters such as solar radiation, temperature and humidity need to be supplied.

3.3.3. Shortcomings of k-means clustering

We used original prepared data and its various transformed data (centering scaled, 0–100 normalized, and percent rank normalized) to test *k*-means clustering. *k*-means clustering yields "regular" PNSD peaks (Figs. S20 and S21), but firstly, the concentration discrimination is several orders of magnitude worse than a usual peak by PMF. Secondly, although the linear scale PNSD is better distinguished, its peak values are more concentrated in the Aitken mode, so it has little practical significance. Thirdly, criteria pollutants cannot be distinguished due to their too small values and need to be normalized, while normalization cannot fundamentally solve its limitations. Fourthly, the apportionment of sources by time series clustering does not match the objective situation that multiple sources generally exist concurrently. To conveniently test and compare the performance of different models, we used 6 clusters. The best number of clusters found by R package NbClust (Charrad et al., 2014) is 3 according to its judgement based on 26 indices. However, the limitations of *k*-means clustering are almost the same for both 6 clusters and the best 3 clusters (Figs. S22 and S23). Therefore, the well-recognized *k*-means clustering (Beddows et al., 2009; Charron et al., 2008; Dall'Osto et al., 2012; Hussein et al., 2014; Masiol et al., 2017a; Masiol et al., 2016; Wegner et al., 2012) is not good and needs to be discarded for the cases here. We had to turn to other methods such as factor analyses including PCA and FA in addition to PMF. *k*-means clustering can be good for cases near sources such as airports (Masiol et al., 2017a) and for identifying groups of samples with similar size distribution spectra.

3.3.4. Results and shortcomings of PCA and FA

PCA and FA combine the PNSD peaks, correlations of component loadings with CP, time patterns, correlations of standardized scores and meteorology, and CWT to apportion the traffic sources, Accumulation process (hereinafter the mode names in sources are for processes not for modes) and Nucleation (Figs. S24 and S25). During heating period, CWT and wind speed complement each other: it corresponds to Nucleation and fresh traffic sources when the CM of trajectory pressure is small, while those of trajectory height and wind speed are large. Otherwise, it generally corresponds to Accumulation source. However, PCA and FA cannot apportion coal heating. Moreover, they produce 'mixed' (mixed source) that is difficult to be determined as a certain major source. Furthermore, there are negative numbers in the loadings of PCA and FA, making the percentages of species, trimodal ratio (Nucleation:Aitken:Accumulation, N/A/A), and SO_2/NO_2 ratio difficult to be determined. Consequently, PCA and FA are useful but not so suitable for PNC source apportionment.

3.3.5. Results and comparison of PMF and NMF

For PMF, we computed the uncertainty according to a published method (Thimmaiah et al., 2009). The units of PNC and mass concentration of CP are # cm^{-3} and μg m^{-3} respectively. Specifically, the uncertainties of PNC (x represents concentration) are determined by 1 + x^0.5 + 0.1*x. The uncertainties of CP (O_3, SO_2, NO_2, and $PM_{2.5}$) are determined by 0.5*min(x) + 0.1*x. The uncertainty of CO is calculated from 0.5*100 + 0.1*x. It should have been 0.5*min(x) + 0.1*x for CO, but its concentration in the raw data is in mg m^{-3} and only has one digit after the decimal point. Its minimum is 0, but actually the true value would be around 0.1. Therefore, we set the minimum of CO as 0.1 mg m^{-3}, namely 100 μg m^{-3}. The data are processed all together for uncertainties but are processed separately case by case for base model runs. We set variables (size bins and CP) with S/N < 6 and the total variable TPNC (namely total particle number concentrations, generally with S/N > 6) as weak for all cases. There are around 10 weak variables out of the total 104 variables in each case. No extra modeling uncertainty is added.

Recommended number of runs (20) in PMF (Norris et al., 2014) is used for every case. Generally, the number of absolute scaled residuals beyond 3 is very limited in size bin variables but relatively greater in CP (Table 2). This reflects the ability of PMF in distinguishing sources of PNC and mass concentrations of CP. A major PNC source sometimes could be an infrequent source for some CP. For the regressions, the average values of R^2 are not lower than 0.96 for size bins and not higher than 0.65 for CP respectively. For 6 factors of each case, at least 52.9% and 80% of the run profiles are inside of the bootstrap (n = 200) interquartile range for size bins and CP respectively. Mapping of bootstrap factors to base factors in 8 periods is generally over 95%. There are no factor swaps. The decreases in Q are all much <1%. The displacement range of mixed sources is relatively larger than that of non-mixed sources. Specifically, the average ratios of displacement range difference (DISP Max - DISP Min) to contribution (Base Value) are 0.81 and 0.3 for mixed sources and non-mixed sources respectively. G-space is helpful to distinguish factors and verify optimal solutions. It is always consistent with the differences of PNSD peaks of factors. For different numbers of factors (4–10 factors) of THU-15, only the solution with 10 factors contains swaps that are relatively huge compared with other solutions without swaps and among five factors instead of only between two factors.

For NMF, we used default parameters and methods in the R package NMF (Gaujoux and Seoighe, 2010; Gaujoux and Seoighe, 2018) except setting 6 factors and 20 runs.

Unlike PCA and FA, PMF can apportion coal heating (Fig. S26). Species percentages, trimodal ratio N/A/A and SO_2/NO_2 ratio can be determined as auxiliary. However, PMF still has more mixed sources than NMF (Figs. 5 and 6). Besides, the operation speed of PMF is 11–28 times slower than NMF with ability of parallel computations (Gaujoux and Seoighe, 2010; Gaujoux and Seoighe, 2018). True uncertainties would really help in accounting for the confidence in measurement by weighting individual points (Norris et al., 2014). For example, points below detection would have less influence on the solution. However, uncertainties input into PMF are artificially estimated (calculated) and vary along with different estimation formulae or equations proposed by different researchers. These estimated uncertainties themselves are uncertain, can be higher or lower than the true uncertainties, and thus could bring invisible or unquantifiable side effects. Therefore, the uncertainties of species calculated for PMF (Thimmaiah et al., 2009) may cause complexity because of the side effects, producing mixed sources. For the 8 cases, uncertainty-input PMF produced 8 mixed sources while non-uncertainty-input NMF produced no mixed sources. Actually every source identified by receptor models would more or less contain a fraction of other sources. It is the dominant source in a factor or type that is named. When the dominant source is hard to determine or the factor obviously contains multiple dominant sources, it is a mixed source. The difference between PMF and NMF in the heating period is slight (Fig. S27). The differences in traffic and Accumulation can be explained by the mixed source. The main difference between the two methods in non-heating period is Nucleation. In 2015 and 2017, Nucleation processes could not be apportioned by PMF but would be explained by mixed sources. Changing the number of factors may avoid the mixed sources. For example, decreasing the number from 6 to 4 or 5 will avoid the mixed source (F6) for the case THU-H15 (Fig. 5).

The bases for naming the factors N1 to N6 (Fig. 6) are specified as follows.

Table 2
Diagnostics and error estimations of PMF results.

Cases	Number of factors	Percent (%) of absolute scaled residuals >3		Regression R^2		Percent (%) of base run profile within bootstrap IQR		Mapping (%) of bootstrap factors to base factors	Counts of factor swaps	Decrease in Q (%)	Contribution and displacement ranges (%)
		Size bins	CP	Size bins	CP	Size bins	CP				
THU-H15	6	0.18	0.72	0.99	0.65	54.2	96.7	99.7	0	−0.001132	4.3–29.4 (0.6–41.7)
THU-NH15	6	0.39	0.74	0.96	0.43	63.6	93.3	98.2	0	−0.000944	1.6–28.8 (0.7–32.3)
IAP-H17	6	0.28	1.45	0.98	0.42	81.8	96.7	95.6	0	0	2.3–28.7 (2.1–29.8)
IAP-NH17	6	0.22	0.75	0.96	0.58	83.7	93.3	99.3	0	0	9.1–32.4 (7–33.5)
NJ-H18	6	0.12	1.07	0.99	0.35	75.1	90	95.6	0	0	5.8–25.9 (5.6–26.7)
NJ-NH18	6	0.32	1.16	0.97	0.42	100	100	100	0	0	8–24.7 (7.9–25)
NJ-H19	6	0.08	0.9	0.99	0.41	60.6	80	92	0	−0.000031	4.1–34.7 (3.4–36.3)
NJ-NH19	6	0.36	1.98	0.96	0.18	52.9	83.3	87.1	0	0	5.2–27.6 (4.4–29)
THU-H15	4	0.95	2.07	0.96	0.46	78.3	95	99.5	0	−0.000231	13.4–34.7 (9.5–39.2)
THU-H15	5	0.37	1.44	0.98	0.55	40.2	44	96.6	0	0	8.5–30.1 (7.1–34.1)
THU-H15	7	0.12	0.7	0.99	0.66	70	85.7	94.6	0	−0.002067	3–24.9 (0.5–37)
THU-H15	8	0.12	0.28	0.99	0.78	58.3	75	94.7	0	−0.279801	1.8–22.9 (0.3–35.9)
THU-H15	9	0.1	0.07	1	0.81	69.9	84.4	94.9	0	−0.002547	1.7–22.1 (0.8–35.9)
THU-H15	10	0.06	0.06	1	0.81	60.2	86	93.7	294	−0.002563	0.3–19.6 (0.1–35.2)

- Size bins include TPNC here.

N1 is coal heating, which is mainly based on: 1. The peak of PNSD appears in the range of 100–300 nm; 2. The ratio of sulfur and nitrogen (SO_2/NO_2) is high; 3. The peak of normalized contribution appears at night; 4. The normalized contribution and the temperature are mainly inversely correlated. The wind direction and backward trajectory show the factor is mainly from Beijing-Tianjin-Hebei and the air mass migration distance is short.

N2 is fresh traffic, mainly based on: 1. The peak of PNSD appears at about 30 nm; 2. The ratio of sulfur and nitrogen (SO_2/NO_2) is low; 3. The peak of normalized contribution appears at morning and evening rush hours; 4. CM_{WS} is high, conducive to the rapid spread of fresh traffic particles to the monitoring site. Longer trajectories and stronger winds bring more fresh emissions, while shorter trajectories and weaker winds allow emissions more aging time. Therefore, regionally, there are much less short trajectories from west for fresh traffic than for the two aged traffic factors (N3 and N4). Directionally, with stronger northwest winds, the normalized contributions (NC) of fresh traffic are the most obvious (dark red) in northwest. With weaker winds, the NC of the two aged traffic factors are the most obvious in southeast, reflecting the impact of the whole urban Beijing on the aged traffic in THU.

N3 is aged traffic 1. It is mainly based on: 1. The peak of PNSD appears at about 100 nm; 2. The ratio of sulfur and nitrogen (SO_2/NO_2) is low; 3. The peak of normalized contribution appears at the morning and evening rush hours; 4. The normalized contribution from the main urban (southeast) area is strong (darker color of the rose chart) despite low wind frequency; 5. CM_{WS} is small, so the traffic particles have aging time during the diffusion process.

N4 is aged traffic 2. It is mainly based on: 1. The peak of PNSD appears in the range of 30–100 nm; 2. The peak of normalized contribution appears in the morning and evening rush hours; 3. The normalized contribution from the main urban area is strong despite low wind frequency; 4. CM_{WS} is small, hence there is aging time for the traffic particles during diffusion.

N5 is Accumulation (process source), which is mainly based on: 1. The peak of PNSD appears at above 300 nm or even 500 nm; 2. The peak of normalized contribution appears at night; 3. The positive correlation between normalized contribution and relative humidity and the negative correlations of normalized contribution with air pressure and wind speed are very regular in diurnal changes; 4. CMtp is the highest, while CMth and CM_{WS} are the lowest, indicating stagnant meteorological condition that facilitates the Accumulation process of particles; 5. Its contributions to CO and $PM_{2.5}$ is the largest, which confirms the Accumulation process.

N6 is Nucleation (process source), mainly based on: 1. The peak of PNSD appears below 20 nm; 2. The peak of normalized contribution appears at noon; 3. The negative correlation of normalized contribution with relative humidity and the positive correlation of normalized contributions with air pressure and wind speed are diurnally regular; 4. CMtp is the lowest, CMth and CM_{WS} are the highest. It is a typical windy weather with active air mass migration, which promotes the Nucleation process; 5. It contributes the most to O_3, which confirms the Nucleation process.

Among these 6 sources, the particle size peak of fresh traffic is smaller than that of aged traffic, which is consistent with the aging process of vehicle emissions accompanied by particle growth (Liu et al., 2020). According to CWT and wind rose, coal heating mainly comes from neighboring cities and local areas (outside the Fifth-Ring Road), especially in areas to southwest and southeast of Beijing. The results of the Pollution Permeation Index (PPI) in 2015 also verified the input of SO_2 (important marker of coal burning) from these directions to Beijing, namely Baoding (to the southwest of Beijing) and Tangshan (to the southeast of Beijing) are the donors of SO_2, while Beijing is the SO_2 receptor (Liang et al., 2016b).

Considering time cost, the PMF calculation program me2gfP4_1345c4.exe is not as reasonable as the nmf function of R's NMF package. NMF is the best and advised to be adopted in PNC source apportionment because NMF produces more certain results and runs significantly faster than PMF that is better than *k*-means clustering, PCA, and FA.

3.4. Source apportionment

Results of NMF applied to the other 7 cases are shown in Figs. S28–S34. N/A/A ratio complements the shortcomings of PNC peaks in quantifying the structures of three modes. SO_2/NO_2 ratios are generally higher in heating periods than in non-heating periods. In heating periods, SO_2/NO_2 ratios of coal heating are generally the highest among all sources.

The contributions of NMF factors are between 8.9% and 22.6% (Fig. S35). The contributions of traffic vary much more in heating periods (54.4%–78.5%) than in non-heating periods (68.1%–74.4%) (Fig. 7). Coal heating generally contributes less and less year by year from 2015 to 2019. The change of coal heating contribution between 2017 and 2018 is the most obvious, demonstrating the efficiency of robust measures such as "2 + 26 cities" taken since 2017 (Chen and Chen, 2019). One of the goals of this measure is to shift coal to gas and

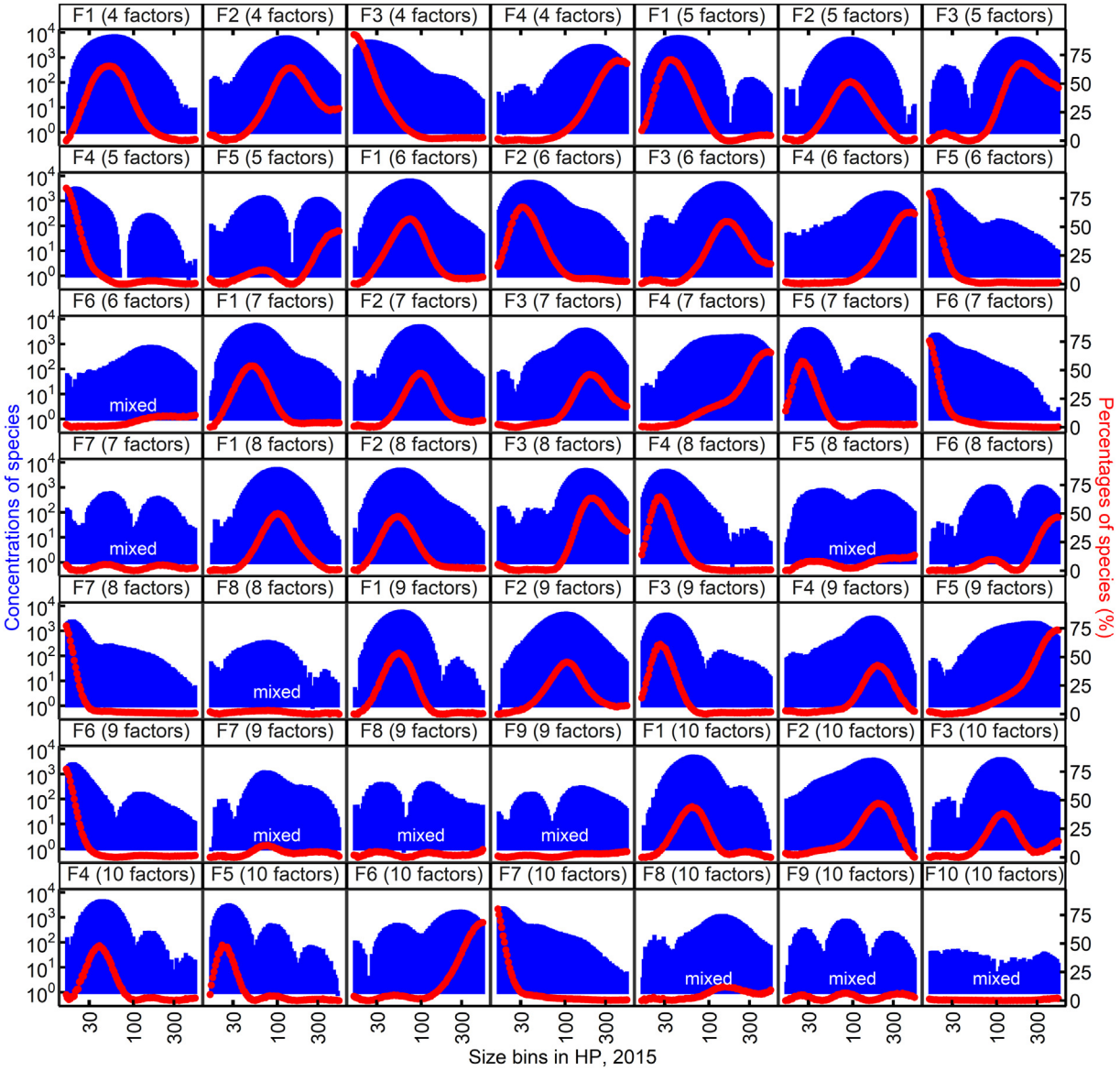


Fig. 5. PMF results in heating period of 2015 with different numbers of factors.

electricity. At ordinary sites THU and IAP, daytime Nucleation can be apportioned, while at NJ it cannot, indicating the difficulty of daytime new particle formation (NPF) at traffic site. However, nighttime peaks of particles below 20 nm are easy to occur at NJ (Figs. S31–S34). We set high requirements for episode classification in the examples, which can be changed to get more episodes. THU and IAP have much lower PNC of Nucleation (Fig. 3b), so the Nucleation episodes (can occur in both daytime and nighttime) are many fewer (only one) in THU and IAP (Fig. S16), compared with NJ.

With the increase of the sample size, the advantages of NMF in factor discrimination and parallel computing running speed become more obvious and important, making NMF suitable for popularization.

3.5. Typical sources in heating periods

3.5.1. Traffic and coal heating

Traffic correlates most strongly with particles of 30–100 nm and coal heating correlates most strongly with particles of 100–300 nm (Fig. 8). In the transportation site (year 2019), traffic also correlates most strongly with particles below 30 nm. The correlation of traffic with NO_2 is much weaker than the correlation of coal heating with SO_2 (Fig. S36). Meanwhile, coal heating is more positively correlated with SO_2/NO_2 than traffic is. Coal heating and traffic are approximately symmetric about the line $R = 0$, indicating the distinct types of the sources.

3.5.2. Nucleation and Accumulation

Compared with ordinary sites (years 2015 and 2017), the correlation of O_3 with particles below 30 nm in the transportation site (year 2019) is much weaker (Fig. S37). As its counterpart, there is no Nucleation source (for the particles below 30 nm) in transportation site. Only daytime Nucleation concentration peaks are regarded as the result of Nucleation source in this work. However, the strong correlations of $PM_{2.5}$ with particles above 100 nm are very alike among the three cases including ordinary and transportation sites. Correspondingly, there are Accumulation sources in the three different sites. In transportation site, $PM_{2.5}$ is partially strongly correlated with particles below 30 nm besides above 100 nm, suggesting the partial homology (same source type) of emitting small particles and big particles in transportation site. Nucleation is much more positively correlated with O_3 and Accumulation is much more positively correlated with $PM_{2.5}$ (Fig. S38). The

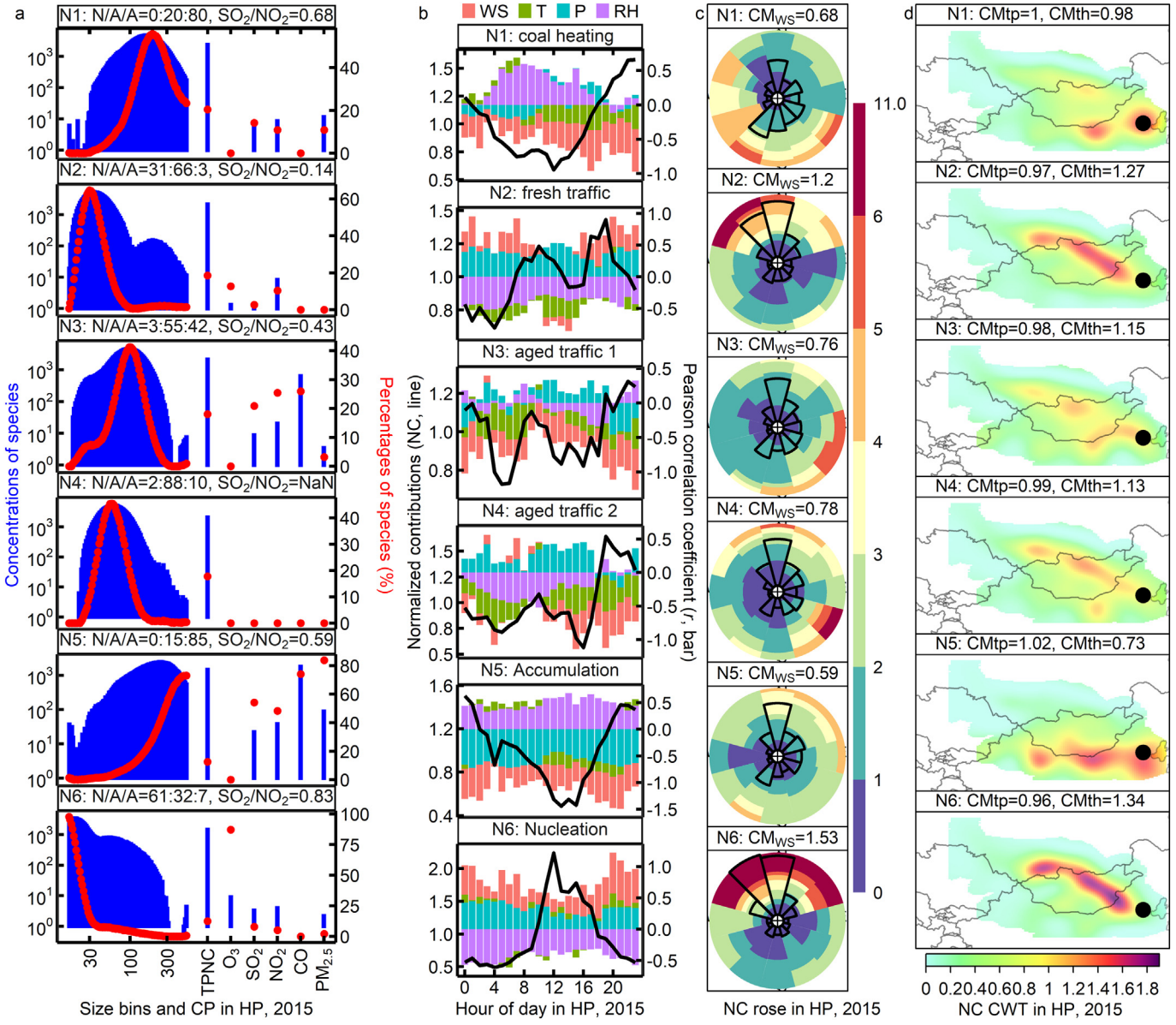


Fig. 6. NMF results in heating period of 2015.

correlation coefficients of Accumulation and Nucleation with O_3 and $PM_{2.5}$ are generally opposite numbers, showing that Accumulation and Nucleation represent different sources and processes in ordinary sites.

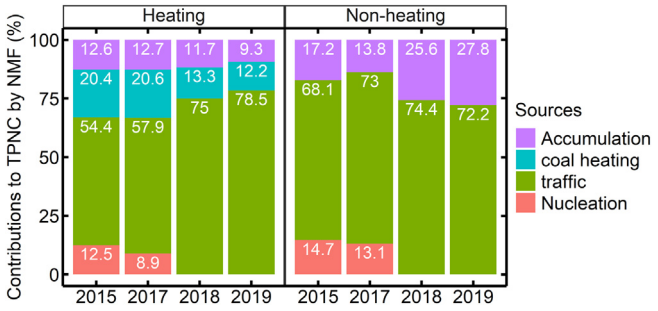


Fig. 7. Variation of sources by NMF.

3.5.3. Actual impact of coal heating

Coal heating is regional for its wide use in north China and contribution to particles bigger than 100 nm, so its interannual trend from the three sites within a 15 km radius is convenient to assess. For the three modes, both coal heating and $PM_{2.5}$ are most relevant to the Accumulation mode (Figs. S39 and S40). According to these realities, the impact of coal heating from the perspective of big particles can be investigated. Specifically, the contributions can be normalized by big particles.

Based on the fraction of Accumulation F_{Accu} and the contribution percentage to total concentration TPNC (P_{TPNC}), the normalized contribution percentages are defined as follows:

Accumulation normalized contribution percentage:

$$F_{Accu} = Accu/TPNC \quad (8)$$

$$P_{Accu} = P_{TPNC} * F_{Accu}/mean(F_{Accu}) \quad (9)$$

$PM_{2.5}$ normalized contribution percentage:

$$P_{PM2.5} = P_{Accu} * PM_{2.5}/mean(PM_{2.5}) \quad (10)$$

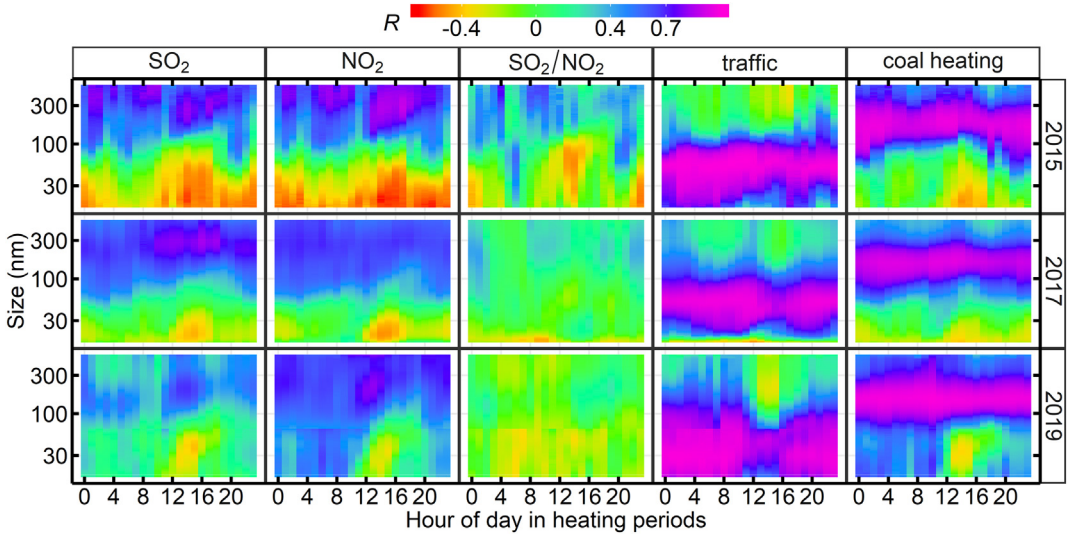


Fig. 8. Hour of day variation of correlation of size bins with SO_2, NO_2, traffic and coal heating in heating periods.

In terms of P_{TPNC}, the contribution of coal heating decreased by 40.2% from 20.4% in 2015 to 12.2% in 2019 (Fig. 9). While according to P_{Accu}, the contribution of coal heating decreased by 74.4% from 27% in 2015 to 6.9% in 2019. Furthermore, referring to $P_{PM2.5}$, the contribution of coal heating decreased by 85.5% from 35.8% in 2015 to 5.2% in 2019. These decreases suggest the coal control for winter particle pollution has been very effective in recent years.

3.5.4. Abundance of Nucleation episodes and absence of Nucleation sources in transportation site

There are Nucleation sources in ordinary sites with no (THU) or few (IAP) Nucleation episodes but no Nucleation sources in transportation site with almost all Nucleation episodes (Figs. S35, 7, S37, S38, and S16). On the one hand, the PNC were uniformly percent rank normalized in all sites, while the PNC of Nucleation mode in transportation site is much higher than those in ordinary sites (Fig. S40). Consequently, Nucleation episodes based on percent rank normalization almost only exist in transportation site. If the normalization is grouped by sites, namely separately normalized for individual sites (one by one), there will be 13 more Nucleation episodes in ordinary sites while only 7 more Nucleation episodes in transportation site (Figs. S41 and S16).

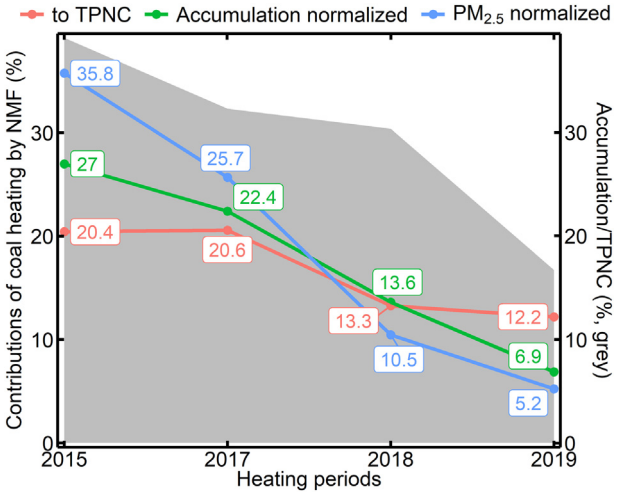


Fig. 9. Big particle normalized contributions of coal heating to TPNC in heating periods.

On the other hand, concentrations of Nucleation mode in ordinary sites peak at midday and rush hours, while in transportation site the Nucleation concentration peaks at nights (Fig. 3b). However, only daytime (photochemical) Nucleation concentration peaks are regarded as the result of Nucleation source here and nighttime (aqueous, heterogeneous or hygroscopic) ones are not, therefore Nucleation sources were not apportioned in the transportation site.

There could be NPF (noontime Nucleation process) sources in the transportation site (NJ). However, compared with nighttime Nucleation process, the noontime Nucleation process is much weaker since the Nucleation concentrations are lowest at noontime and the GMD peaks at noontime in NJ (Fig. 3b). In NJ, the highest PNC levels of Nucleation particles from big trucks at night are further supplemented by commuting vehicles during rush hours. Therefore, when they hit bottom at noontime, the remaining traffic Nucleation particles are still so abundant that they could largely conceal NPF. In other words, the noontime Nucleation process is relatively (compared with traffic Nucleation particles) so weak that observing NPF is difficult in NJ. Besides, high PNC levels of pre-existing particles have high condensation and coagulation sinks that can suppress NPF (Kulmala et al., 2004; Nie et al., 2014). Consequently, the remaining traffic Nucleation particles are still so abundant at noontime that they could suppress NPF.

3.6. Methodological implications for further PNC source apportionment studies

The contributions of this work to further PNC source apportionment studies may include:

1) Introduction of NMF with fast speed and obvious discrimination of factors. The volume of long-term PNC data with around 100 size bins is huge, especially when the time resolution is 5 min or so. Running speed of receptor model for PNC source apportionment is vital. NMF is qualified in speed for its parallel computing. Moreover, NMF shows advantage in factor discrimination that is also important for PNC source apportionment.
2) Proposals of concepts of contribution moment and big particle normalized contribution. Contribution moment quantifies the impact of meteorological parameters such as wind and trajectory. It explicitly shows the degree of meteorological effect on midday Nucleation and nighttime Accumulation processes. Big particle normalized contribution quantitatively links TPNC with number concentrations of different modes and particle mass concentrations (PMC) regarding

the contributions of sources. The $PM_{2.5}$ normalized contribution is highly consistent with the published results of studies based on chemical composition and mass contribution (Chen et al., 2019; P.F. Liu et al., 2017).

3) Comparison of advantages and disadvantages of 5 receptor models. Based on the actual performance of basic functions of each source apportionment receptor model, their advantages and disadvantages can be compared (Table 3). This would be helpful to future PNC source apportionment studies. Combining the advantages of NMF in speed and discriminating factors and the advantages of PMF in solution optimization is ideal. More ideally, the speed of NMF could also be improved to meet the future requirement from real-time PNC source apportionment. Besides, the solution optimization tools should be improved by thoroughly considering the features of PNC data and the specific emissions around sites to narrow the gap between mathematical results and real situations.

4. Conclusions

Based on the PNC data of 3 sites in Beijing involving 33 months from 2015 to 2019 and auxiliary data of CP, Met, and Traj, we developed a series of efficient methods for data preprocessing, episode classification, and source apportionment of PNC.

Table 3
Comparison of 5 receptor models for PNC source apportionment.

Models	Advantages	Disadvantages
k-means clustering	- Ratios of species can be calculated; - Extremely fast.	- Very low discrimination of concentrations; - Low discrimination of peak particle sizes; - Source apportionment by time classification is not practical.
PCA	- Extremely fast.	- Cannot apportion coal heating; - Ratios of species cannot be calculated (contains negative component loadings); - Produces mixed source(s).
FA	- Extremely fast.	- Cannot apportion coal heating; - Ratios of species cannot be calculated (contains negative factor loadings; - Produces mixed source(s).
PMF	- High discrimination of concentrations; - High discrimination of peak particle sizes; - High discrimination of ratios of species; - High discrimination of contribution moments; - Can apportion coal heating; - Multiple solution optimization tools.	- Produces mixed source(s); - Lacks alternative matrix factorization algorithms; - Base run is very slow; - BS-DISP (Bootstrap-Displacement) is extremely slow.
NMF	- High discrimination of concentrations; - High discrimination of peak particle sizes; - High discrimination of ratios of species; - High discrimination of contribution moments; - Can apportion coal heating; - Produces no mixed source(s); - Multiple alternative matrix factorization algorithms; - Very fast (parallel computing).	- Lacks some of the solution optimization tools that PMF has.

For data preprocessing, we developed new methods such as the automatic identification of consecutive NA for more than half an hour based on moving averages and the point-by-point considered outlier removal based on conditional extremum, after overcoming the shortcomings of previous methods that delete too many values, delete inadequately, or generate new outliers. In the original data, there are too many NA at both ends of the particle size, observations with too many consecutive NA in time series, and outliers in each particle size bin. Our developed preprocessing method worked well by firstly deleting unqualified variables and observations, then removing outliers, and finally interpolating for NA. They delete less values, delete more completely, and generate no new outliers.

For characteristics analysis (mainly episode classification), we put forward the definition and automatic division method of PNC pollution episodes, revealed the pollution law involving all parameters, and compiled simple algorithms of GMD and CMD. The episode classification is very flexible. Concentration thresholds are based on percent ranks, which allow convenient comparisons of PNC despite the lack of standards. According to the general situations of pollution, users can get different numbers (vary with strictness) of episodes by adjusting threshold windows, duration hours, and trend constraints.

Based on NMF, coupled with the newly-proposed contribution moment and big particle normalized contribution, an efficient PNC source apportionment and assessment system has been established. For source apportionment of hourly averaged PNC here, *k*-means clustering, PCA, and FA are not so suitable, while PMF and NMF are both suitable. NMF is more certain in results and runs much faster than PMF. Typical sources of heating periods in Beijing include Nucleation, traffic, coal heating, and Accumulation. The PNC source apportionment can well reflect the sources of particles below 100 nm, especially traffic sources. In 2017, the contribution of traffic to TPNC was 68%, which was 23% (relatively 51%) higher than the local contribution of traffic to $PM_{2.5}$, namely 45% (Hou, 2018). From the perspective of TPNC that mainly represent the total concentration of small particles below 100 nm, the contribution of coal heating has decreased by a small part (40%) from 2015 to 2019. However, from the perspectives of Accumulation mode and $PM_{2.5}$ that mainly represent big particles above 100 nm, the contribution of coal heating has decreased by a major part (around 80%). These decreases suggest the effectiveness of coal burning control for winter particle pollution in recent years. Overall, this method system can not only apportion the sources of PNC that mainly represent particles below 100 nm, but also help to explain the sources of PMC that mainly represent particles above 100 nm. The general suggestion for future studies is to take full advantage of NMF's matrix factorization and absorb PMF's self-evaluation tools (diagnostics and error estimation) in parallel computing, especially in the continuously perfected open source platform R since it includes abundant and powerful functions for developing efficient and tailored methods for PNC data. Similarly, the EPA PMF can be improved by absorbing the advantages of NMF's matrix factorization and parallel computing.

CRediT authorship contribution statement

Chun-Sheng Liang: Methodology, Writing - original draft. **Hao Wu:** Investigation, Formal analysis, Writing - review & editing. **Hai-Yan Li:** Investigation, Writing - review & editing. **Qiang Zhang:** Writing - review & editing. **Zhanqing Li:** Writing - review & editing. **Ke-Bin He:** Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors gratefully acknowledge the National Natural Science Foundation of China (41571130035, 41921005, 91544217, 41925022, and 91837204), the National Research Program for Key Issues in Air Pollution Control (DQGG0201), and other funds for field experiments in IAP and NJ (Z. Li et al., 2019). We thank Prof. Jing-Kun Jiang for reading and improving this manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2020.140923.

References

Al-Dabbous, A.N., Kumar, P., 2015. Source apportionment of airborne nanoparticles in a Middle Eastern city using positive matrix factorization. Environmental Science-Processes & Impacts 17, 802–812.
Bache SM, Wickham H. magrittr: A Forward-Pipe Operator for R. Vienna, Austria: The R Foundation. R package version 1.5. https://CRAN.R-project.org/package=magrittr. 2014.
Baldauf, R.W., Devlin, R.B., Gehr, P., Giannelli, R., Hassett-Sipple, B., Jung, H., et al., 2016. Ultrafine particle metrics and research considerations: review of the 2015 UFP workshop. Int. J. Environ. Res. Public Health 13.
Bartczak, D., Goenaga-Infante, H., 2016. Particle number size distribution. In: Tantra, R. (Ed.), Nanomaterial Characterization: An Introduction. John Wiley & Sons, Inc, Hoboken, New Jersey, pp. 63–80.
Bartholomew, D.J., 1995. Spearman and the origin and development of factor analysis. Br. J. Math. Stat. Psychol. 48, 211–220.
Baumer, D., Vogel, B., Versick, S., Rinke, R., Mohler, O., Schnaiter, M., 2008. Relationship of visibility, aerosol optical thickness and aerosol size distribution in an ageing air mass over South-West Germany. Atmos. Environ. 42, 989–998.
Beaudrie, C.E.H., Kandlikar, M., Ramachandran, G., 2016. Chapter 5 - Using Expert Judgment for Risk Assessment. In: Ramachandran, G. (Ed.), Assessing Nanoparticle Risks to Human Health, Second edition William Andrew Publishing, Oxford, pp. 91–119.
Beddows, D.C.S., Dall'Osto, M., Harrison, R.M., 2009. Cluster analysis of rural, urban, and curbside atmospheric particle size data. Environmental Science & Technology 43, 4694–4700.
Beddows, D.C.S., Harrison, R.M., Green, D.C., Fuller, G.W., 2015. Receptor modelling of both particle composition and size distribution from a background site in London, UK. Atmos. Chem. Phys. 15, 10107–10125.
Brines, M., Dall'Osto, M., Beddows, D.C.S., Harrison, R.M., Gomez-Moreno, F., Nunez, L., et al., 2015. Traffic and nucleation events as main sources of ultrafine particles in high-insolation developed world cities. Atmos. Chem. Phys. 15, 5929–5945.
Buonanno, G., Ficco, G., Stabile, L., 2008. Size distribution of ultrafine particles and trends of concentration near a linear (major highway) and point source (waste incinerator). In: Ranzi, E. (Ed.), Aaas08: 2nd Advanced Atmospheric Aerosol Symposium. vol. 16. Aidic Servizi Srl, Milano, pp. 95–104.
Buseck, P.R., Adachi, K., 2008. Nanoparticles in the atmosphere. Elements 4, 389–394.
Bycenkiene, S., Ulevicius, V., Prokopciuk, N., Jasineviciene, D., 2013. Observations of the aerosol particle number concentration in the marine boundary layer over the south-eastern Baltic Sea. Oceanologia 55, 573–597.
Bycenkiene, S., Plauskaite, K., Dudoitis, V., Ulevicius, V., 2014. Urban background levels of particle number concentration and sources in Vilnius, Lithuania. Atmos. Res. 143, 279–292.
Carnerero, C., Pérez, N., Petäjä, T., Laurila, T.M., Ahonen, L.R., Kontkanen, J., et al., 2019. Relating high ozone, ultrafine particles, and new particle formation episodes using cluster analysis. Atmospheric Environment: X 4, 100051.
Carslaw, D., 2019. worldmet: Import Surface Meteorological Data from NOAA Integrated Surface Database (ISD). R package version 0.8.7. https://CRAN.R-project.org/package=worldmet.
Carslaw, D.C., Ropkins, K., 2012. openair - an R package for air quality data analysis. Environ. Model Softw. 27-28, 52–61.
Carslaw, D., Ropkins, K., 2019. openair: Tools for the Analysis of Air Pollution Data. R package version 2.7-0. https://cran.r-project.org/web/packages/openair/.
Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A., 2014. Nbclust: an R package for determining the relevant number of clusters in a data set. J. Stat. Softw. 61, 1–36.
Charron, A., Birmili, W., Harrison, R.M., 2008. Fingerprinting particle origins according to their size distribution at a UK rural site. J. Geophys. Res.-Atmos. 113.
Chen, H., Chen, W., 2019. Potential impact of shifting coal to gas and electricity for building sectors in 28 major northern cities of China. Appl. Energy 236, 1049–1061.
Chen, J., Zhao, C.S., Ma, N., Liu, P.F., Gobel, T., Hallbauer, E., et al., 2012. A parameterization of low visibilities for hazy days in the North China Plain. Atmos. Chem. Phys. 12, 4935–4950.
Chen, X.S., Wang, Z.F., Li, J., Chen, H.S., Hu, M., Yang, W.Y., et al., 2017. Explaining the spatiotemporal variation of fine particle number concentrations over Beijing and surrounding areas in an air quality model with aerosol microphysics. Environ. Pollut. 231, 1302–1313.
Chen, X.S., Wang, Z.F., Li, J., Yang, W.Y., Chen, H.S., Wang, Z., et al., 2018. Simulation on different response characteristics of aerosol particle number concentration and mass concentration to emission changes over mainland China. Sci. Total Environ. 643, 692–703.
Chen, Z.Y., Chen, D.L., Wen, W., Zhuang, Y., Kwan, M.P., Chen, B., et al., 2019. Evaluating the "2+26" regional strategy for air quality improvement during two air pollution alerts in Beijing: variations in PM2.5 concentrations, source apportionment, and the relative contribution of local emission and regional transport. Atmos. Chem. Phys. 19, 6879–6891.
Cheng, J., Karambelkar, B., Xie, Y., 2019. leaflet: Create Interactive Web Maps With the JavaScript 'Leaflet' Library. R package version 2.0.3. https://CRAN.R-project.org/package=leaflet.
Cusack, M., Perez, N., Pey, J., Alastuey, A., Querol, X., 2013. Source apportionment of fine PM and sub-micron particle number concentrations at a regional background site in the western Mediterranean: a 2.5 year study. Atmos. Chem. Phys. 13, 5173–5187.
Dal Maso, M., Kulmala, M., Riipinen, I., Wagner, R., Hussein, T., Aalto, P.P., et al., 2005. Formation and growth of fresh atmospheric aerosols: eight years of aerosol size distribution data from SMEAR II, Hyytiala, Finland. Boreal Environ. Res. 10, 323–336.
Dall'Osto, M., Beddows, D.C.S., Pey, J., Rodriguez, S., Alastuey, A., Harrison, R.M., et al., 2012. Urban aerosol size distributions over the Mediterranean city of Barcelona, NE Spain. Atmos. Chem. Phys. 12, 10693–10707.
Demin, G., 2019. expss: tables, labels and some useful functions from spreadsheets and 'SPSS' statistics. R package version 0.10.1. https://CRAN.R-project.org/package=expss.
Dowle, M., Srinivasan, A., 2019. data.table: Extension of 'data.frame'. R package version 1.12.8. https://CRAN.R-project.org/package=data.table.
Du, W., Zhao, J., Wang, Y.Y., Zhang, Y.J., Wang, Q.Q., Xu, W.Q., et al., 2017. Simultaneous measurements of particle number size distributions at ground level and 260m on a meteorological tower in urban Beijing, China. Atmos. Chem. Phys. 17, 6797–6811.
Dusek, U., Frank, G.P., Hildebrandt, L., Curtius, J., Schneider, J., Walter, S., et al., 2006. Size matters more than chemistry for cloud-nucleating ability of aerosol particles. Science 312, 1375–1378.
EU. Commissions regulation (EU) no. 459/2012, Off. J. Eur. Union. http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32012R0459 2012 (last access: 1 November 2017).
Fellows, I., Stotz, J.P., 2019. OpenStreetMap: Access to Open Street Map Raster Images. R package version 0.3.4. https://CRAN.R-project.org/package=OpenStreetMap.
Forgy, E.W., 1965. Cluster analysis of multivariate data: efficiency vs. interpretability of classifications. Biometrics 21, 768–769.
Frampton, M., Brauer, M., Kleeman, M., Kreyling, W., Ntziachristos, L., Sarnat, S., 2013. Understanding the Health Effects of Ambient Ultrafine Particles. Health Effects Institute, Boston, MA.
Friend, A.J., Ayoko, G.A., Jayaratne, E.R., Jamriska, M., Hopke, P.K., Morawska, L., 2012. Source apportionment of ultrafine and fine particle concentrations in Brisbane, Australia. Environ. Sci. Pollut. Res. 19, 2942–2950.
Friend, A.J., Ayoko, G.A., Jager, D., Wust, M., Jayaratne, E.R., Jamriska, M., et al., 2013. Sources of ultrafine particles and chemical species along a traffic corridor: comparison of the results from two receptor models. Environ. Chem. 10, 54–63.
Gaujoux, R., Seoighe, C., 2010. A flexible R package for nonnegative matrix factorization. BMC Bioinformatics 11.
Gaujoux, R., Seoighe, C., 2018. NMF: Algorithms and Framework for Nonnegative Matrix Factorization (NMF). R package version 0.21.0. https://cran.r-project.org/web/packages/NMF/.
Grolemond, G., Wickman, H., 2011. Dates and times made easy with lubridate. J. Stat. Softw. 40, 1–25.
Gross, J., Hamed, A., Sonntag, A., Spindler, G., Manninen, H.E., Nieminen, T., et al., 2018. Atmospheric new particle formation at the research station Melpitz, Germany: connection with gaseous precursors and meteorological parameters. Atmos. Chem. Phys. 18, 1835–1861.
Gu, J.W., Pitz, M., Schnelle-Kreis, J., Diemer, J., Reller, A., Zimmermann, R., et al., 2011. Source apportionment of ambient particles: comparison of positive matrix factorization analysis applied to particle size distribution and chemical composition data. Atmos. Environ. 45, 1849–1857.
Hameri, K., Hussein, T., Kulmala, M., Aalto, P., 2004. Measurements of fine and ultrafine particles in Helsinki: connection between outdoor and indoor air quality. Boreal Environ. Res. 9, 459–467.
Harrison, R.M., Shi, J.P., Xi, S.H., Khan, A., Mark, D., Kinnersley, R., et al., 2000. Measurement of number, mass and size distribution of particles in the atmosphere. Philosophical Transactions of the Royal Society of London Series a-Mathematical Physical and Engineering Sciences 358, 2567–2579.
Harrison, R.M., Beddows, D.C.S., Dall'Osto, M., 2011. PMF analysis of wide-range particle size spectra collected on a major highway. Environmental Science & Technology 45, 5522–5528.
Hartigan, J.A., Wong, M.A., 1979. Algorithm AS 136: a k-means clustering algorithm. J. R. Stat. Soc.: Ser. C: Appl. Stat. 28, 100–108.
Heintzenberg, J., Wehner, B., Birmili, W., 2007. 'How to find bananas in the atmospheric aerosol': new approach for analyzing atmospheric nucleation and growth events. Tellus B: Chemical and Physical Meteorology 59, 273–282.
Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. 24, 417–441.
Hou, L.-Q., 2018. Beijing Releases List of Top Air Polluting Sources. Beijing Municipal Environmental Monitoring Center http://www.chinadaily.com.cn/a/201805/15/WS5afa22eca3103f6866ee8561.html 2018-05-15.
Hussein, T., Hameri, K.A., Aalto, P.P., Paatero, P., Kulmala, M., 2005. Modal structure and spatial-temporal variations of urban and suburban aerosols in Helsinki - Finland. Atmos. Environ. 39, 1655–1668.

Hussein, T., Molgaard, B., Hannuniemi, H., Martikainen, J., Jarvi, L., Wegner, T., et al., 2014. Fingerprints of the urban particle number size distribution in Helsinki, Finland: local versus regional characteristics. Boreal Environ. Res. 19, 1–20.
Hussein, T., Atashi, N., Sogacheva, L., Hakala, S., Dada, L., Petäjä, T., et al., 2020. Characterization of urban new particle formation in Amman—Jordan. Atmosphere 11, 79.
Iannone, R., 2018. SplitR: use the HYSPLIT model inside R and do more with it. R package version 0.4. https://github.com/lhenneman/SplitR.
Joutsensaari, J., Ozon, M., Nieminen, T., Mikkonen, S., Lahivaara, T., Decesari, S., et al., 2018. Identification of new particle formation events with deep learning. Atmos. Chem. Phys. 18, 9597–9615.
Kabacoff, R.I., 2015. R in Action: Data Analysis and Graphics With R 2nd Edition. Manning Publications, Shelter Island, NY.
Kasumba, J., Hopke, P.K., Chalupa, D.C., Utell, M.J., 2009. Comparison of sources of submicron particle number concentrations measured at two sites in Rochester, NY. Sci. Total Environ. 407, 5071–5084.
Khan, M.F., Latif, M.T., Amil, N., Juneng, L., Mohamad, N., Nadzir, M.S.M., et al., 2015. Characterization and source apportionment of particle number concentration at a semi-urban tropical environment. Environ. Sci. Pollut. Res. 22, 13111–13126.
Kim, E., Hopke, P.K., Larson, T.V., Covert, D.S., 2004. Analysis of ambient particle size distributions using unmix and positive matrix factorization. Environmental Science & Technology 38, 202–209.
Kittelson, D.B., 1998. Engines and nanoparticles: a review. J. Aerosol Sci. 29, 575–588.
Krecl, P., Larsson, E.H., Strom, J., Johansson, C., 2008. Contribution of residential wood combustion and other sources to hourly winter aerosol in Northern Sweden determined by positive matrix factorization. Atmos. Chem. Phys. 8, 3639–3653.
Krecl, P., Targino, A.C., Johansson, C., Strom, J., 2015. Characterisation and source apportionment of submicron particle number size distributions in a busy street canyon. Aerosol Air Qual. Res. 15, 220–233.
Kulkarni, P., Baron, P.A., Willeke, K., 2011. Aerosol Measurement: Principles, Techniques, and Applications. John Wiley & Sons.
Kulmala, M., Vehkamaki, H., Petaja, T., Dal Maso, M., Lauri, A., Kerminen, V.M., et al., 2004. Formation and growth rates of ultrafine atmospheric particles: a review of observations. J. Aerosol Sci. 35, 143–176.
Kumar, P., Morawska, L., Birmili, W., Paasonen, P., Hu, M., Kulmala, M., et al., 2014. Ultrafine particles in cities. Environ. Int. 66, 1–10.
Lee, D.D., Seung, H.S., 1999. Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791.
Li, H.Y., Duan, F.K., He, K.B., Ma, Y.L., Kimoto, T., Huang, T., 2016. Size-dependent characterization of atmospheric particles during winter in Beijing. Atmosphere 7.
Li, H.Y., Cheng, J., Zhang, Q., Zheng, B., Zhang, Y.X., Zheng, G.J., et al., 2019a. Rapid transition in winter aerosol composition in Beijing from 2014 to 2017: response to clean air actions. Atmos. Chem. Phys. 19, 11485–11499.
Li, Z., Wang, Y., Guo, J., Zhao, C., Cribb, M.C., Dong, X., et al., 2019b. East Asian Study of Tropospheric Aerosols and their Impact on Regional Clouds, Precipitation, and Climate (EAST-AIRCPC). vol. 124 pp. 13026–13054.
Liang, C.S., Yu, T.Y., Chang, Y.Y., Syu, J.Y., Lin, W.Y., 2013. Source apportionment of PM2.5 particle composition and submicrometer size distribution during an Asian dust storm and non-dust storm in Taipei. Aerosol Air Qual. Res. 13, 545–554.
Liang, C.-S., Duan, F.-K., He, K.-B., Ma, Y.-L., 2016a. Review on recent progress in observations, source identifications and countermeasures of PM2.5. Environ. Int. 86, 150–170.
Liang, C.-S., Liu, H., He, K.-B., Ma, Y.-L., 2016b. Assessment of regional air quality by a concentration-dependent Pollution Permeation Index. Sci. Rep. 6, 34891.
Lianou, M., Chalbot, M.C., Kotronarou, A., Kavouras, I.G., Karakatsani, A., Katsouyanni, K., et al., 2007. Dependence of home outdoor particulate mass and number concentrations on residential and traffic features in urban areas. J. Air Waste Manage. Assoc. 57, 1507–1517.
Liu, S., Hu, M., Wu, Z.J., Wehner, B., Wiedensohler, A., Cheng, Y.F., 2008. Aerosol number size distribution and new particle formation at a rural/coastal site in Pearl River Delta (PRD) of China. Atmos. Environ. 42, 6275–6283.
Liu, Z.R., Hu, B., Liu, Q., Sun, Y., Wang, Y.S., 2014. Source apportionment of urban fine particle number concentration during summertime in Beijing. Atmos. Environ. 96, 359–369.
Liu, H., Qi, L., Liang, C., Deng, F., Man, H., He, K., 2020. How aging process changes characteristics of vehicle emissions? A review. Crit. Rev. Environ. Sci. Technol. 50, 1796–1828.
Liu, Z.R., Wang, Y.S., Hu, B., Ji, D.S., Zhang, J.K., Wu, F.K., et al., 2016. Source appointment of fine particle number and volume concentration during severe haze pollution in Beijing in January 2013. Environ. Sci. Pollut. Res. 23, 6845–6860.
Liu, P.F., Zhang, C.L., Xue, C.Y., Mu, Y.J., Liu, J.F., Zhang, Y.Y., et al., 2017a. The contribution of residential coal combustion to atmospheric PM2.5 in northern China during winter. Atmos. Chem. Phys. 17, 11503–11520.
Liu, Z., Hu, B., Zhang, J., Xin, J., Wu, F., Gao, W., et al., 2017b. Characterization of fine particles during the 2014 Asia-Pacific economic cooperation summit: number concentration, size distribution and sources. Tellus Series B-Chemical and Physical Meteorology 69, 19.
Lloyd, S.P., 1982. Least squares quantization in PCM. IEEE Trans. Inf. Theory 28, 129–137.
MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. vol. 1, pp. 281–297 Oakland, CA, USA.
Masiol, M., Vu, T.V., Beddows, D.C.S., Harrison, R.M., 2016. Source apportionment of wide range particle size spectra and black carbon collected at the airport of Venice (Italy). Atmos. Environ. 139, 56–74.
Masiol, M., Harrison, R.M., Vu, T.V., Beddows, D.C.S., 2017a. Sources of sub-micrometre particles near a major international airport. Atmos. Chem. Phys. 17, 12379–12403.
Masiol, M., Hopke, P.K., Felton, H.D., Frank, B.P., Rattigan, O.V., Wurth, M.J., et al., 2017b. Source apportionment of PM2.5 chemically speciated mass and particle number concentrations in New York City. Atmos. Environ. 148, 215–229.
Mcelroy, M.W., Carr, R.C., Ensor, D.S., Markowski, G.R., 1982. Size distribution of fine particles from coal combustion. Science 215, 13–19.
Meier, J., Wehner, B., Massling, A., Birmili, W., Nowak, A., Gnauk, T., et al., 2009. Hygroscopic growth of urban aerosol particles in Beijing (China) during wintertime: a comparison of three experimental methods. Atmos. Chem. Phys. 9, 6865–6880.
Meng, X., Ma, Y.J., Chen, R.J., Zhou, Z.J., Chen, B.H., Kan, H.D., 2013. Size-fractionated particle number concentrations and daily mortality in a Chinese City. Environ. Health Perspect. 121, 1174–1178.
Mertens, J., Lepaumier, H., Rogiers, P., Desagher, D., Goossens, L., Duterque, A., et al., 2020. Fine and ultrafine particle number and size measurements from industrial combustion processes: primary emissions field data. Atmos. Pollut. Res. 11, 803–814.
Morawska, L., Zhang, J.F., 2002. Combustion sources of particles. 1. Health relevance and source signatures. Chemosphere 49, 1045–1058.
Morawska, L., Thomas, S., Gilbert, D., Greenaway, C., Rijnders, E., 1999. A study of the horizontal and vertical profile of submicrometer particles in relation to a busy road. Atmos. Environ. 33, 1261–1274.
Morissette, L., Chartier, S., 2013. The k-means clustering technique: general considerations and implementation in Mathematica. Tutorials in Quantitative Methods for Psychology 9, 15–24.
Müller, K., Wickham, H., 2019. tibble: simple data frames. R package version 2.1.3. https://CRAN.R-project.org/package=tibble.
Nie, W., Ding, A.J., Wang, T., Kerminen, V.M., George, C., Xue, L.K., et al., 2014. Polluted dust promotes new particle formation and growth. Sci. Rep. 4.
Norris, G., Duvall, R., Brown, S., Bai, S., 2014. EPA Positive Matrix Factorization (PMF) 5.0 Fundamentals and User Guide. U.S. Environmental Protection Agency & Sonoma Technology, Inc.
Oberdorster, G., Oberdorster, E., Oberdorster, J., 2005. Nanotoxicology: an emerging discipline evolving from studies of ultrafine particles. Environ. Health Perspect. 113, 823–839.
Ogulei, D., Hopke, P.K., Zhou, L.M., Pancras, J.P., Nair, N., Ondov, J.M., 2006. Source apportionment of Baltimore aerosol from combined size distribution and chemical composition data. Atmos. Environ. 40, S396–S410.
Ogulei, D., Hopke, P.K., Chalupa, D.C., Utell, M.J., 2007. Modeling source contributions to submicron particle number concentrations measured in Rochester, New York. Aerosol Sci. Technol. 41, 179–201.
Paatero, P., 1997. Least squares formulation of robust non-negative factor analysis. Chemom. Intell. Lab. Syst. 37, 23–35.
Paatero, P., Tapper, U., 1994. Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. Environmetrics 5, 111–126.
Pearson, K., 1901. LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2, 559–572.
Pey, J., Querol, X., Alastuey, A., Rodriguez, S., Putaud, J.P., Van Dingenen, R., 2009. Source apportionment of urban fine and ultra-fine particle number concentration in a Western Mediterranean city. Atmos. Environ. 43, 4407–4415.
Pey, J., Alastuey, A., Querol, X., Rodriguez, S., 2010. Monitoring of sources and atmospheric processes controlling air quality in an urban Mediterranean environment. Atmos. Environ. 44, 4879–4890.
Price, H.D., Arthur, R., BeruBe, K.A., Jones, T.P., 2014. Linking particle number concentration (PNC), meteorology and traffic variables in a UK street canyon. Atmos. Res. 147, 133–144.
Rao, C.R., 1964. The use and interpretation of principal component analysis in applied research. Sankhyā: The Indian Journal of Statistics, Series A 329–358.
Revelle, W., 2020. psych: procedures for personality and psychological research. R package version 1.9.12.31. https://CRAN.R-project.org/package=psych.
Rivas, I., Beddows, D.C.S., Amato, F., Green, D.C., Jarvi, L., Hueglin, C., et al., 2020. Source apportionment of particle number size distribution in urban background and traffic stations in four European cities. Environ. Int. 135.
Robinson, D., Hayes, A., 2019. broom: convert statistical analysis objects into tidy Tibbles. R package version 0.5.3. https://CRAN.R-project.org/package=broom.
Rodriguez, S., Van Dingenen, R., Putaud, J.P., Dell'Acqua, A., Pey, J., Querol, X., et al., 2007. A study on the relationship between mass concentrations, chemistry and number size distribution of urban fine aerosols in Milan, Barcelona and London. Atmos. Chem. Phys. 7, 2217–2232.
Sarkar, D., 2008. Lattice: Multivariate Data Visualization With R. Springer, New York.
Sarkar, D., 2018. lattice: trellis graphics for R. R package version 0.20-38. https://cran.r-project.org/web/packages/lattice/.
Seinfeld, J.H., Pandis, S.N., 2006. Atmospheric Chemistry and Physics: From Air Pollution to Climate Change. John Wiley & Sons, Hoboken.
Sloane, C.S., Watson, J., Chow, J., Pritchett, L., Richards, L.W., 1991. Size-segregated fine particle measurements by chemical-species and their impact on visibility impairment in Denver. Atmospheric Environment Part a-General Topics 25, 1013–1024.
Slowikowski, K., 2019. ggrepel: automatically position non-overlapping text labels with 'ggplot2'. R package version 0.8.1. https://CRAN.R-project.org/package=ggrepel.
Sowlat, M.H., Hasheminassab, S., Sioutas, C., 2016. Source apportionment of ambient particle number concentrations in central Los Angeles using positive matrix factorization (PMF). Atmos. Chem. Phys. 16, 4849–4866.
Spearman, C.E., 1904. "General intelligence" objectively determined and measured. Am. J. Psychol. 15, 201–292.
Spinu, V., Grolemund, G., Wickham, H., 2018. lubridate: make dealing with dates a little easier. R package version 1.7.4. https://cran.r-project.org/web/packages/lubridate/.

Squizzato, S., Masiol, M., Emami, F., Chalupa, D.C., Utell, M.J., Rich, D.Q., et al., 2019. Long-term changes of source apportioned particle number concentrations in a metropolitan area of the northeastern United States. Atmosphere 10.

Stanier, C.O., Khlystov, A.Y., Pandis, S.N., 2004. Ambient aerosol size distributions and number concentrations measured during the Pittsburgh Air Quality Study (PAQS). Atmos. Environ. 38, 3275–3284.

Statheropoulos, M., Vassiliadis, N., Pappa, A., 1998. Principal component and canonical correlation analysis for examining air pollution and meteorological data. Atmos. Environ. 32, 1087–1095.

Sun, Y.W., Zhou, X.H., Wang, W.X., 2016. Aerosol size distributions during haze episodes in winter in Jinan, China. Particuology 28, 77–85.

Tan, J.H., Duan, J.C., Chai, F.H., He, K.B., Hao, J.M., 2014. Source apportionment of size segregated fine/ultrafine particle by PMF in Beijing. Atmos. Res. 139, 90–100.

Tan, Z.F., Rohrer, F., Lu, K.D., Ma, X.F., Bohn, B., Broch, S., et al., 2018. Wintertime photochemistry in Beijing: observations of ROx radical concentrations in the North China Plain during the BEST-ONE campaign. Atmos. Chem. Phys. 18, 12391–12411.

Teetor, P.R., 2011. Cookbook: Proven Recipes for Data Analysis, Statistics, and Graphics. O'Reilly Media, Inc.

The R Core Team, 2019. R: A Language and Environment for Statistical Computing. Version 3.6.2. R Foundation for Statistical Computing, Vienna, Austria https://www.R-project.org/.

Thimmaiah, D., Hovorka, J., Hopke, P.K., 2009. Source apportionment of winter submicron Prague aerosols from combined particle number size distribution and gaseous composition data. Aerosol Air Qual. Res. 9, 209–236.

Trojanowski, R., Fthenakis, V., 2019. Nanoparticle emissions from residential wood combustion: a critical literature review, characterization, and recommendations. Renew. Sustain. Energy Rev. 103, 515–528.

Tunved, P., Strom, J., Hansson, H.C., 2004. 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.

Ushey, K., 2018. RcppRoll: efficient rolling/windowed operations. R package version 0.3.0. https://CRAN.R-project.org/package=RcppRoll.

von Bismarck-Osten, C., Birmili, W., Ketzel, M., Massling, A., Petaja, T., Weber, S., 2013. Characterization of parameters influencing the spatio-temporal variability of urban particle number size distributions in four European cities. Atmos. Environ. 77, 415–429.

Vu, T.V., Delgado-Saborit, J.M., Harrison, R.M., 2015. Review: particle number size distributions from seven major sources and implications for source apportionment studies. Atmos. Environ. 122, 114–132.

Wåhlin, P., Palmgren, F., Van Dingenen, R., 2001. Experimental studies of ultrafine particles in streets and the relationship to traffic. Atmos. Environ. 35, S63–S69.

Wang, Z.B., Hu, M., Wu, Z.J., Yue, D.L., He, L.Y., Huang, X.F., et al., 2013. Long-term measurements of particle number size distributions and the relationships with air mass history and source apportionment in the summer of Beijing. Atmos. Chem. Phys. 13, 10159–10170.

Wang, Y.Y., Li, Z.Q., Zhang, R.Y., Jin, X.A., Xu, W.Q., Fan, X.X., et al., 2019. Distinct ultrafine- and accumulation-mode particle properties in clean and polluted urban environments. Geophys. Res. Lett. 46, 10918–10925.

Weber, S., Kuttler, W., Weber, K., 2006. Flow characteristics and particle mass and number concentration variability within a busy urban street canyon. Atmos. Environ. 40, 7565–7578.

Weber, S., Kordowski, K., Kuttler, W., 2013. Variability of particle number concentration and particle size dynamics in an urban street canyon under different meteorological conditions. Sci. Total Environ. 449, 102–114.

Wegner, T., Hussein, T., Hameri, 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.

Whitby, K.T., 1978. Physical characteristics of sulfur aerosols. Atmos. Environ. 12, 135–159.

Wichmann, H.E., Peters, A., 2000. Epidemiological evidence of the effects of ultrafine particle exposure. Philosophical Transactions of the Royal Society of London Series a-Mathematical Physical and Engineering Sciences 358, 2751–2768.

Wickham, H., 2007. Reshaping data with the reshape package. J. Stat. Softw. 21, 1–20.

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York.

Wickham, H., 2017. reshape2: flexibly reshape data: a reboot of the reshape package. R package version 1.4.3. https://cran.r-project.org/web/packages/reshape2/.

Wickham, H., 2019a. forcats: tools for working with categorical variables (factors). R package version 0.4.0. https://CRAN.R-project.org/package=forcats.

Wickham, H., 2019b. stringr: simple, consistent wrappers for common string operations. R package version 1.4.0. https://CRAN.R-project.org/package=stringr.

Wickham, H., Henry, L., 2019. tidyr: tidy messy data. R package version 1.0.0. https://CRAN.R-project.org/package=tidyr.

Wickham, H., Seidel, D., 2019. scales: scale functions for visualization. R package version 1.1.0. https://CRAN.R-project.org/package=scales.

Wickham, H., Chang, W., Henry, L., Pedersen, T.L., Takahashi, K., Wilke, C., et al., 2019a. ggplot2: create elegant data visualisations using the grammar of graphics. R package version 3.2.1. https://cran.r-project.org/web/packages/ggplot2/.

Wickham, H., Francois, R., Henry, L., Müller, K., 2019b. dplyr: a grammar of data manipulation. R package version 0.8.3. https://CRAN.R-project.org/package=dplyr.

Wiedensohler, A., Birmili, W., Nowak, A., Sonntag, A., Weinhold, K., Merkel, M., et al., 2012. Mobility particle size spectrometers: harmonization of technical standards and data structure to facilitate high quality long-term observations of atmospheric particle number size distributions. Atmospheric Measurement Techniques 5, 657–685.

Wilke, C.O., 2019. cowplot: streamlined plot theme and plot annotations for 'ggplot2'. R package version 1.0.0. https://CRAN.R-project.org/package=cowplot.

Wu, Z.J., Hu, M., Liu, S., Wehner, B., Bauer, S., Ssling, A.M., et al., 2007. New particle formation in Beijing, China: statistical analysis of a 1-year data set. J. Geophys. Res.-Atmos. 112.

Yu, G., 2019. ggplotify: convert plot to 'grob' or 'ggplot' object. R package version 0.0.4. https://CRAN.R-project.org/package=ggplotify.

Yue, W., Stolzel, M., Cyrys, J., Pitz, M., Heinrich, J., Kreyling, W.G., et al., 2008. Source apportionment of ambient fine particle size distribution using positive matrix factorization in Erfurt, Germany. Sci. Total Environ. 398, 133–144.

Zeileis, A., Grothendieck, G., 2005. zoo: S3 infrastructure for regular and irregular time series. J. Stat. Softw. 14, 1–27.

Zeileis, A., Grothendieck, G., Ryan, J.A., 2019. zoo: S3 infrastructure for regular and irregular time series (Z's ordered observations). R package version 1.8-6. https://cran.r-project.org/web/packages/zoo/.

Zhao, P.S., Du, X., Su, J., Ding, J., Dong, Q., 2020. Aerosol hygroscopicity based on size-resolved chemical compositions in Beijing. Sci. Total Environ. 716.

Zhou, L.M., Kim, E., Hopke, P.K., Stanier, C.O., Pandis, S., 2004. Advanced factor analysis on Pittsburgh particle size-distribution data. Aerosol Sci. Technol. 38, 118–132.

Zhou, L.M., Kim, E., Hopke, P.K., Stanier, C., Pandis, S.N., 2005. Mining airborne particulate size distribution data by positive matrix factorization. J. Geophys. Res.-Atmos. 110.

Zong, Y.C., Botero, M.L., Yu, L.Y.E., Kraft, M., 2019. Size spectra and source apportionment of fine particulates in tropical urban environment during southwest monsoon season. Environ. Pollut. 244, 477–485.