

Comparison of Two Approaches to Modeling Atmospheric Aerosol Particle Size Distributions

Vladimír Ždímal^{1*}, Marek Brabec^{2,3}, Zdeněk Wagner⁴

¹ *Laboratory of Aerosol Chemistry and Physics, Institute of Chemical Process Fundamentals of the AS CR, v. v. i., Rozvojová 135, Praha 6, 165 02, Czech Republic*

² *Department of Biostatistics and Informatics, National Institute of Public Health, Šrobárova 48, Praha 10, 100 42*

³ *Department of Nonlinear Modeling, Institute of Computer Science, Pod Vodárenskou věží 2, Praha 8, 182 07, Czech Republic*

⁴ *E. Hála Laboratory of Thermodynamics, Institute of Chemical Process Fundamentals of the AS CR, v. v. i., Rozvojová 135, Praha 6, 165 02, Czech Republic*

Abstract

This paper compares two approaches to modeling (smoothing) aerosol particle size distribution (particle counts for specified diameter intervals): i) the semiparametric approach based on a maximum likelihood fitting of lognormal (LN) mixtures at each time separately, followed by smoothing parameter tracks, ii) the nonparametric approach based on a kernel-like smoothing as an application of the gnostic theory of uncertain data. The specific advantages and disadvantages of both the semiparametric and nonparametric approaches are discussed and illustrated using real data containing a day-long time series of size spectra measurements.

Keywords: Particle size distribution; Lognormal mixture; Semiparametric modeling; Nonparametric modeling; Gnostic theory of uncertain data.

INTRODUCTION

The data describing the dynamics of particle size distribution (PSD) are inevitably complex and might be modeled statistically from

various perspectives. There are many potential approaches that differ in the level of sophistication, computational load, preliminary information requirements, or assumptions needed to justify them. However, we can think of two basic classes into which they can be categorized: nonparametric and semiparametric models.

*Corresponding author: Fax: +420-220920661,

Tel: +420-220390246

E-mail address: zdimal@icpf.cas.cz

The semiparametric models of our interest consist of two parts. The parametric part postulates a mathematical formula to describe PSD feasibly and uses it, together with the measurement error distribution specification, to fit the PSD at each measurement time (e. g., by maximum likelihood). The time dynamics of parameters are then smoothed nonparametrically. The model specification involves assumptions that always behave as a two-edged sword: they can improve efficiency if they are approximately correct, but can spoil the analysis if they are not plausible. In real data fitting, such a model cannot have too many parameters and will not follow every little detail of the data. Rather, it should be able to capture the most important features and skip the minor ones, in line with the philosophy embodied in the aphorism by George E. P. Box (Box and Draper, 1987): “All models are wrong. But some of them are useful.” On the other hand, nonparametric models try to escape the danger of possibly incorrect parametric model specifications by making as few assumptions as possible. Typically, only very basic notions are employed, mainly the smoothness of the fitted distribution. The resulting solution then makes a compromise between the quality of the fit and its smoothness. Such an approach is certainly appealing since it creates the impression that it represents a sort of ideal, fool-proof or automatic tool that does not require any substantial assumption-making and hence precludes preliminary thinking about model choice. Unfortunately, nothing is for free and the impression is not correct for several reasons.

Generally, the efficiency is less important (the approximately correct parsimonious parametric model can estimate things in a more efficient and stable way). More influential is the fact that the balance between smoothness and the goodness-of-fit is not easy to establish. A wrong setting of procedural details can easily lead to over/underfitting, which is usually not substantial for the overall performance of the smoother, but can be disastrous for the fit of the local details of substantial interest (e. g., local maxima size and location). When focusing on these details, special techniques and/or a lot of fine tuning might be required (Marron and Chaudhuri, 1998) that are not easy to implement for large spectral time series data, or a lot of personal experience with similar data is required (as well as some subjective judgment).

In this work we discuss the semi- and nonparametric model properties and illustrate their capabilities on the data from a real measurement campaign. As a representative of the semiparametric methodology we used a lognormal (LN) mixture for the description of the PSD, followed by a loess nonparametric smoother (to smooth the parameter dynamics). LN mixtures were selected because mixtures with few lognormal components (1, 2 or 3) have traditionally been used in aerosol research both for theoretical computations (Seinfeld and Pandis, 1998) and for practical data analysis (Makela *et al.*, 2000; Voutilainen and Kaipio, 2002). We chose a flexible gnostic smoother as a representative of purely nonparametric methods. It was used as an example of a technique that is similar to traditional kernel smoothing methods, but enjoys robustness and

interesting motivation built on first principles (Kovanic, 1986). Its scale parameter, which defines the bandwidth, is estimated in such a way that the entropy of the original data is equal to the entropy of the calculated distribution function. The smoothness of the gnostic distribution function is thus determined by the data adopting the rule “let the data speak for themselves.”

In this text, we will try to compare both approaches to show their weak and strong points. We will stress namely those features that the potential user should be aware of.

MATERIAL AND MEASUREMENT METHODS

We analyzed data obtained from one day of a measurement campaign that took place at the Finokalia Station on the island of Crete, Greece as a part of the EC-funded 5th FP project called SUB-AERO (Lazaridis *et al.*, 2006). The chosen subset of data was obtained using a Scanning Mobility Particle Sizer (SMPS, Model 3934, TSI Inc., USA). This aerosol spectrometer consists of two major parts. In the Electrostatic Classifier (Model EC 3071A), one narrow fraction of particles having the same electrical mobility is selected. This fraction then enters the Condensation Particle Counter (CPC 3022A) where particles grow by the condensation of n-butanol on their surface and are counted optically. The field strength in the EC changes continuously in order to scan the whole available mobility range. After the

scan is completed, a sophisticated program provides the aerosol number size distribution in the sample. In this experiment, each scan lasted for 90 seconds and was followed immediately by another scan. The data matrix has $T = 953$ spectra, measured on January 12, 2001 in equidistant time intervals (1.5 minutes apart) from 00:00:10 to 23:59:59. Each scan went through 103 mobility/size channels. At the t -th time we got a measurement of the cumulative particle count (N_{it}), meaning the count of particles with a diameter smaller than or equal to the i -th size limit, d_i (i.e., the count in the interval $\langle d_0, d_i \rangle$) $i = 1, 2, \dots$, where $d_0 = 7.23$ nm and the boundaries increased exponentially, $d_i = d_{i-1} \delta$ with $\delta = 1.0367$, $I = 103$. Consequently, the measured diameter range spanned the interval $\langle 7.23, 294.2 \rangle$ nm.

SEMIPARAMETRIC MODEL—LOGNORMAL MIXTURES

The PSD is modeled parametrically (as an LN mixture), separately at each measurement time. The mixture parameters are estimated via maximum likelihood. These rough estimates of the time tracks are then smoothed nonparametrically to obtain a clearer picture of trends and more or less local systematic changes in their values.

Model

The parametric model of the size distribution at time $t, t=1$, can be formulated as:

$$N_{it} = \left(A_t \sum_{k=1}^K p_{tk} \frac{\left[\Phi\left(\frac{\log(d_i) - \mu_{tk}}{\sigma_{tk}}\right) - \Phi\left(\frac{\log(d_0) - \mu_{tk}}{\sigma_{tk}}\right) \right]}{\left[\Phi\left(\frac{\log(d_i) - \mu_{tk}}{\sigma_{tk}}\right) - \Phi\left(\frac{\log(d_0) - \mu_{tk}}{\sigma_{tk}}\right) \right]} \right) \varepsilon_{it} \quad (1)$$

The modeled quantity (N_{it}) is the cumulative count of particles with a diameter smaller than or equal to the i -th size limit d_i , $i=1,2, \dots$, as provided by the measurement device (see Section 2 for details). The systematic part is then composed of an appropriately scaled sum of the lognormal cumulative distribution function (cdf) differences. $\Phi(\cdot)$ denotes the standard normal cdf. We use $\log(\cdot)$ for the natural (e -base) logarithm. The parametrization is such that the parameter A corresponds to the total particle number concentration in the scanned interval $\langle d_0, d_i \rangle$, and the sum gives the (normalized) size distribution within this interval. Each of the K terms in the sum corresponds to the contribution from the k -th truncated lognormal component of the mixture. This component has the weight p_{tk} (with obvious restrictions $\sum_k p_{tk} = 1$ and $0 \leq p_{tk} \leq 1$) and the parameters μ_{tk} , σ_{tk} , which describe the shape of the component. To assure identifiability, we assume (without loss of generality) that the components $k=1, \dots$ are ordered according to the μ -parameters in ascending order (so that $\mu_{tk} < \mu_{t,k+1}$). This mixture represents a systematic part of the model.

Further, there is a multiplicative error, ε_{it} capturing not only the measurement error, but

also local irregularities (e. g., wind gusts bringing air masses of unusual composition) and the lack of the lognormal mixture model fit. It is assumed to be lognormally distributed, i.e. $\varepsilon_{it} \sim LN(0, \sigma_t^2)$, independently across the t 's and i 's. One can explain this behavior, for example, by considering the fact that the error distribution should be skewed and have a positive support. In practical terms, it is helpful that the model (1) changes to the easy-to-handle homoscedastic case when the logarithm of both sides is taken [as an application of the TBS methodology, Carroll and Ruppert (1988)].

For a fixed K , the model parameters (that need to be estimated from the data to identify the model completely) are: A_t , $\sigma_{t,k}$, μ_{tk} , p_{tk} , $k=1, \dots$. According to suggestions from aerosol theory and previous practical work (Seinfeld and Pandis, 1998; Makela et al., 2000; Voutilainen and Kaipio, 2002), we fitted the lognormal mixtures with one, two, or three components ($K=1,2,3$), which we will denote by LN1, LN2, LN3, respectively. If the size spectra behave “reasonably”, each of the parameters should be a smooth function of time, so that for the observed times $t=1, \dots, =953$, it represents a time series. The “only” problem is that these series are not directly observable and need to be estimated. Estimations were done in two steps: i) by obtaining the rough time-by-time parameter estimates, via maximum likelihood, using the parametric

model (1) and the implied likelihood function, and: ii) by smoothing the parameter time-tracks via a flexible nonparametric smoother *loess* (Cleveland and Devlin, 1988). The two steps represent a semi-parametric approach (the size distribution is modeled parametrically, the time dynamics nonparametrically), in an attempt to combine the useful properties of both. The basic idea is that the available knowledge about the distributional shape should be used to improve the efficiency and restrict the distributional shape range to exclude wild shapes that might be otherwise suggested by the count distortion introduced; e. g., by the measurement error. On the other hand, very little is known about the time-dynamics of the spectra a priori, so that the time evolution should be specified as flexibly as possible.

Let us consider one of the tasks typical for aerosol spectra analysis, namely the search for the number and location of distributional modes. In the LN mixture model, the number of peaks arises as a consequence of the parameter configuration. It is important to note that it is *always less than or equal to* the number of LN components. This is because some components can be wide enough and located closely enough so that they blend together into one peak. In this sense, the parametric model is much more informative than the plain smoothing of observed counts. It can discover components drifting apart before they are distant enough to be visible on the surface.

Estimation

Model (1) can be fitted relatively easily via maximum likelihood for each time t and a fixed K . From a practical point of view, all that is required is the numerical maximization of the logarithm of the likelihood. We used quasi-Newton-Raphson here. Log likelihood behaves in a relatively decent way, as long as K is not large (which was our case with $1 \leq K \leq 3$ choices). For numerical and substantive reasons, we reparametrized the component proportions p_{ik} 's via a cumulative logit transformation (Agresti, 1990), so that obvious p_{ik} properties were automatically enforced.

Convergence was generally good when we used the following scheme to construct the (time-dependent) starting values. For LN1 we started from the observed count in $\langle d_0, d_1 \rangle$ for the initial A estimate, with the average and standard deviation of logarithms of the class interval centers (weighted by the time-averaged particle count observed in the particular class) for μ and σ . For LN2, we started from the LN1 results and added an additional component, whose location was suggested by the analysis of the residuals after the LN model. For LN3, we started from the LN2.

For a given K , we estimated the parameter tracks (i.e., smoothed them) by the robust variant of the *loess* smoother (Cleveland and Devlin, 1988). As it is quadratic only locally, its shape is allowed to change with time, so that the resulting track estimate is generally highly

nonlinear. The smooth fit represents a compromise between a perfect (but rough and inefficient) fit and smoothness. The “exchange rate” in this compromise is given by the span parameter. The span gives a proportion of the data that are used locally to estimate the value at a particular time (a larger span means more smoothness). We chose a span from 0.1-0.25 (after some experimentation and inspection of the quality and smoothness of the fit). The tracks were smoothed for each parameter separately. To get the \hat{A}_t estimate, we smoothed the logs of the original time-specific MLEs and exponentiated the results. The \hat{p}_{ik} s were obtained via renormalization of the smoothed estimates (so that they add to one, as they should). The other parameters were smoothed directly.

In the previous text, we proceeded as if K was known. Obviously, in practice K is not known and must be selected somehow. Since we restricted the values of K to $1 \leq K \leq 3$ and hence fitted lognormal mixtures with one, two and three components, we had to select one of them. To this end, we used a (crude) procedure mimicking the likelihood ratio test at the level of 0.10 [to alleviate the possible problems with the parameter space boundary (McCulloch and Searle, 2001)]. The decision scheme was as follows: i) if the test of LN2 vs. LN1 was not significant, we selected LN1, ii) if the test in i) was significant, we conducted a test of LN3 vs. LN2 and selected LN3 or LN2 when it was/was not significant, respectively. Clearly, such a procedure generally leads to a time-varying number of lognormal component’s estimates (K_t).

Although one might see the time-varying K_t as a realistic feature, the time-dynamics smoothing becomes *much* more complicated with the parameter space dimension changes. Even though there are elegant Bayesian approaches to this (Green, 1995), they tend to be quite complicated and computationally demanding even in much simpler set-ups than we encounter here. Therefore, we employed two simpler approaches. The first was to use the results of the largest model considered (LN3) and to smooth its parameters in a straightforward way. Admittedly, this might lead to occasional overfitting (when a less parsimonious model would be chosen by the selection procedure) and hence, a somewhat inefficient parameter estimation. Nevertheless, the extent of the problems should not be typically too large for two reasons: the parameter values are smoothed anyway, so that occasional excesses related to unstable estimation should be suppressed (as long as the LN3 model is approximately valid most of the time), and one is often not interested directly in the parameters but in their functions (like location and/or size of the peaks of the size spectra) which are much more stable than the parameters themselves.

When the interest is predominantly in peak location dynamics, our second approach can be utilized—that is time-by-time testing can be applied to select (time-varying) the dimensionality estimate K_t , then peak locations can be smoothed instead of parameters.

All the computations were done in S-plus (Venables and Ripley, 1994). For the

minimization of the negative log likelihood, we used the built-in function `nlinb`. A two-stage estimation (estimating the LN parameters time-by-time and then smoothing the estimates by a nonparametric smoother like `loess`) is just one way to get to the resulting estimates. It is certainly not the most efficient way, but it offers a relatively easily usable tool for processing masses of data. It is potentially more efficient to build time smoothing into the model explicitly; e. g., by postulating a time series model in parameters [e. g., in multivariate ARIMA style, or even more elegantly in a state-space framework, (Voutilainen and Kaipio, 2001)]. This is appealing from both the theoretical and the statistical point of view, but the approach certainly presents two obstacles: i) the computational complexity, ii) the need for a reasonable dynamic statistical model specification. We did not go to the development of the time series model (which would be rather specific for the Crete space-time location), because of not having enough aerosol-dynamics-related information on hand to be able to build such a model, and wanting to demonstrate mainly the difference between the totally nonparametric (section 4) and the more specific semiparametric modeling approaches, while simultaneously staying relatively general (so that the finding can be applied mode widely). One possible consequence might be that we lost some efficiency that the more-specific model could probably have gained. Nevertheless, we do not expect that the time series specification would increase the efficiency dramatically.

Results of the Semiparametric Model

When processing the real data, we replaced the occasional zero observed counts in the first several intervals by small numbers (by halves of the minimum nonzero count observed in that particular interval during the observed period). Since the spectrum for the first observed time ($t=1$) was rather aberrant, we excluded it from further analyses (and proceeded with a relabeled time index, running from $t=1, \dots, (n-1)$). Fig. 1 compares the results of the LME fit of LN1, LN2 and LN3 to the size spectrum obtained on January 12, 2001 at 00:42:09. The open circles correspond to $\frac{dN_t(d)}{d \log(d)}$ estimated internally by the measurement device at the size interval midpoints (by taking the derivatives of N_{it} s smoothed internally by the measurement device); i.e., to the quantities that we *did not* use for fitting. We used a raw (empirical) cumulative count N_t instead to avoid the arbitrary pre-smoothing step and the arbitrary placement to the interval midpoint where only the total interval count is available. As a consequence, the open circles and lines are only approximately comparable.

The plot illustrates how the fit improves as we go from LN1 (which is forced to have one peak only, so that it tries to do its best by fitting the main peak and then seriously distorting the lower and upper end of the distribution), to LN2 (which still misses the lower peak even though it can theoretically fit two peaks—because both available components are spent to describe the complicated main peak shape) and LN3 (which picks both peaks fairly

well, while smoothing the peak densities a little bit due to the inherent measurement variability). Note also that the fitted lines correspond to median responses and not to their expected values (the two are quite different quantities under the lognormal distribution of ε_{it} 's in 1).

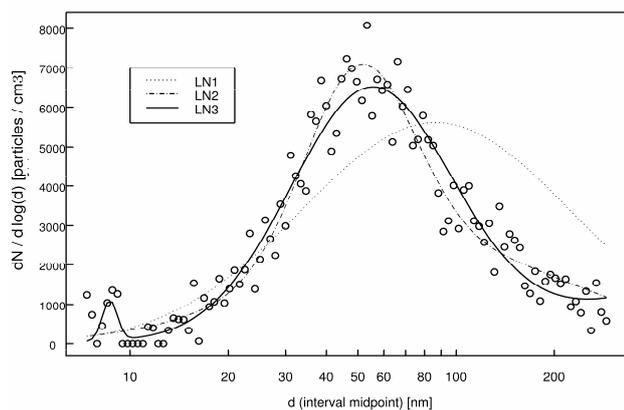


Fig. 1. The fits of a particle size distribution obtained on January 12, 2001 at 00:42:09 by the semiparametric model with one, two and three LN components (lines), compared with the data (points).

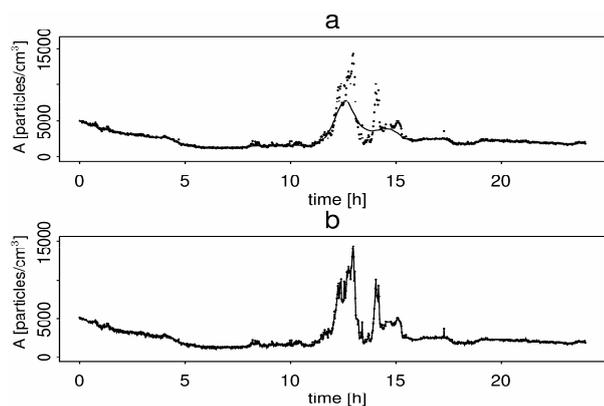


Fig. 2. The time track of the total particle number concentration \hat{A}_τ estimated using the LN3 model. Dots represent the measured points in both cases; the curve represents the LN3 model fit: a) after smoothing, and b) before smoothing with the loess smoother.

Fig. 2 illustrates the dynamics of the fitted total particle number concentrations (\hat{A}_τ 's) in the monitored diameter interval $\langle d_0, d_I \rangle$, estimated from the LN3 model. τ is measured in hours (0-24) and corresponds to the actual measurement time. Fig. 2(a) shows the situation after smoothing implied by the nonparametric part of the model, while 2(b) shows results of time-by-time LN3 estimation without any smoothing of parameters along the time line.

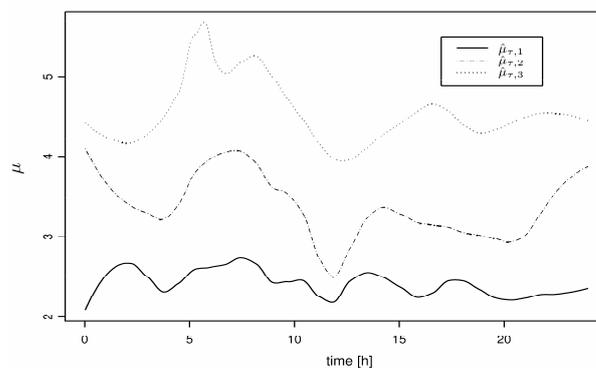


Fig 3. LN3, the smoothed time tracks of the $\hat{\mu}_{\tau,1}$, $\mu_{\tau,2}$, $\hat{\mu}_{\tau,3}$ estimates.

Fig. 3 compares the smoothed estimates of the $\mu_{\tau,k}$, $k=1,2,3$ parameters. Notice that the smoothed estimates respect the $\hat{\mu}_{\tau,k} < \mu_{\tau,k+1}$ restriction (as they should). $\hat{\mu}_{\tau,1}$ tends to remain low and pretty stable over time. There are much more dynamics in $\mu_{\tau,2}$, $\mu_{\tau,3}$ behavior. All three components show a quite abrupt change of the behavior close to 11 a.m. Even this observation alone might offer interesting insights into the aerosol size distribution evolution. Notice however, that $\hat{\mu}_{\tau,k}$ s

obviously do not represent estimates of the spectrum peak locations on the log scale. In fact, there can easily be a smaller number of peaks than 3 even if the fitted LN3 is non-trivial. This is because the peak locations depend in a complicated way on relative location of $\hat{\mu}_{\tau,k}$ s as well as on other parameters. We will discuss peak locations later.

From $\hat{p}_{\tau,k}$ comparison, we can see that in terms of particle count it is representative, the first component is rather “small.” Nevertheless, its behavior might be of utmost interest (e.g., in connection with the nucleation mode development). The (almost) inverse relation between $\hat{p}_{\tau,2}$ and $\hat{p}_{\tau,3}$ is obviously given by the sum-to-one restriction (and the fact that the $\hat{p}_{\tau,1}$ is small).

The other possible smoothing strategy mentioned in the section Estimation is to fit LN1, LN2, LN3 and to decide for one of them by an LRT-like testing, compute local maxima for the selected mixture and finally smooth the local maxima. One would expect that this approach should lead to more parsimonious time-dependent K_{τ} . The potential gains in efficiency (achieved by sparing a few parameters) might be offset by: i) the random nature of the selection procedure (the test necessarily commits both type I and type II errors), ii) occasional jumps in maxima locations before smoothing (induced by dimensionality (K_{τ}) changes), and hence it is not entirely clear whether it really pays off to go with the more difficult-to-handle model. The percentages of times when LN1, LN2 or

LN3 was selected by the test-based procedure were 6.4 %, 18.9% and 74.7 %, respectively. In other words, the LN3 prevails (so that the previous approach with always fitting LN3 should not be too bad); but the percentage of the simpler mixtures is not totally negligible—providing some motivation for time-varying K_{τ} attempts. Not only that the improvement provided by going from LN2-LN3 is not always the same, but also that it is distributed unevenly along the time axis. Greatest improvements occur around 11 and 15 hours.

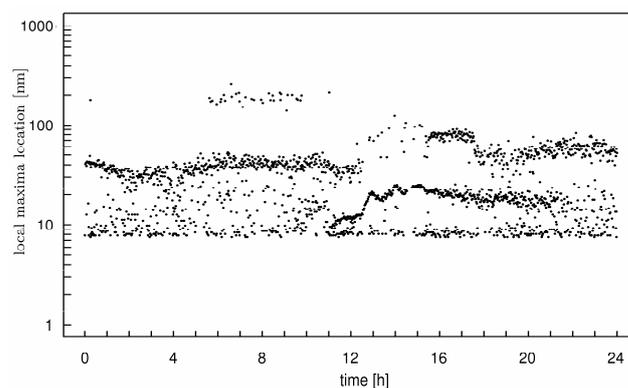


Fig. 4. The locations of non-smoothed local maxima on particle distribution functions and their evolution in time (as determined by the semiparametric model).

When we take all the smoothed parameters together, we can easily use them to derive other quantities of practical interest. The position of local maximum might be one of them. Many others can be computed just as easily; for example, local maxima sizes. But even subtler ones, like the locations of maxima of the first derivative of the spectral density, can be computed as well. When we selected one of the

fitted models by the testing procedure described earlier and found the local peaks (numerically), we plotted their rough (not-smoothed) peak locations in Fig. 4. Specifically, the number of peaks is not forced to be 3 (in fact, it varies between 1, 2 and 3).

Resulting local maxima of the regression function are plotted separately for each time. The proportions of the times when one, two or three peaks were found in the fitted are 18.9%, 43.1%, 37.9%, respectively. Note also that even this relatively easy procedure clearly visualized modal changes occurring in the late morning (around 11 a. m.). These changes are connected to a complicated transitional phenomenon, so called “particle formation event” (see also section Results of the Nonparametric Model), which is sometimes observed in atmospheric aerosol sampling campaigns. It is not the aim of this paper to deal with its mechanism in detail. However, a widely accepted explanation says that there are some thermodynamically stable clusters (TSC) present in the atmosphere. As the size of these clusters is under the lower detection limits of most aerosol spectrometers, they are practically “invisible.” Under favorable conditions, e. g. when a sufficient amount of condensable vapors is available in the atmosphere, these vapors condense on the cluster's surface causing its growth. When the cluster becomes visible for the spectrometer, it is detected. On the measured spectra, it looks like a new mode coming from the lowest sizes and increasing with time. Various parameters may then be extracted from its time behavior; e. g., the growth rate, production rate of the

condensable vapor, and so on (Kulmala *et al.*, 2004).

NONPARAMETRIC MODEL

The PSD is modeled by means of tools taken from the gnostic theory of uncertain data developed by Kovanic (1986). Over the years of development, the gnostic theory has grown into a set of tools, each of which finds its application in some branch of data analysis. The tools differ in the kind of robustness. Tools taking advantage of the quantifying distribution function possess the robustness with respect to inliers (outer robustness) and are useful for signal processing and similar applications.

The opposite case; i.e., the robustness with respect to outliers (inner robustness) is a feature of the tools based upon estimating distribution functions. Two kinds of estimating distribution functions are available, a global and a local one.

The global estimating distribution function is by definition unimodal. It can be used to verify the homogeneity of a data sample or to describe properties of such a sample. However, the PSD of atmospheric aerosols is usually multimodal, the global estimating distribution function thus cannot be used. The local estimating distribution function is a kernel estimate and can therefore describe multimodal distributions. In the following text, we will use the local estimating distribution function only.

Summary of the gnostic theory of uncertain data

In this section we summarize the most important features of the gnostic theory of uncertain data. We limit explanation to the local estimating distribution function. This work extends the methodology of using the local estimating distribution function so that the optimal bounds of a finite data support can be estimated. This will be explained in the section Data Support

The modeled quantity is the number of particles of the particular size class. We will model both variations of the particle number in the time domain, as well as the PSD at a specific time. The equations will therefore be given in an abstract way and both types of usage will be explained later.

The first axiom of the gnostic theory states that the measured value a can be expressed as a sum of an ideal value a_0 , which is the mathematical model of the true value, and uncertainty, which is a product of a scale parameter S and normalized uncertainty Φ :

$$a = a_0 + S\Phi \tag{2}$$

Using transformations

$$z = \exp(a), \quad z_0 = \exp(a_0) \tag{3}$$

a multiplicative model is obtained:

$$z^{1/S} = z_0^{1/S} \exp(\Phi) \tag{4}$$

Since conversion between the additive and the multiplicative model is straightforward, all

the following equations will be written for the multiplicative model only.

The basic properties are defined for each individual observation and the equations have been derived by Kovanic (1986) from the above-mentioned axiom. Having the observation z , we can calculate the probability of an ideal value being less or equal to z_0 :

$$p(z_0) = \frac{1-h}{2} \tag{5}$$

where h is irrelevance defined as

$$h = \frac{q^2 - 1/q^2}{q^2 + 1/q^2} \tag{6}$$

$$\text{and } q = \left(\frac{z}{z_0} \right)^{1/S} \tag{7}$$

After substituting z and z_0 into Eq. (5) and differentiating, we get the density

$$\frac{dp}{d \log z_0} = \frac{4}{S} \left[\left(\frac{z}{z_0} \right)^{2/S} + \left(\frac{z}{z_0} \right)^{-2/S} \right]^{-2} \tag{8}$$

It can be proven that Eq. (8) satisfies all conditions required for Parzen's kernels (Parzen, 1962). The shape of the kernel is determined by the value of the scale parameter. If $S \rightarrow 0$, the kernel converges to a δ -function. Entropy can be calculated both for the data sample z_k , $k=1, \dots$, and the smoothed kernel estimate. We will select such value of S which yields equal entropy in both cases. The value of the scale parameter is thus

obtained by solving the following equation:

$$\frac{\sin(\pi S / 2)}{\pi S / 2} = \frac{\sum_{k=1}^K W_k f_k}{\sum_{k=1}^K W_k} \quad (9)$$

where W_k are apriori weights and f_k are gnostic weights defined by equation

$$f = \frac{2}{q^2 + 1/q^2} \quad (10)$$

evaluated using $S=1$ for each particular $z = z_k$. It can be shown that there always exists a unique solution $0 \leq S \leq 2$. The exact meaning of the apriori weights W_k as well as the z_k values, depends on the application of the equation and will be explained later.

Data support

In practical modeling, the domain of the distribution functions is often taken to be the set of all real numbers R^1 or the set of positive real numbers R_+ . However, these sets represent mathematical abstractions. The physical quantities do not reach infinite values and are often bounded. In some cases the bounds are known in advance. In other cases we only know that the bounds exist, but their values are unknown. This is the case of PSD. The smallest particles must be considerably larger than molecules. Very large particles cannot survive in aerosol because of fast sedimentation. The actual bounds usually depend on a great many unknown factors. We thus intend to obtain such values of bounds which best describe the data.

The equations above are, however, derived for variables whose domains are R^1 or R_+ ,

respectively. We will apply the following transformation:

$$z = \frac{z' - Z_L}{1 - z' / Z_U} \quad (11)$$

where $z \in R_+$, $Z_L < z' < Z_U$. (12)

and Z_L, Z_U are the lower and upper bounds of the finite data support, respectively.

Procedure

The particle numbers in each size class are determined from the raw cumulative count reported by the measuring device. These values may be influenced by gross measurement errors, which may cause problems when estimating the scale parameter. Other quantities determined from the distribution function, such as the positions of local extrema and concentration of particles in a particular size range would be incorrect. The scale parameter is not directly connected to any physical quantity. Its smoothness could be expected but not justified by any natural law. Construction of the gnostic distribution function requires not only the value of the scale parameter, but also the original raw counts. A filtration technique applied to the scale parameter will therefore not solve the problem either. On the contrary, the raw cumulative counts represent the direct physical measurement. If there is no nearby source of particles, the concentration changes smoothly with time, and rapidly occurring variations are caused by various kinds of measurement errors. The cumulative counts are therefore first filtered in the time domain.

One of the simplest smoothers is the application of an average in a moving window. We use a similar gnostic smoother. The modeled variable is the particle number of a particular size class. Since the particle number is allowed to be zero, we make use of the additive model, Eq. (2) and transform it using Eq. (11) into the multiplicative model, Eq. (4). The moving window is formed by five consecutive values. These values are used in Eqs. (10) and (9) in order to calculate the scale parameter. Quantity z_0 in Eq. (7) is the particle number at the end of the moving window transformed to the multiplicative data support. All weights W_i in Eq. (9) are set to 1. The scale parameter enables us to construct a gnostic distribution function of particle numbers and the quantile for probability equal to 0.5 (gnostic median) is used as a filtered value. This procedure automatically removes outliers. They could be marked as such by further analysis of the differences between the original counts and the filtered values, but it was not done because such information is not important for the determination of a distribution function.

The PSD function is then determined at each time t using the filtered values. The determination of the distribution function consists in an estimation of the scale parameter and the bounds of the data support. The agreement of the gnostic estimation local distribution function with the empirical distribution function is not a good criterion. If the scale parameter is changed and the distribution function possesses different number of modes, the differences between the calculated and empirical distribution functions

change only slightly. Changes of much-greater magnitude can be found by comparing the calculated and empirical distribution densities. We therefore decided to find the optimum parameters by minimizing the maximum difference between the calculated and empirical distribution density.

The filtered values N_{kt} represent the particle numbers of size class d_k , $k = 1, \dots, K$ at time t . The weighted empirical distribution function is constructed as

$$p_{it} = \frac{\sum_{k=1}^i N_{kt}}{\sum_{k=1}^K N_{kt} + (N_{1t} + N_{Kt}) / 2} \quad (13)$$

Such definition allows for probability $p(d > d_K) > 0$, as well as $p(d < d_1) > 0$.

The empirical density is obtained by numerical differentiation. A polynomial of the third order fitted to seven neighboring points is used in order to suppress the oscillations caused by uncertainties in the experimental data.

The bounds of the data support and the scale parameter are obtained by minimizing

$$L = \max_d \left| \Delta \frac{dp}{d \log d} \right| \quad (14)$$

where $|\Delta(dp / d \log d)|$ is the absolute value of the difference between the weighted empirical density and the density evaluated from the local estimating distribution function. For each approximation of the bounds Z_L and Z_U , the particle sizes d_k are transformed to the infinite data support using Eq. (11). The scale

parameter is calculated from Eq. (9). The filtered particle numbers N_{kt} are used in place of the apriori weights W_k and the gnostic weights f_k are evaluated from the particle sizes transformed into the infinite data support. Variable z_0 required in Eq. (10) is the weighted median, the definition of which is given in the following paragraph.

Let p_i , $i=1, \dots, N$ be the probabilities defined by the weighted empirical distribution function, Eq. (13). Let P be a given value such that $p_1 \leq P \leq p_N$. Then let p_r be the greatest value satisfying $p_r \leq P$ and p_s be the least value satisfying $P \leq p_s$. The quantile for probability P is defined as a value obtained by inverse linear interpolation using p_r and p_s . The weighted median is defined as the weighted quantile for $P = 0.5$.

The definition does not allow for an evaluation of the weighted quantiles for $P < p_1$ and $P > p_N$. We must be able to calculate weighted quantiles before the bounds of the data support are determined. Extrapolation below and above the experimental data is therefore impossible.

Before the weighted median can be used in Eq. (10), it must be converted into the infinite data support by means of Eq. (11).

The objective function, Eq. (14) may possess several local minima. In order to determine the global minimum, the generalized random search with alternating heuristics (GCRS/ALTH) developed by Tvrdík *et al.* is used (Krivý and Tvrdík, 1995; Tvrdík and Krivý, 1999; Tvrdík *et al.*, 2001). The MatLab implementation of the algorithm is available from the author's website (Tvrdík) and the

function can also be used with Octave (Octave).

For mostly numerical reasons Eq. (11) is not used directly. We first normalize the values of particle diameters:

$$z' = \exp \left[2 \frac{\log(d / d_{min})}{\log(d_{max} / d_{min})} - 1 \right] \quad (15)$$

where d_{min} and d_{max} are the minimum, and maximum values of the particle diameter d and z' is the normalized multiplicative value such that $1/e \leq z' \leq e$. We now find optimum values of the bounds in the normalized data support. The GCRS/ALTH algorithm requires limits of the box type. The limits, such as $10^{-6} \leq Z_L \leq e^{-1.1}$ and $e^{1.1} \leq Z_U \leq 10^6$, are wide enough so that the optimum value is not missed. Bounds d_L and d_U are then evaluated from Eq. (15).

The $Z_L - Z_U$ space extends over 12 orders of magnitude and is asymmetric. Its metrics prefers Z_L close to data and Z_U distant from data. We therefore perform minimization for transformed values:

$$X_L = 2 \frac{\log(Z_L / 10^{-6})}{\log(e^{-1.1} / 10^{-6})} - 1 \quad X_U = 2 \frac{\log(Z_U / e^{1.1})}{\log(10^6 / e^{1.1})} - 1 \quad (16)$$

Results of the Nonparametric Model

The gnostic filter applied to the time series of particle counts is very simple. We prefer a fast response to the quality of filtration. The main purpose is removal of measurements that are evidently wrong and replacing them by a

smoothed value. Filter performance is presented in Fig. 5. You can see that the line corresponding to the filtered values slightly reduces the scatter of the experimental data while following the main trend of the time series. Two outliers were removed.

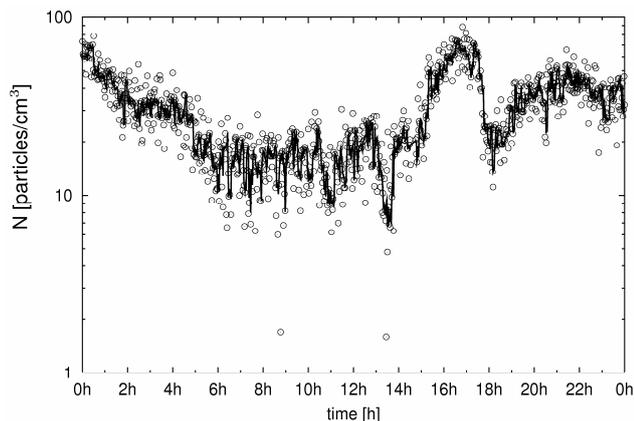


Fig. 5. Example: the time series of the concentration of particles in the size bin 109.4 nm; the measured values (points) and a smoothed line estimated by the gnostic filter.

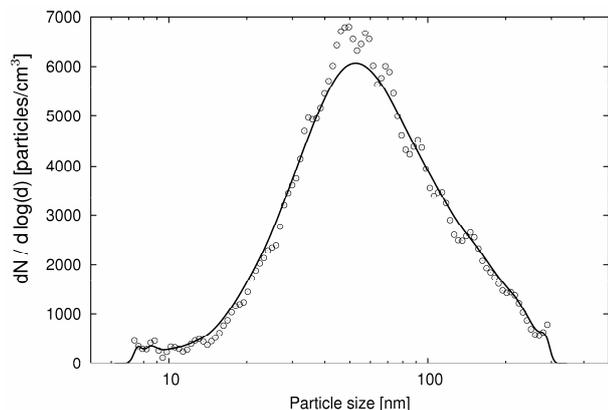


Fig. 6. A fit of a particle size distribution obtained on January 12, 2001 at 00:42:09 by the nonparametric model (line), compared with the time-filtered data (points).

In this work we used only a distribution function with a constant scale parameter where

z_0 needed in Eq. (10) is obtained as a weighted median converted to the infinite data support. The typical distribution is depicted in Fig. 6. It can be seen that small peaks were smoothed out.

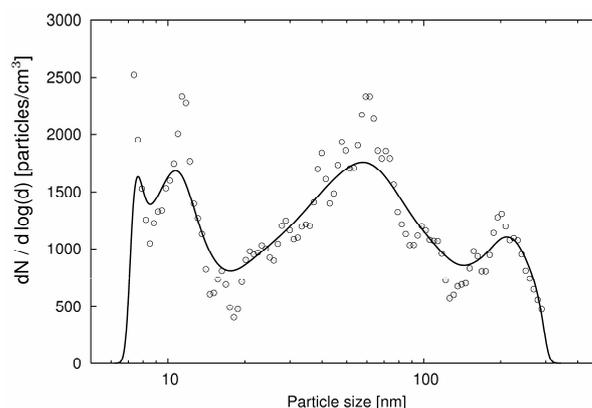


Fig. 7. The distribution density function (curve) determined by the gnostic method. Experimental data are denoted by points. The sample contains a relatively high concentration of small particles.

The character of the distribution is changed in the period from 11 h to 16 h during a particle formation event that has already been discussed in the section Results of the Semiparametric Model. The concentration of small particles is increased, and they often form two separate local maxima. Such a distribution is depicted in Fig. 7.

COMPARISON

Parametric modeling offers an excellent tool if the model is in agreement with the real data. If the model is based upon theoretical assumptions, the parameters estimated from the data provide a useful insight, which can

help in understanding phenomena occurring in the atmosphere. The method, however, fails if the data depart considerably from the model.

The gnostic approach presents a nonparametric method. The value of the scale parameter does not depend on the analyst, it is determined from the condition of entropy equality and thus the risk of under- or over-smoothing is relatively low. The use of the finite data support improves agreement between the empirical and gnostic distribution functions at the edges of the size interval. The number of local maxima is then obtained from an analysis of the distribution function.

The comparison of the original data and its fit by a semiparametric and a nonparametric method are shown in Figs. 1 and 6. The semiparametric method is represented by 3 curves, corresponding to the LN1, LN2 and LN3 models. It can be seen that in this case both LN3 and the nonparametric model describe the measured data reasonably well, recovering both the main peak at 60 nm and also the small one below 10 nm.

Figs. 4 and 8 then show how the two approaches succeeded in capturing the positions of local maxima on the particle distributions' densities as they move in time during the analyzed day. Again, both methods behaved similarly. It seems, however, that the nonparametric approach yielded a more robust result (the positions of local maxima do not change that much with time). The reason is obvious: the nonparametric method performed data filtration in the time domain at first, and it removed a lot of noise from the data. Both approaches revealed the particle formation

event starting around 11 a.m., both followed the growth of these particles until the evening when the concentration maxima connected to these particles disappeared. However, as this new maximum was superimposed on a distribution that already had 3 other maxima, the LN3 model naturally failed to describe the fourth one.

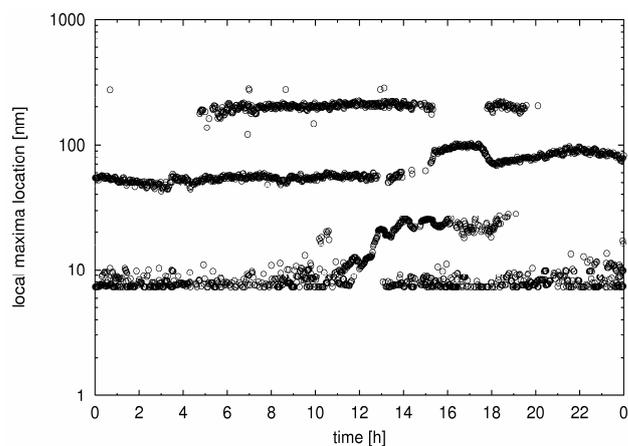


Fig. 8. The locations of local maxima on particle distribution functions and their evolution in time (as determined by the nonparametric model).

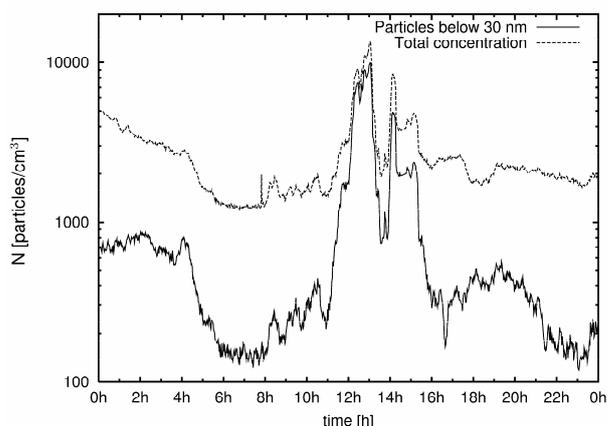


Fig. 9. The concentration of particles with a size below 30nm and the total concentration of particles.

Fig. 9 reveals that within the mentioned time period the aerosol is formed mostly by the newly appearing small particles with sizes below 30 nm. Since this local maximum dominates the distributions, the maxima at other sizes are sometimes changed into shoulders on another peak. This causes short gaps in Fig. 8.

Another important difference is found when calculating the concentration of particles. Let's assume that there are two sources, each of which produces particles with lognormal size distributions. Parametric approach gives us the concentrations for each source of particles. Such information is not available if the nonparametric approach is used. It is only possible to find a local minimum between the two modes and evaluate the concentrations of the particles with sizes lower and greater than the size at the minimum.

A related issue is that of information compression. The parametric model with a few LN components compresses the empirical information into a compact form—parameter estimates. Consider for instance the “compression ratio”—that is the reciprocal of the ratio between the number of datapoints in the empirical spectra (number of channels) and the number of parameters in the LN mixture that is used to fit them. For our data, it is $9/103 = 0.0874$ for LN3 and it is obviously even better for simpler mixtures. The nonparametric methodology can offer compression if we are interested only in the number and positions of modes. If we later need some other type of information, we either have to repeat the calculation, or the calculated

parameters must be stored in addition to the measured data. This is the price paid for higher flexibility in nonparametric estimates.

CONCLUSIONS

This paper presents two methods of modeling atmospheric PSDs: the parametric approach and the nonparametric. Our general view is that in this comparison there is no absolute winner; both approaches have their merits. However, in order to give our kind reader some hints, we will try to summarize pros and cons of both approaches.

The semiparametric method based on lognormal mixtures has these advantages: a) it is more efficient and less computationally demanding; b) it gives excellent results if the model agrees with the data; c) it provides a high data compression ratio; d) it gives concentration of aerosol particles corresponding to each of the modes; and f) it is more easily programmable since its lognormal components are available as built-in functions in most statistical packages.

The drawbacks of the semiparametric method are: a) if the data depart considerably from model assumptions, the method fails to arrive at correct and physically sound results, but this fact may not be recognized from the results itself; b) the method becomes more complicated when the number of LN distributions in the mixture (K) changes with time; and c) it does not follow fast changes in the PSDs well even if number K is kept constant (but that depends more on the

nonparametric smoothing of parameter tracks than on the LN parametric part).

The nonparametric method based on the gnostic theory of uncertain data offers these advantages: a) the method is based on fewer assumptions and does not force the data to have any particular shape; b) it keeps the amount of information contained in the data by fulfilling the condition of preserving the entropy; c) the method is robust to outliers; d) it follows fast changes in the PSD; e) it provides good agreement with the data even at the edges of the PSD.

The main disadvantages of this method are: a) it is not directly available in the statistical packages and so it has to be programmed; b) it is more computationally demanding because the use of a reliable global nonlinear optimization algorithm is inevitable; c) it does not compress the data, even if it provides practically important results as integral concentrations in a given size range or number and position of distributional modes.

ACKNOWLEDGEMENTS

This work was supported by these grants of the EC: ENVK2-1999-00052 and EUSAAR contract RII3-CT-2006-026140, by Czech Science Foundation grant No. 205/03/1560, and partly by the Institutional Research Plan AV0Z10300504 "Computer Science for the Information Society: Models, Algorithms, Applications."

REFERENCES

- Agresti, A. (1990). *Categorical Data Analysis*. John Wiley, New York.
- Box, G.E. and Draper, N.R. (1987). *Empirical Model-Building and Response Surfaces*. John Wiley, New York.
- Carroll, R.J. and Ruppert, D. (1988). *Transformation and Weighting in Regression*. Chapman and Hall, London.
- Cleveland, W.S. and Devlin, S.J. (1988). Locally-Weighted Regression: An Approach to Regression Analysis by Local Fitting. *J. Am. Statist. Assoc.* 83: 596–610.
- Green, P.J. (1995). Reversible Jump Markov chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika*, 82: 711–732.
- Kovanic, P. (1986). A New Theoretical and Algorithmical Basis for Estimation, Identification and Control. *Automatica*. 22: 657–674.
- Krivý, I. and Tvrdík, J. (1995). The Controlled Random Search Algorithm in Optimizing of Regression Models. *Comp. Stat. & Data Analysis*. 20: 229–234.
- Kulmala, M., Vehkamaäki, H., Petaäjä, T., Dal Maso, M., Lauri, A., Kerminen, V.-M., Birmili, W. and McMurry, P. H. (2004). Formation and Growth Rates of Ultrafine Atmospheric Particles: A Review of Observations. *J. Aerosol Sci.* 35: 143–176.
- Lazaridis, M., Eleftheriadis, K., Smolik, J., Colbeck, I., Kallos, G., Drossinos, Y., Ždimal, V., Vecera, Z., Mihalopoulos, N., Mikuska, P., Bryant, C., Housiadas, C., Spyridaki, A., Astitha, M. and Havranek, V.

- (2006). Dynamics of Fine Particles and Photo-Oxidants in the Eastern Mediterranean (SUB-AERO). *Atmos. Environ.* 40: 6214–6228.
- Lee, C.-K., Lin, S.-C. (2008) Chaos in Air Pollutant Concentration (APC) Time Series. *Aerosol Air Qual. Res.* 8: 381-391.
- Makela, J.M., Koponen, I.K., Aalto, P. and Kulmala, M. (2000). One-year Data of Submicron Size Modes of Tropospheric Background Aerosol in Southern Finland. *J. Aerosol Sci.* 31: 595–611.
- Marron, J.S. and Chaudhuri, P. (1998). When is a Feature Really There? The SiZer Approach. In Sadjadi, F. A., Editor, *Autom. Target Recognition VII, Proc. of SPIE*, 3371: 306–312.
- McCulloch, C.E. and Searle, S.R. (2001). *Generalized, Linear, and Mixed Models*. John Wiley, New York.
- Octave. URL <http://www.octave.org>.
- Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *Ann. Math. Statist.* 33: 1065–1076.
- Seinfeld, J.H. and Pandis, S.N., 1998. *Atmos. Chem. and Physics*. J. Wiley & Sons, New York.
- Tvrđík, J. URL <http://albert.osu.cz/tvrdik/>.
- Tvrđík, J. and Krivý, I. (1999). Simple Evolutionary Heuristics for Global Optimization. *Computational Statistics and Data Analysis*, 30: 345–352.
- Tvrđík, J., Krivý, I. and Misík, L. (2001). Evolutionary Algorithm with Competing Heuristics. In *Proceedings of Mendel 2001*. pages 58–64. VUT Brno, 2001. Also available from http://albert.osu.cz/tvrdik/index_pic_en.html.
- Venables, W.N. and Ripley, B.D. (1994). *Modern Applied Statistics with S-Plus*. Springer, New York.
- Voutilainen, A. and Kaipio, J.P. (2002). Estimation of Time-Varying Aerosol Size Distributions—Exploitation of Modal Aerosol Dynamical Models. *J. Aerosol Sci.* 33: 1181–1200.
- Voutilainen, A. and Kaipio, J.P. (2001). Estiation of Non-Stationary Aerosol Size Distributions Using the State-Space Approach. *J. Aerosol Sci.* 32: 631–648.

Received for review, July 21, 2008

Accepted, November 5, 2008