Global NEST Journal, Vol 8, No 3, pp 241-251, 2006
Copyright© 2006 Global NEST
Printed in Greece. All rights reserved

AN APPLICATION OF THEORETICAL PROBABILITY DISTRIBUTIONS, TO THE STUDY OF PM_{10} AND $PM_{2.5}$ TIME SERIES IN ATHENS, GREECE

I. GAVRIIL[1]
G. GRIVAS[1]
P. KASSOMENOS[2]
A. CHALOULAKOU[1,*]
N. SPYRELLIS[1]

[1]*National Technical University of Athens*
School of Chemical Engineering
Heroon Polytechniou 9
GR-15780 Zografos, Athens, Greece
[2]*University of Ioannina, Faculty of Physics*
Department of Astrogeophysics
University Campus, GR-45110, Ioannina, Greece

*to whom all correspondence should be addressed:
e-mail: dchal@central.ntua.gr

Received: 08/02/06
Accepted: 02/05/06

ABSTRACT
Probability density functions (pdf) have been used in the analysis of the distribution of pollutant data, for examining the frequency of high concentration events. There have been very few studies on the concentration distribution of PM in urban areas. The distribution of PM concentrations has an impact on human health effects and the setting of PM regulations. Eight probability distribution functions were fitted to measured concentrations of PM_{10} and $PM_{2.5}$ in order to determine the shape of the concentration distribution. The “goodness-of-fit” of the probability density functions, to the data, was evaluated, using various statistical indices (including Chi-square and Kolmogorov-Smirnov tests). The evaluation was conducted for two separate years and the results indicated that the Pearson type VI pdf provided a better fit to the measured data. Other functions exhibiting high accuracy of fit were the inverse Gaussian, the lognormal and Pearson type V.

The possibility to use probability density functions for predicting the daily high concentration percentiles to less than everyday sampling scenarios is also shown. The differences in the distribution of concentrations under these scenarios are important for regulatory compliance. When trying to detect the high concentrations there is significant possibility of missing the events and thus, underestimating the number of exceedances occurred. Significant deviations from actual daily measurements of PM_{10} and $PM_{2.5}$ concentration percentiles were observed, when infrequent sampling scenarios were examined. The differences were higher for the 1-in-6 sampling schedules and reached 2.8% for mean PM_{10} and 8% for $PM_{2.5}$ while for the maximum concentrations the respective differences were 21.3% and 31.9%. Differences between the frequency distributions of everyday and non-everyday sampled concentrations were observed, while lognormal and inverse Gaussian functions provided a better approximation of the upper percentiles.

Fitting infrequent data on continuous probability functions for the improvement of the approximation to the real statistical values provided good results regarding the 90th percentile, which corresponds to the E.U. provision of 35 annual exceedances of 24-h limit PM_{10} values. In the case of the extreme 98th and 99th percentiles, the method provided satisfactory results for both the PM_{10} infrequent sampling scenarios.

KEYWORDS: Air Pollution, PM_{10}, $PM_{2.5}$, probability density functions.

INTRODUCTION

Given the stochastic nature of atmospheric processes, concentrations of air pollutants can be treated as random variables with measurable statistical properties. If certain conditions are met, the statistical characteristics of pollutant concentrations can be described by theoretical probability density functions. The imposition of these particular mathematical forms can represent the distribution of underlying data. Probability density functions (pdf) have been extensively used in the past years in a variety of applications, where data smoothing, interpolation or extrapolation is needed (Wilks, 1995).

Specifically, in atmospheric sciences the most characteristic applications include the approximation of the frequency of exceedances of critical concentration levels and the estimation of reduction in emissions, required for attainment of AQS (Air Quality Standard) objectives (Georgopoulos and Seinfend, 1982; Abatzoglou *et al.,* 1996; Burkehardt *et al.*, 1998; Morel *et al.*, 1999). In most of the cases, pollutant concentrations have been found to better fit to the lognormal distributions This fact is in agreement with past assumptions that deviation from log-normality should be attributed to sampling error (DeNevers *et al.,* 1979), or with the theory of random dilutions (Ott, 1990). However, it should be stressed out that a strong natural basis is not a prerequisite for the selection of a specific distribution.

Recently, the use of frequency distributions has been extended to ambient particulate matter concentrations (Kao and Friedlander, 1996; Rumburg *et al.*, 2001; Lu, 2002; Gomiscek *et al.,* 2002; Lu and Fang, 2003; Grivas *et al.*, 2004; Lu, 2004; Karaca *et al.*, 2005). Air pollution research has been increasingly focused to particulate matter during the last years, in view of the severe threat it poses on human health, being related to detrimental respiratory and cardiovascular impairments (Dockery and Pope, 1994; Schwartz *et al.*, 1996; Katsouyanni *et al.*, 1997)

Results of systematic and extensive research that commenced during the end of the previous decade, indicated that particulate matter (PM_{10} and $PM_{2.5}$) related atmospheric pollution would emerge as one of the primary environmental issues in the area of Athens (Chaloulakou *et al.*, 2003a; Grivas *et al.*, 2004; Chaloulakou *et al.,* 2005). The research findings were verified when official monitoring of PM_{10} begun in various locations, indicating severe exceedances of EU-established concentration limit values (Grivas and Chaloulakou, 2006).

In the present work, PM_{10} and $PM_{2.5}$ concentration time series from a central measurement site in Athens were statistically examined for the determination of their parent frequency distribution, with the comparative evaluation of several theoretical probability distributions. Probability density functions were also fit to concentration data corresponding to less than daily sampling frequencies, in order to investigate whether this procedure can lead to an improved estimation of high concentration events, as compared to the actual less than daily measurement program. This additional objective of this study becomes of particular importance taking into account that less than daily sampling scenarios are mentioned in both the EU Air Quality Directives, with 14% annual data coverage for indicative measurement (CEC,1999) and in the USEPA PM_{10} and $PM_{2.5}$ air quality standards (1-every-3 days minimum sampling frequency). Moreover, the new proposal to the European parliament for a directive on ambient air quality and cleaner air for Europe (CEC, 2005), which now includes standards for $PM_{2.5}$, clearly mentions the potential of following less than daily sampling routines and measurement campaigns of short duration during a year (supplemented by modeling), in cases where annual mean concentrations remain below an upper assessment threshold (14 μg m^{-3} for PM_{10} and 10 μg m^{-3} for $PM_{2.5}$).

DATA AND METHODS

Particulate matter measurements

PM_{10} and $PM_{2.5}$ sampling was conducted in central Athens (Aristotelous str.). Samplers were placed in a station of the National Air Pollution Monitoring Network, at the building of the Ministry of Public Health. The sampling location is characterized by heavy vehicular traffic, as well as intense commercial and human activity. Twenty-four hour PM_{10} and $PM_{2.5}$ samples (midnight to midnight) were collected daily and simultaneously, using two low-volume reference-equivalent samplers (US EPA-approved Partisol Model 2000, Rupprecht & Patashnick). Sampler inlets were located 6.7 meters above ground. Particles were collected on 47mm Pallflex TX40 filters (Teflon-coated glass fiber filters). Particle concentrations were determined gravimetrically using an electronic microbalance (Mettler Toledo AT201), with a resolution of 0.01 mg. Both blank and field filter samples were conditioned at constant temperature (22±3°C) and relative humidity (40±5%) for at least 24 hours prior to being weighted. The precision of the measurements was determined with parallel sampling and was found equal to 2.5%. The limit of detection (LOD) was estimated at 3 times the standard deviation of field blank filters collected (Burton *et al.,* 1996) and was found 5.4 μg m^{-3} for PM_{10} and 6.3 μg m^{-3} for $PM_{2.5}$.

More details on the measurement protocol as well as extensive analysis of the particle pollution problem of Athens have already been reported in literature (Chaloulakou *et al.*, 2003a; Grivas *et al.*, 2004; Chaloulakou *et al.*, 2005; Manalis *et al.*, 2005).

Everyday measurements were carried out between June1999 - May2001. The measurement schedule resulted in a dataset with a coverage exceeding 90%. For the filling of the missing values a neural network-based imputation technique was implemented. This routine has also been implemented successfully in PM time series management and analysis before (Gavriil *et al.*, 2005). Details on the modeling methodology can be found at Chaloulakou *et al.* (2003c).

Probability density functions

The obtained PM_{10} and $PM_{2.5}$ concentrations (*x)* were fitted to eight selected parent probability distributions, for two years of measurements separately. The examined distributions are described by the following functions:

- *Lognormal*

$$f_L(x) = \frac{1}{\sqrt{2\pi}(x-\gamma)\sigma}\exp\left[-\frac{(\ln(x-\gamma)-\mu)^2}{2\sigma^2}\right],\ x>\gamma;\ -\infty<\mu<\infty;\ \sigma>0;\ \gamma\geq 0 \qquad (1)$$

where μ and σ are the scale and shape parameters of the distribution representing the geometric mean and standard geometric deviation respectively, while γ is the location parameter. Fitting data to a lognormal distribution evaluates the assumption that the logarithmic transformed data values follow a Gaussian distribution.

- *Gamma*

$$f_G(x)=\frac{(x-\gamma)^{\lambda-1}}{\sigma^{\lambda}\Gamma(\lambda)}\ \exp[-(\frac{x-\gamma}{\sigma})],\ x\geq\gamma;\ \sigma>0;\ \lambda>0;\ \gamma\geq 0 \qquad (2)$$

where σ and λ are the scale and shape parameters of the distribution, γ is the location parameter and Γ is the *gamma* function:

$$\Gamma(\lambda)=\int_0^{\infty} t^{\lambda-1}e^{-t}dt \ \text{(Eulerian integral form)} \qquad (3)$$

- *Weibull*

$$f_W(x) = \frac{\lambda}{\sigma}\left(\frac{x-\gamma}{\sigma}\right)^{\lambda-1}\ \exp[-(\frac{x-\gamma}{\sigma})^{\lambda}],\ x\geq\gamma;\ \sigma>0;\ \lambda>0;\ \gamma\geq 0 \qquad (4)$$

where σ and λ are the scale and shape parameters of the distribution, and γ is the location parameter. If *λ*=1 the Weibull distribution is identical with the Gamma distribution.

- *Beta*

$$f_B(x)=\frac{(\beta-\alpha)^{1-\sigma-\lambda}(x-\alpha)^{\sigma-1}(\beta-x)^{\lambda-1}}{B(\sigma,\lambda)},\quad \alpha< x <\beta;\ \sigma>0;\ \lambda>0;\ \beta> \alpha >0 \qquad (5)$$

where σ and λ are the scale and shape parameters of the distribution, [α,β] is the concentration range and B is the *beta* function (Legendre's solution to the Eulerian integral of the first kind).:

$$B(\sigma,\lambda)=\frac{\Gamma(\sigma)\Gamma(\lambda)}{\Gamma(\sigma+\lambda)}=\frac{(\sigma-1)!(\lambda-1)!}{(\sigma+\lambda-1)!} \qquad (6)$$

- *Inverse Gaussian*

$$f_{IG}(x)= \left(\frac{1}{2\pi\, x^{3}\sigma}\right)^{1/2} \exp[-\frac{1}{2x}\,(\frac{x-\mu}{\mu\sigma})^{2}],\ 0<x<\infty;\ \mu>0;\ \sigma>0 \qquad (7)$$

where μ is the data mean and σ is a scale parameter. The inverse Gaussian is also known as the Wald distribution.

- *Log-logistic*

$$f_{LL}(x)= \frac{1}{\sigma}\,\frac{\exp\frac{\ln(x-\gamma)}{\sigma}}{[1+\exp\frac{\ln(x-\gamma)}{\sigma}]^{2}}],\ x\geq\gamma;\ \sigma>0;\ \gamma\geq 0 \qquad (8)$$

where σ is a scaling parameter and *γ* is the location parameter. The log-logistic distribution is derived from the logistic distribution, through logarithmic tranformation of the data.

- *Pearson type V*

$$f_{PV}(x)=\frac{\lambda^{\sigma}}{\Gamma(\sigma)(\chi-\gamma)^{\sigma+1}}\ \exp(-\frac{\lambda}{x-\gamma}),\ x\geq\gamma;\ \sigma>0;\ \lambda>0;\ \gamma\geq 0 \qquad (9)$$

where *Γ* is the Gamma function, σ and λ are the scale and shape parameters, and *γ* is the location parameter.

- *Pearson type VI*

$$f_{PVI}(x)=\frac{a^{s}\,(x-\gamma)^{\beta-1}}{B(\beta,\sigma)(\alpha+(\chi-\gamma))^{\sigma+\beta}},\ x\geq\gamma;\ a>0;\ b>0;\ m>0;\ \gamma\geq 0 \qquad (10)$$

where B is the *beta* function, α, β are the shape parameters, σ the scale parameter, and *γ* is the location parameter. The last two distributions are part of the wider Pearson frequency distribution system which is comprised by 12 theoretical distributions. Pearson distributions are classified by the parameter *κ*:

$$\kappa=\frac{\beta_1(\beta_2+3)^{2}}{4(2\beta_2-3\beta_1-6)(4\beta_2-3\beta_1)} \qquad (11)$$

where β_1 β_2 are the squared skewness and squared kurtosis, respectively. For *κ*=1 the distribution is called Pearson type V and for *κ*>1 Pearson type VI (in fact the Gamma distribution is the Pearson type III distribution where $\kappa\rightarrow\infty$)

Parameter Estimation

The specific nature of theoretical distribution is determined by the particular values of their parameters. The optimal values of the scale and shape parameters of the distributions were estimated using the method of maximum likelihood.

This method aims to calculate θ_k parameters of a *k*-parameter distribution in order to maximize the likelihood function *L(θ)*.

$$L(\theta)=L(x_1,x_2,\ldots,x_n;\theta_1,\theta_2,\ldots,\theta_\kappa)=f(x_1;\theta_1,\theta_2,\ldots,\theta_\kappa)f(x_2;\theta_1,\theta_2,\ldots,\theta_\kappa)\ldots f(x_n;\theta_1,\theta_2,\ldots,\theta_\kappa)=$$

$$\prod_{i=1}^{n} f(x_i;\theta_1,\theta_2,\ldots,\theta_\kappa) \qquad (12)$$

where $(x_1,x_2,\ldots,x_n)$ are the independent observations from a random sample deriving from a population following a distribution described by a *κ*-parameter probability density function $f(x;\theta_1,\theta_2,\ldots,\theta_\kappa)$.

If $f(x_1;\ \theta_1,\theta_2,\ldots,\theta_\kappa)$,$f(x_2;\ \theta_1,\theta_2,\ldots,\theta_\kappa),\ldots,\ f(x_n;\ \theta_1,\theta_2,\ldots,\theta_\kappa)$ are the probability functions of each of each of the sample values then the maximum likelihood function describes the joint probability function of the random sample.

The likelihood function being differentiable at $\theta_1, \theta_2, \ldots, \theta_\kappa$ the estimation of the parameters with the maximum likelihood method is made by taking the partial derivatives of *L(θ)* by each parameter and solving the resulting *κ* equations to zero. Usually, computations are made using the logarithm of the likelihood function, and since the logarithm is also a strictly increasing function the same parameters values will maximize both the likelihood and the log-likelihood functions.

$$\frac{\partial \ln L(\theta)}{\partial \theta_\kappa} = 0 \tag{14}$$

The method of maximum likelihood is considered advantageous for parameter estimation in comparison with the simple methods of moments (which is also occasionally used), as the latter can lead to misleading extrapolations and interferences. On the other hand, since the method of maximum likelihood requires ample processing power due to the complex numerical calculations involved, when large data sets are analyzed, computational time increases substantially.

The location parameter *γ* was set to zero for the continuous distribution functions since it was desired that concentrations would exhibit behaviour with physical meaning (it should be noted that accepting or stating a certain value as a global or regional particle concentration background is rather difficult and will not be attempted). The upper bound *b* for the beta distribution was set to 250 μg m^{-3} for PM_{10} and to 150 μg m^{-3} for $PM_{2.5}$ since these concentration levels have never been reached in the Greater Area of Athens, on a daily basis, even in the occurrence of severe episodic conditions. The optimum scale parameter for the Pearson type VI distribution was determined using an iterative trial and error process.

Goodness-of-fit tests

The comparative evaluation of the above presented functions was made using the Kolmogorov-Smirnov (K-S), Anderson-Darling (A-D) and Chi-Square (χ^2) goodness-of-fit tests.

- The K-S statistic is defined as the maximum difference between the sample cumulative distribution function -S(x) and the examined theoretical function -F(x).

$$D_0 = \max\left|F(x) - S(x)\right| \tag{15}$$

The D_0 value is compared with the $D_{n,a}$ value, which is the largest difference acceptable at the significance level a for a n-sized sample. If $D_0 < D_{n,a}$ the hypothesis that the sample can be described by the fitted theoretical distribution is accepted at the a significance level.

- The A-D statistic controls the hypothesis that the sample derives from a distribution which is described by the fitted density function, using the A^2statistic:

$A^2 = -N - S$, where:

$$S = \sum_{i=1}^{N} \frac{2i-1}{N}[\ln F(X_i) + \ln(1 - F(X_{N+1-I})] \tag{16}$$

and $X_1 \ldots X_N$ are the sample values sorted in order of magnitude. The A-D statistic is considered more dependable than the K-S statistic since it emphasizes at the upper tails of the distribution functions where the larger discrepancies are expected.

- The χ^2 test divides the data range in a predefined number of k independent bins and then compares the number (n) of actual observations in each bin (n_i) with the number theoretically assigned to this bin by the fitted function. The χ^2 test essentially compares the data histogram with the probability density function. The comparison is made with the statistic:

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} \tag{17}$$

where p_i is the probability of occurrence in the ith bin for the theoretical distribution. Under the null hypothesis that the data were drawn from the examined distribution, the test statistic follows the chi-square distribution with v degrees of freedom (v=number of bins - number of parameters- 1). In fact, the chi-square distribution is a special case of the aforementioned

Gamma distribution with λ=v/2 and σ=2. Small values of the statistic support the null hypothesis. It is noted that the determination of number of bins for the χ^2 tests should be done in a way that at least 5 expected counts are included in each concentration interval (Wilks, 1995).

Statistic indicators

A series of error indices and correlation measures were also utilized. If P_i is the expected value from the probability density function and O_i is the measured value then the following indices are defined (Wilmott, 1982; Chaloulakou *et al.*, 2003b):

Mean Absolute Error (MAE): $\mathrm{MAE}= \frac{1}{N}\sum_{i=1}^{N} |Oi - Pi|$ (18)

Root Mean Square Error (RMSE): $\mathrm{RMSE}= (\frac{1}{N}\sum_{i=1}^{N} |Oi - Pi|^2)^{1/2}$ (19)

Correlation Coefficient (R): $R = \frac{\sum_{i=1}^{N}(Pi - \overline{Oi})^2}{\sum_{i=1}^{N}(Oi - \overline{Oi})^2}$ (20)

Index of Agreement (d_2): $d_{2}= 1 - \frac{\sum_{i=1}^{N}(Pi - Oi)^2}{\sum_{i=1}^{N}\left(\left|Pi - \overline{Oi}\right| + \left|Oi - \overline{Oi}\right|\right)}$ (21)

Software

The bulk of the probability distribution statistical treatment was conducted using the specialized software suite, ExpertFit 6.0 (Averill M. Law & Associates). Supplementary statistical runs were made with SPSS 11.0 and numerical calculations with Wolfram Research Mathematica 5.0.

RESULTS AND DISCUSSION

Fitting distribution functions to the entire dataset

The aforementioned conventional statistical indicators (MAE, RMSE, R^2, d_2) were used at an initial stage for the evaluation and selection of the above eight parent distributions, which are presented in this study, from the larger ensemble of continuous univariate theoretical distributions. When tested with the traditional statistical error and correlation measures they all exhibited an above average approximation capability and they were subsequently evaluated using the goodness of fit statistics. The best-fit probability density functions are selected based on the combined results of goodness of fit statistics.

The results of the evaluation of probability density functions and their rankings are displayed on Table 1 for PM_{10} and on Table 2 for $PM_{2.5}$. The best overall performing pdf for both PM_{10} and $PM_{2.5}$ was the tri-parametric Pearson type VI. Good results were also obtained for Pearson type V, inverse Gaussian and lognormal frequency distributions. The performance of the log-logistic pdf was marginally satisfactory while beta, gamma and Weibull distributions did not produce quite an accurate fit to the data. The results are supported by the findings, recently presented by Karaca *et al.* (2005) for PM_{10} and $PM_{2.5}$ data in Turkey, which evaluate a wide set of candidate theoretical probability distributions. Earlier work has also shown the precedence of lognormal frequency distributions over gamma and Weibull distributions (Rumburg *et al.,* 2001, Grivas *et al.*, 2004). A general conclusion is that the distribution of measured PM data, which present higher frequencies of low to mid range concentrations but also considerable high concentration events, were best approximated by strongly right-skewed continuous probability density functions. This is reasonable since particulate matter concentrations as atmospheric variables present a high degree of asymmetry, being physically limited on the left, since they are constrained to be nonnegativite, but they also tend to include distinct high values (episodic events).

Table 1. Goodness-of it statistics of examined probability distributions to daily PM_{10} concentrations, for the two years of measurement. Rankings listed in parentheses

	1st year			*2nd year*		
	K-S	A-D	χ^2	K-S	A-D	χ^2
Beta	0.091(7)	3.256(7)	13.68(7)	0.101(8)	4.712(7)	27.20(7)
Gamma	0.069(6)	1.594(6)	7.89(6)	0.067(6)	1.998(6)	12.96(6)
Inverse Gaussian	0.045(4)	0.512(3)	4.80(3)	**0.039(1)**	0.520(3)	6.98(5)
Log-logistic	0.044(3)	0.890(5)	7.25(4)	0.049(5)	0.803(5)	5.43(3)
Lognormal	0.046(5)	0.573(4)	7.74(5)	0.040(3)	0.586(4)	6.86(4)
Pearson V	0.034(2)	0.316(2)	**3.91(1)**	0.042(4)	0.278(2)	4.26(2)
Pearson VI	**0.032(1)**	**0.315(1)**	4.60(2)	0.040(2)	**0.267(1)**	**4.06(1)**
Weibull	0.102(8)	5.057(8)	39.94(8)	0.096(7)	5.213(8)	29.65(8)

Table 2. Goodness-of it statistics of examined probability distributions to daily $PM_{2.5}$ concentrations, for the two years of measurement. Rankings listed in parentheses

	1st year			*2nd year*		
	K-S	A-D	χ^2	K-S	A-D	χ^2
Beta	0.084(7)	3.209(7)	26.27(7)	0.092(8)	4.161(7)	25.46(7)
Gamma	0.058(6)	1.516(6)	12.67(6)	0.078(6)	2.004(6)	12.39(6)
Inverse Gaussian	0.045(3)	**0.551(1)**	**5.26(1)**	0.048(3)	0.527(3)	**5.17(1)**
Log-logistic	0.053(5)	1.021(5)	9.71(5)	0.049(4)	1.111(5)	10.57(5)
Lognormal	0.047(4)	0.619(3)	5.53(2)	0.051(5)	0.685(4)	6.04(2)
Pearson V	**0.043(1)**	0.643(4)	6.91(4)	0.047(2)	0.477(2)	8.40(4)
Pearson VI	**0.043(1)**	0.580(2)	5.77(3)	**0.044(1)**	**0.455(1)**	7.31(3)
Weibull	0.090(8)	4.888(8)	42.01(8)	0.084(7)	4.403(8)	30.18(8)

Figures 1a, 1b show the Pearson type VI distribution overplot, superimposed on the histogram of measured PM_{10} and $PM_{2.5}$ data, respectively. The parameters of the distribution are displayed along with measures of statistical agreement between measured and expected concentrations. The low values of the error indices (MAE, RMSE) and the close to unity values of R^2 and d_2 verify the suitability of the function for describing the distribution of measured data.

Statistics for infrequent sampling scenarios

Table 3 presents basic statistics for PM_{10} and $PM_{2.5}$ concentrations measured during 2000 as well as the range of statistics of the time series corresponding to hypothetic non-daily sampling frequencies. A discussion of the severity of PM_{10} and $PM_{2.5}$ levels observed in central Athens can be found elsewhere (Chaloulakou *et al.*, 2003a; Grivas *et al.,* 2004; Chaloulakou *et al.*, 2005; Grivas and Chaloulakou, 2006). However, it is worth mentioning the severe deviation of the $PM_{2.5}$ annual mean concentration from the oncoming EU concentration cap of 25 μg m^{-3} (CEC, 2005). This value (based on rolling three year averages) has to be achieved by 2010. In view of the requirements of this new directive, the Greek authorities have established a 4-station $PM_{2.5}$ monitoring network.

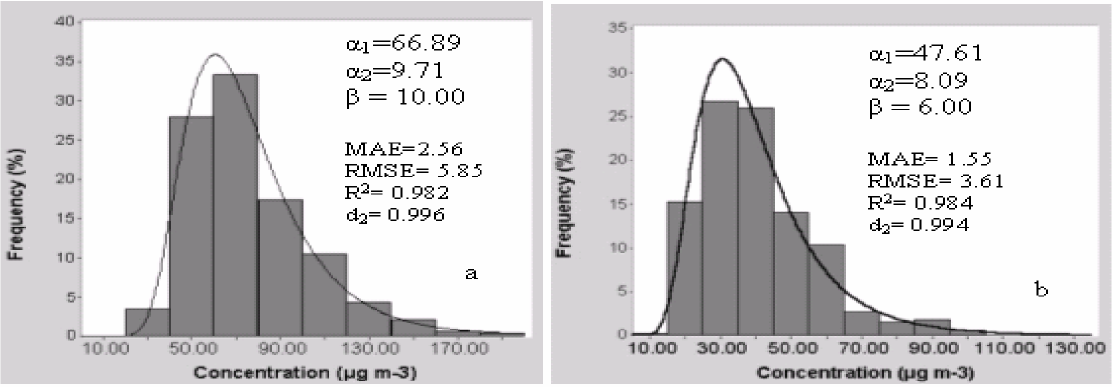

Figure 1. Pearson type VI overplot on density histograms of PM_{10} (a) and $PM_{2.5}$ (b) concentrations

Table 3. Descriptive statistics for everyday and infrequent sampling during the calendar year of 2000

	PM_{10}			$PM_{2.5}$		
	Daily	1 in 3 days (range)	1 in 6 days (range)	Daily	1 in 3 days (range)	1 in 6 days (range)
Mean	76.7	75.7-77.5	74.5-78.9	40.2	39.9-40.1	37 -41.4
Max.	207	185-207	163-207	135	108-135	92-135
90th	125	115-132	110-135	65.3	63-71	57-77
98th	171	157-170	141-177	91.2	91-92	82-92
99th	183	174-183	152-190	95.0	92-97	86-113
s	34.3	33.3-36.2	27.6-39.5	20.2	18.6-21.5	17.1-22.8
Skewness	1.15	1.07-1.27	0.90-1.52	1.38	1.24-1.62	0.78-1.98
Kurtosis	1.14	0.60-1.67	0.11-3.46	2.44	1.70-3.96	0.61-5.82

As expected, the discrepancies in the results increase as sampling frequency decreases. The differences are less pronounced for the average concentrations (reaching 2.8% for PM_{10} and 8% for $PM_{2.5}$ at the 1-in-6 sampling schedules) than for the higher percentiles (discrepancies for the maxima as high as 21.3% and 31.9% for PM_{10} and $PM_{2.5}$, respectively). Large differences are also characteristic for scewness and kurtosis coefficients and indicate that the underlying distributions of data sets obtained from infrequent sampling present differences, as compared to those of daily sampling. It is characteristic, that the error for the moments about the arithmetic mean increases with the rank of the moment (standard deviation is the second, skewness is the third and kurtosis the fourth moment).

Evaluation of extreme percentiles with fitted probability density functions

The evaluation procedure was repeated for the less-than-daily sampling scenarios. These involve 1-every-6 days (old USEPA) and 1-every-3 days (current USEPA) sampling frequencies. The sampling frequency of 1-every-6 days is close to the required frequency by the E.U. for indicative sampling of PM_{10}. Higher percentiles of concentrations (relative to existing limit values) are calculated using the theoretical distributions and compared to those actually measured.

The datasets coming from infrequent sampling were fit to the four best performing functions described above (Pearson type VI, Pearson type V, inverse Gaussian, lognormal). It was examined if the high percentiles calculated by the distributions provide a better approximation to those of everyday sampling in comparison to the high percentiles deriving from actual infrequent sampling. The root mean square error was used as a measure of agreement, since more than one datasets correspond to infrequent sampling scenarios. The results are presented on Table 4.

Table 4. Root mean square error between concentration statistics based on everyday sampling and statistics based on infrequent sampling (measured and fitted). Values in μg m^{-3}

	Measured	Pearson type V	Pearson type VI	Inverse Gaussian	Lognormal
PM_{10}, 1-in-3					
Mean	0.7	0.8	0.8	0.7	0.8
90th	7.1	3.7	4.1	4.7	5.4
98th	8.1	14.9	12.6	6.7	6.7
99th	6.7	33.6	29.7	6.2	6.5
PM_{10}, 1-in-6					
Mean	1.7	2.0	1.9	1.7	1.8
90th	9.6	8.5	8.8	8.6	9.0
98th	19.2	24.8	23.1	15.9	16.3
99th	20.1	43.5	39.3	18.5	19.4
$PM_{2.5}$, 1-in-3					
Mean	0.2	0.6	0.4	0.2	0.3
90th	3.8	3.4	2.9	2.2	2.0
98th	0.2	10.3	19.6	15.6	6.2
99th	2.6	37.7	33.1	12.4	13.1
$PM_{2.5}$, 1-in-6					
Mean	1.5	1.5	1.5	1.5	1.5
90th	6.2	4.3	4.3	3.8	3.8
98th	5.5	21.1	19.4	8.4	8.6
99th	10.3	39.1	35.6	13.8	14.8

In the case of infrequent sampling the lognormal and inverse Gaussian and lognormal pdf provided a better fit especially for the higher concentration range. The suitability of the lognormal probability density function for data obtained by an 1-every-6 days sampling procedure has been already documented in literature (Kao and Friedlander, 1995; Rumburg *et al.,* 2001; Grivas *et al.*, 2004). It appears that the most sophisticated Pearson type V and VI function can capitalize on more lengthy datasets (e.g. the full datasets previously examined) rather than on datasets of reduced size.

The results indicate that for the 90th percentile (which approximately corresponds to the E.U. provision of 35 annual exceedances of 24-h limit PM_{10} values), the statistical treatment of measured data can provide a better approximation to the concentration which derives from everyday sampling and is required for investigation of compliance with legislated air quality standards. Thus, it is proposed that in the case of infrequent sampling, a post-hoc statistical treatment can lead to an improved estimate of the threshold concentration over which the critical number of 35 annual exceedances appears.

In the case of the 98th (already used by the USEPA for $PM_{2.5}$ and is relevant to the E.U. provision of 7 annual exceedances of 24-h limit PM values, which has been proposed for the second stage of the implementation of the air quality daughter directive) and 99th (used by the USEPA for PM_{10}) percentiles, the inverse Gaussian and lognormal function appeared to improve the approximation for both of the PM_{10} infrequent sampling scenarios. The improvement (expressed in RMSE) for the 98th and 99th percentiles respectively, using the inverse Gaussian function was 17.2% and 7.5% for 1-every-3 days sampling, while in the case of 1-every-6 days sampling the improvement was 17.4% and 7.9% for the two percentiles.

Regarding $PM_{2.5}$ the 98th and 99th percentiles calculated from fitted data appear to deviate from the actual values. Rumburg *et al.* (2001) also report similar findings for $PM_{2.5}$ concentrations. The inability of the theoretical distributions to accurately reproduce the variability of the actual dataset in the extreme value area range of $PM_{2.5}$ concentrations should be attributed to the abnormality in the right tail of their actual distribution. This is also apparent from Figure 1b. Indeed, an increased (by 24%) frequency of concentrations around 90μg m^{-3} (where the extreme percentiles are calculated) is observed, in comparison with the preceding concentration range. The procedure should be reevaluated in the future when a lengthier $PM_{2.5}$ time-series will become available.

5. CONCLUSIONS

Selected probability distribution functions were fitted to PM_{10} and $PM_{2.5}$ concentration data measured for two years at a central location in Athens. As evaluated with goodness- of-fit measures, it appeared that the most appropriate probability density functions were the Pearson type VI and V, the inverse Gaussian and the lognormal functions. We conclude on the suitability of continuous, positively skewed distributions for describing PM data in areas with increased concentration levels.

Significant deviations from actual daily measurements for the higher PM_{10} and $PM_{2.5}$ concentration percentiles were obtained, when infrequent sampling scenarios were examined. The use of theoretical probability distributions on infrequent data for the improvement of the approximation to the real statistical values yielded good results regarding the crucial 90th concentration percentile for both PM_{10} and $PM_{2.5}$ and also improved the approximation to 98th and 99th percentiles for PM_{10}. For $PM_{2.5}$ the application did not appear efficient for the extreme 98th and 99th percentiles. It is proposed that more specific analysis should be conducted for the right tails of particulate matter frequency distributions, including the consideration of extreme value probability density functions.

The presented specific application of probability density function in the field of particulate matter study is one of the numerous possible applications and highlights the importance of this promising method. It is suggested that further research is conducting by exploring the statistical characters of particulate matter for pursuing critical tasks, as the prediction of exceedances of limit values and assessment thresholds, the estimation of emission source reduction and evaluation of proposed air quality control policies.

REFERENCES

Abatzoglou G., Chaloulakou A., Assimacopoulos D., Lekkas T. (1996) Prediction of air pollution episodes: extreme value theory applied in Athens, *Environmental Technology*, **17**, 349-359.

Burkhardt J., Sutton M.A., Milford C., Storeton-West R.L., Fowler D. (1998) Ammonia concentrations at a site in Southern Scotland from 2 yr. of continuous measurements, *Atmospheric Environment*, **32**, 325-331.

CEC- Commission of the European Communities (1999) Council Directive 1999/30/EC, relating to limit values for sulphur dioxide, nitrogen dioxide and oxides of nitrogen, particulate matter and lead in ambient air, *Official Journal of European Communities*, **L163,** 41-60.

CEC- Commission of the European Communities (2005) Proposal for a Directive of the European Parliament and of the Council on Ambient on ambient air quality and cleaner air for Europe, Brussels, 21.9.2005.

Chaloulakou A., Kassomenos P., Spyrellis N., Demokritou P., Koutrakis P. (2003), Measurements of PM_{10} and $PM_{2.5}$ particle concentrations in Athens, Greece, *Atmospheric Environment*, **37**, 649-660.

Chaloulakou A., Saisana M., Spyrellis N. (2003) Comparative assessment of neural networks and regression models for forecasting summertime ozone in Athens, *The Science of the Total Environment*, **313**, 1-13.

Chaloulakou A., Grivas G., Spyrellis N. (2003) Neural Network and Multiple Regression Models for PM_{10} Prediction in Athens. A Comparative Assessment, *Journal of the Air & Waste Management Association***, 53**, 1183-1190.

Chaloulakou A., Kassomenos P., Grivas G., Spyrellis N. (2005) Particulate Matter and Black Smoke concentration levels in Central Athens, Greece, *Environment International*, **31**, 651-659.

De-Nevers N., Lee K.W., Frank N.H. (1979) Patterns in TSP distribution', *Journal of the Air Pollution Control Assossiation*, **29**, 32-35.

Dockery D.W., Pope III C.A. (1994) Acute respiratory effects of particulate air pollution, *Annual Review of Public Health,* **15**, 107-132.

Gavriil I., Grivas G., Diapouli E., Kanouta V., Chaloulakou A., Spyrellis N. (2005) $PM_{2.5}$ concentration time-series in Athens, Greece, European Aerosol Conference, September 2005, Ghent, Belgium, 387.

Georgopoulos P.G., Seinfeld J.H. (1982) Statistical distributions of air pollutant concentrations, *Environmental Science and Technology*, **16**, 401A-416A.

Gomiscek B., Hauck H., Stopper S., Preining O. (2004) Spatial and temporal variations of PM_1, $PM_{2.5}$, PM_{10} and particle number concentration during the AUPHEP- project*, Atmospheric Environment*, **38**, 3917-3934.

Grivas G., Chaloulakou A., Samara C., Spyrellis N. (2004) Spatial and Temporal Variation of PM_{10} mass concentrations within the Greater Area of Athens, Greece, *Water Air and Soil Pollution*, **158**, 357-371.

Grivas G., Chaloulakou A. (2006) Artificial Neural Network Models for Prediction of PM_{10} Hourly Concentrations, in the Greater Area of Athens, Greece, *Atmospheric Environment*, **40**, 1216-1229.

Kao A.S., Friedlander S.K. (1995) Frequency distributions of PM10 chemical components and their sources, *Environmental Science and Technology*, **29**, 19-28.

Karaca F., Alagha O., Erturk F. (2005) Statistical characterization of atmospheric PM_{10} and $PM_{2.5}$ concentrations at a non-impacted suburban site of Istanbul, Turkey, *Chemosphere*, **59**, 1183-1190.

Katsouyanni K., Touloumi G., Spix C., Schwartz J., Balducci F., Medina S., Rossi G., Wojtyniak B., Sunyer J., Bacharova L., Schouten J.P., Ponka A., Anderson H.R. (1997) Short term effects of ambient sulphur dioxide and particulate matter on mortality in 12 European cities: results from time series data from the APHEA project, *British Medical Journal*, **314**, 1658-1663.

Lu H.C. (2002) The statistical characters of PM10 concentration in Taiwan area, *Atmospheric Environment*, **36**, 491-502.

Lu H.C., Fang G.C. (2002) Estimating the frequency distributions of PM_{10} and $PM_{2.5}$ by the statistics of wind speed at Sha-Lu, Taiwan, *Science of the Total Environment*, **298**, 119-130.

Lu H.C., Fang G.C. (2003) Predicting the exceedances of a critical PM_{10} concentration- a case study in Taiwan, *Atmospheric Environment*, **37**, 3491-3499.

Lu H.C. (2004) Estimating the emission source reduction of PM10 in central Taiwan, *Chemosphere*, **54**, 805-814.

Manalis N., Grivas G., Protonotarios V., Moutsatsou A., Samara C., Chaloulakou A. (2005) Toxic metal content of particulate matter (PM_{10}) within the Greater Area of Athens, *Chemosphere*, **60**, 557-566.

Morel B., Yen S., Cifuentes L. (1999) Statistical distribution for air pollutant applies to the study of the particulate problem in Santiago, *Atmospheric Environment*, **33**, 2575-2585.

Nagahara Y. (2004) A method of simulating multivariate nonnormal distributions by the Pearson distribution system and estimation, *Computational Statistics and Data Analysis*, **47**, 1-29.

Ott W.R.A. (1990) Physical explanation of the lognormality of pollutant concentrations, *Journal of the Air & Waste Management Association*, **40**, 1378-1383.

Rumburg B., Alldredge R., Claiborn C. (2001) Statistical distributions of particulate matter and the error associated with sampling frequency, *Atmospheric Environment*, **35**, 2907-2920.

Schwartz J., Dockery D.W., Neas L.M. (1996) Is daily mortality associated specifically with fine particles?, *Journal of the Air & Waste Management Association*, **46**, 927-939.

Wilks D.S. (1995) *Statistical methods in the atmospheric sciences*, Academic Press, San Diego, CA.

Willmott C.J. (1982) Some comments on the evaluation of model performance, *Bulletin of the American Meteorology Society*, **63**, 1309-1313.