CHEMICAL ENGINEERING TRANSACTIONS
VOL. 57, 2017
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.aidic.it/cet
Guest Editors: Sauro Pierucci, Jiří Jaromír Klemeš, Laura Piazza, Serafim Bakalis

ISBN 978-88-95608- 48-8; **ISSN** 2283-9216
DOI: 10.3303/CET1757082

Application of Probability Density Functions in Modelling Annual Data of Atmospheric NOx Temporal Concentration

Wesley H. Prieto, Marco A. Cremasco

Department of Process Engineering – School of Chemical Engineering – University of Campinas
Albert Einstein Avenue, 500 – ZIP CODE: 13083-852 – Campinas - SP – Brazil
wesley@feq.unicamp.br

Currently it is observed, in many countries, an increasing concern by environmental agencies to monitor and control the air pollutants levels and, in this scenario, nitrogen oxides mainly arising from combustion processes deserve special attention. In large cities, the concentration and dispersion of NOx should be monitored not only by its toxicity, but also to be associated with photochemical production of tropospheric ozone, fine particulates, and its participates in the production of free radicals in atmosphere. In this context, the importance of understanding this phenomenon is grounded not only for to understand the complex dynamics involved in air pollution, but also the indispensability of the study of modeling and forecasting methodologies that can provide information for decision making with regard to the control of this compound in atmosphere. Thus, the present study aims to model, by probability density functions (PDF), the annual concentrations of NOx obtained in the period of 2010 to 2015 at the monitoring station of Ibirapuera Park, Sao Paulo, belonging to the Environmental Sanitation Technology Company of the State of Sao Paulo, Brazil. Initially, temporal data were exported directly from the electronic platform of Sao Paulo's agency of air pollution control. The variation of annual NOx concentration is expressed in time series, with 1 hour of acquisition frequency and a total of 8,600 points/year. After obtaining the time series, the original data were organized into classes, and the maximum and minimum intervals determined by Sturges rule. In order to choose the most representative statically bin, it was evaluated the coefficient of variation of the mean to determine the point from which there are no more significant variations of the mean values of concentration of each time series. After this step, fourteen probability density functions were evaluated, and the fitting of the models were assessed by the Kolmogorov-Smirnov test. From the analyzes, it was concluded that the evaluated data showed clear positive displacement and leptokurtic distribution, indicating the Gumbel probability density function as the most representative among those evaluated in this study.

1. Introduction

One of the biggest problems associated with a country's economic development is air pollution. The relationship between the increasing concentration of air pollutants and the incidence of various health problems of the population is becoming increasingly clear, making air pollution a serious public health problem (Braga, 1999). In both developed and developing countries, the motor vehicles and the increasing emissions of toxic pollutants from industrial chimneys (Derisio, 1992), cause high concentrations of harmful substances that are responsible for low visibility and various respiratory problems in living beings (Pope et al., 2002). After the advent of the Industrial Revolution little was done to control or study these effects, the episodes of thermal inversion related to meteorological events being those responsible for impelling the scientific community to verify, certify and relate mortality rates in urban centers to atmospheric pollution (Dockery and Pope, 1994). In this aspect, the oxides of nitrogen (NOx) stand out. Automotive vehicles account for 96.3% of all NOx emissions in the São Paulo Metropolitan Region (CETESB, 2004) and, therefore, produce more nitrogen oxides than any other human activity. It is clear that it is necessary to control the emission of pollutants and in this sense several methodologies are applied in order to quantify, control the concentrations of these compounds and generate indicators of air quality. In this scenario, the probabilistic methodologies are highlighted. Probability density functions have been applied successfully in many physical phenomena such

Please cite this article as: Prieto W., Cremasco M., 2017, Application of probability density functions in modeling annual data of atmospheric nox temporal concentration, Chemical Engineering Transactions, 57, 487-492 DOI: 10.3303/CET1757082

as river discharges, wind speed, rainfall, and air quality (Harikrishna and Arun, 2003; Kan and Chen, 2004; Oguntude et al., 2014). Studying the dispersion of pollutants through a probabilistic approach is important because when the parent probability distribution of air pollutants is correctly chosen, the specific distribution can be used to predict the mean concentration and probability of exceeding a critical concentration (Oguntunde et al., 2014). Therefore, the present work aims to study the atmospheric dispersion of NOx in the years 2010 to 2015 by means of time series obtained at the Monitoring Station of the Environmental Sanitation Company of the State of São Paulo, located in the Ibirapuera Park, in the city of São Paulo. The approach used was purely probabilistic, focusing on the interpolation of the best probability density function for the data modeling.

2. Materials and methods

The time series of NOx concentrations were directly exported from the electronic platform of the Environmental Sanitation Technology Company of the State of São Paulo (CETESB). All analyzes were performed for the years 2010 to 2015. The sampling frequency is 1 hour, with 8,760 total points (N) in each series. Therefore, the present work is divided in two parts. In the first one, the data were divided into bins (K) using as limits the Sturges Rules presented in Equations 1 and 2. To validate the choice of the optimal bin, the coefficient of variation of the mean (CV = σ/μ) was calculated for K = 3, 5, 8, 10, 12, 15, 20, 25, 30, 35, 40, 50, 60 e 70.

Sturges Rules (N < 25): $K = 1 + 3{,}3 \log N$ (1)

Sturges Rules (N > 25): $K = \sqrt{N}$ (2)

For all time series, mean (μ), standard deviation (σ), variance (σ^2), skewness (A) and kurtosis (C) were obtained according to Equations 3, 4, 5, 6 and 7.

Mean: $\mu = \dfrac{\sum_{j=1}^{k} f_j x_j}{N}$ (3)

Standard Deviation: $\sigma = \sqrt{\dfrac{\sum_{j=1}^{N} \left(x_j - \mu\right)^2}{N}}$ (4)

Variance: $\sigma^2 = \sqrt{\dfrac{\sum_{j=1}^{N} \left(x_j - \mu\right)^2}{N}}$ (5)

Skewness: $A = \dfrac{E\left(X - \mu\right)^3}{\sigma^3}$ (6)

Kurtosis: $C = \dfrac{E\left(X - \mu\right)^4}{\sigma^4}$ (7)

The second stage of this study consisted in adjusting the probability density functions (PDF) present in Equations 8 to 21. In all, 14 functions were evaluated: Normal, Log-Normal, Weibull, Exponential, Gamma, Pearson III, Beta (Singh, 1998; Walck, 2007), Logistic, Moyal, Gumbel, Cauchy, Chi-square, Rayleigh, Maxwell (Walck, 2007). The adjustment method used was least squares and the quality of the adjustments was evaluated using the Kolmogorov-Smirnov methodology. The Kolmogorov-Smirnov (K-S) test consists of the non-parametric analysis of two univariate and continuous distributions (Stephens, 1970; Press, 1992). The K-S method starts from the comparison between a critical difference of the cumulative distribution and the model with the theoretical critical parameter associated with the desired significance. All calculations performed in the steps described above were performed in Excel 2013 software (64 bits).

Normal: $f(x) = \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\dfrac{\left(x-\mu\right)^2}{2\sigma^2}\right)$ (8)

Log-Normal: $$f(x)=\frac{1}{x\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right) \quad (9)$$

Moyal: $$f(x)=\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}+\exp\left(-\frac{x-\mu}{\sigma}\right)\right)\right] \quad (10)$$

Gumbel: $$f(x)=\frac{1}{a}\exp\left[-\frac{x-b}{a}-\exp\left(\frac{x-b}{a}\right)\right] \quad (11)$$

Weibull: $$f(x)=\frac{\eta}{b}\left(\frac{x}{b}\right)^{(\eta-1)}\exp\left[-\left(\frac{x}{b}\right)^{\eta}\right] \quad (12)$$

Gama: $$f(x)=a(ax)^{b-1}\frac{1}{\Gamma(b)}\exp(=ax) \quad (13)$$

Rayleigh: $$f(x)=\frac{1}{a^2}\exp\left[-\frac{1}{2}\left(\frac{x}{a}\right)^2\right] \quad (14)$$

Maxwell: $$f(x)=\frac{1}{a^3}\left(\frac{2}{\pi}\right)^{\frac{1}{2}}x^2\exp\left[-\frac{1}{2}\left(\frac{x}{a}\right)^2\right] \quad (15)$$

Logistic: $$f(t)=\frac{1}{k}\exp\left(\frac{t-\alpha}{k}\right)\left(1+\exp\left(\frac{t-\alpha}{k}\right)\right)^{-2} \quad (16)$$

Cauchy: $$f(x)=\frac{1}{\pi(1+x^2)} \quad (17)$$

Pearson III: $$f(x)=\frac{1}{a\Gamma(b)}\left(\frac{x-c}{a}\right)^{b-1}\exp\left[-\left(\frac{x-c}{a}\right)\right] \quad (18)$$

Exponential: $$f(x)=\frac{1}{a}\exp\left[-\left(\frac{x}{a}\right)\right] \quad (19)$$

Chi-Square: $$f(x)=\left(\frac{x}{2}\right)^{\left(\frac{a}{2}-1\right)}\frac{1}{2\Gamma\left(\frac{a}{2}\right)}\exp\left(-\frac{x}{2}\right) \quad (20)$$

Beta: $$f(x)=\frac{1}{\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}}x^{(a-1)}(1-x)^{(b-1)} \quad (21)$$

3. Results and Discussion

In Figure 1, the time series of NOx concentration for the years 2010 to 2015 are presented. It is verified that there is no clear variation or tendency around an average value and noises are observed. In order to extract more information about the behavior of the series, Table 1 shows the mean, standard deviation, variance, skewness and kurtosis calculations for each of the years studied. It is quite evident the lack of homogeneity in the data of each time series, because in all cases, it is observed that the standard deviation presents mean amplitude and a very high variance. These facts are corroborated by the fact that these measurements are highly sensitive to atmospheric conditions, sudden physical changes at measurement sites and other changes that make the series very heterogeneous and unpredictable in the long run. With regard to the skewness, it is possible to notice a situation of asymmetry in all cases, with right or positive displacement. In the case of the

kurtosis coefficient, in situations where $C = 3$ the kurtosis is denominated mesocurtica (normal curve), $C > 3$ the curve is leptokurtic, and for $C < 3$ it is called the platicuric curve (Neckel, 2016). In the case of the data under study, for all series evaluated, $C > 3$, therefore are leptokurtic curves. It is important to note that, for the data under analysis, the symmetry issue is quite clear, as will be seen below, but sometimes it is visually subtle and it is difficult to assert any conclusion through simple graphical observation (Neckel, 2016).

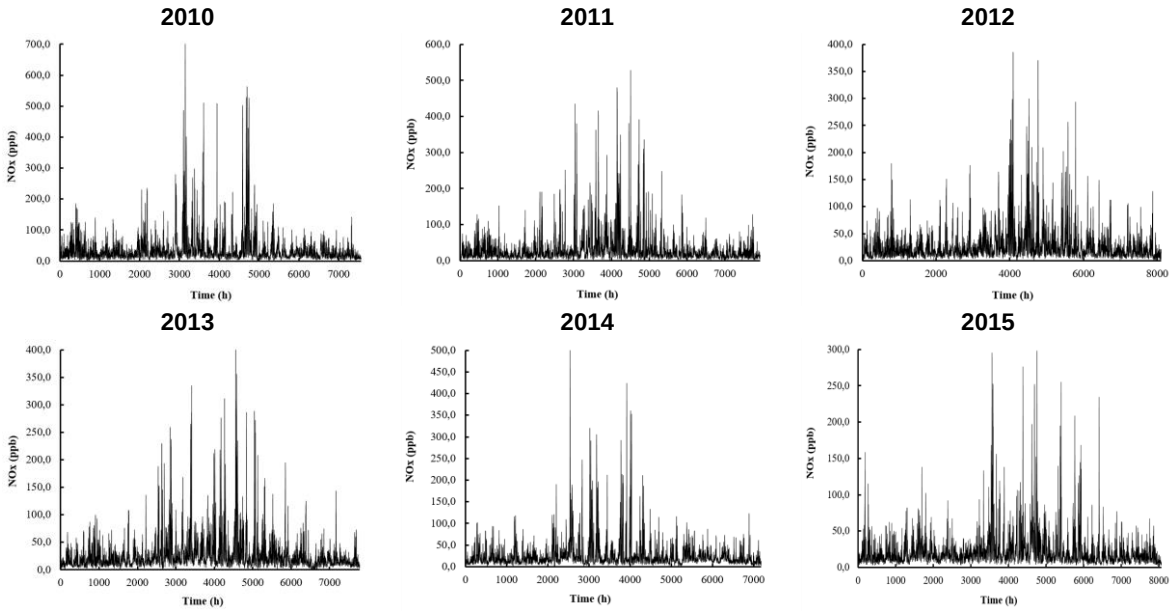

Figure 1: Time series of NOx concentration for the years 2010 to 2015.

Table 1: Mean, standard deviation, variance, skewness and kurtosis calculated for each time series.

	2010	2011	2012	2013	2014	2015
μ (ppb)	35.90	31.44	26.78	26.72	26.38	22.09
σ (ppb)	47.44	41.17	29.27	31.79	32.28	24.11
σ²	2250.23	1694.79	856.77	1010.86	1042.15	581.14
A	5.37	4.61	4.03	4.40	5.06	4.54
C	41.14	29.70	25.18	27.14	38.54	30.18

After this analysis, the optimization of the best bin to represent the distribution of the data to be adjusted was performed. In this sense, Figure 2 shows the evolution of the coefficient of variation of the average with the increase of the bins. By means of the graphical observation it is evident that there is no more significant variation in the mean for values of $K \geq 20$. Therefore, it was adopted the distribution in 20 bins in the modeling.

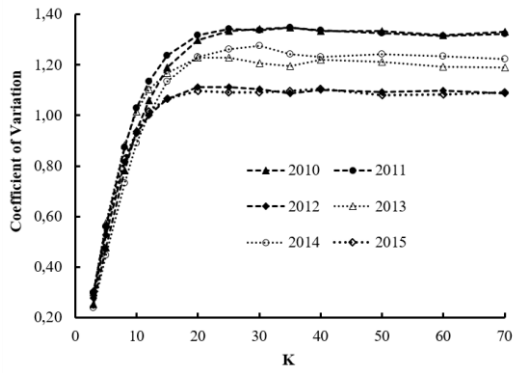

Figure 2: Evolution of the coefficient of variation of the average with the increase of the bins.

Once the bin was defined, the adjustment of the 14 probability density functions was performed using the least squares methodology, and then the quality of the adjustments was verified through the Kolmogorov-Smirnov

test and the evaluation of the sum of the quadratic errors ($S(e^2)$). Table 2 presents the K-S test results for the probability density functions that best fit the data and also the parameter values of the models. In this table it is possible to compare the critical deviation value (Dc), at a significance level of 1%, with the value of the maximum deviation (D) for each adjusted functions. In cases where $D < Dc$, the adjustment is approved according to the K-S criteria. Approved functions are highlighted in the table. It is also possible to observe that all the adjustments approved by the K-S test presented very low values for the maximum quadratic error ($S(e^2)$), thus corroborating such adjustments as being the best for the respective series.

Table 2: Results of the Kolmogorov-Smirnov test and the parameters of the approved models.

	Weibull	Gumbel	Moyal	Log-Normal	Gama	Weibull	Gumbel	Moyal	Log-Normal	Gama
Dc	**0.019**	**0.050**	**0.078**	**0.015**	**0.035**	**η; b**	**a; b**	**μ; σ**	**μ; σ**	**a; b**
2010	0.014	0.037	0.031	0.069	0.041	0.92	19,85	4.92	-	-
$S(e^2)\times10^5$	5	64	82	12,212	23,785	26.27	0	8.29	-	-
2011	0.018	0.047	0.060	0.085	0.087	0.97	16.69	4.61	-	-
$S(e^2)\times10^5$	30	77	95	11,6690	226,814	22.78	0	8.18	-	-
2012	0.039	0.020	0.042	0.013	0.048	-	16.18	4.39	2.40	-
$S(e^2)\times10^5$	33	29	26	6	37	-	0	8.84	0.82	-
2013	0.018	0.037	0.014	0.092	0.245	1.15	15.89	5.10	-	-
$S(e^2)\times10^5$	420	38	33	9,596	158,391	21.12	0	8.50	-	-
2014	0.010	0.037	0.039	0.083	0.123	0.96	14.80	4.33	-	-
$S(e^2)\times10^5$	11	52	64	11,787	255,862	19.73	0	7.19	-	-
2015	0.098	0.048	0.113	0.012	0.027	-	11.85	-	2.69	0.093
$S(e^2)\times10^5$	36	29	44	7	28	-	4.63	-	0.74	1.59

Finally, Figure 3 shows the comparison of the adjustments of the best probability density functions with the respective original series (Experimental). By means of this figure it is possible to note, again, the leptokurtic tendency with positive displacement of both the adjusted models and the experimental data. Also, the small differences between the experimental profile and those of the adjusted functions are more clearly demonstrated, since in this reconstruction, different from what occurs with the cumulated frequency, the errors are no longer damped by the accumulation of frequencies, so that discrete discrepancies become sharper. Such discrepancies do not invalidate the results, since the representativeness presented by the models is higher than expected given the heterogeneity of the series. It is important to make it clear that the only probability density function that was repeated for all cases was Gumbel suggesting, therefore, that this is the most generic PDF to express the data studied.

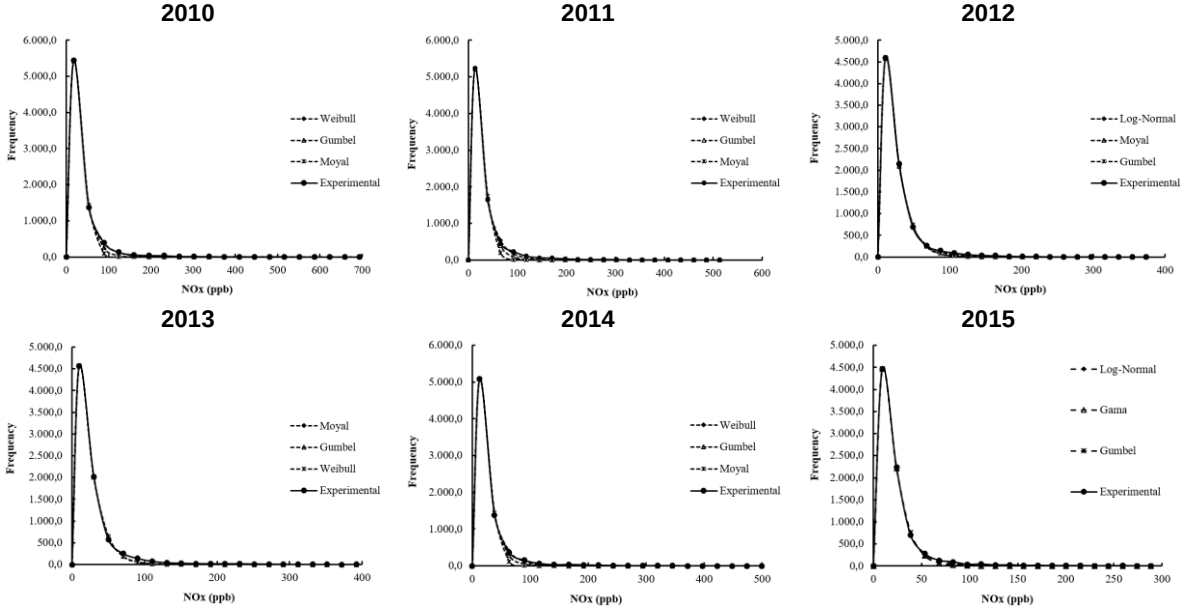

Figure 3: Comparison of fitted models with experimental data.

4. Conclusion

Through the study, it can be concluded that all the time series studied presented skewness with positive displacement and leptokurtic curve. After extensive analysis of the variation of the bin, the best grouping occurred for K = 20. Regarding the adjustment of the best model of probability distribution, it was verified that for each year a set of PDFs are satisfactorily adjusted, but the Gumbel model appeared in all evaluated years, suggesting that this is the most generic PDF to express the atmospheric NOx concentration variation in the monitoring station of Ibirapuera Park, São Paulo, Brazil.

List of Symbols

Latin Symbols

A – skewness;
a, b, c – PDF parameters.
C – Kurtosis;
E(X) – *expected value*;
f(x) – probability density function.
f_j – frequencies of appearance of a certain value;
K – bin;
N – total number of points in a set;
n – order of the moving average;
X – element of a random sample;
x_j – data that will compute the mean.

Greek Symbols

μ – Arithmetic mean of a set*;*
σ – mean standard deviation;
σ^2 – variance.

References

Braga, A. L. F.; Conceição, G. M. S.; Pereira, L. A. A.; Kishi H.; Pereira, J. C. R.; Andrade, M. F.; et al. Air pollution and pediatric respiratory hospital admissions in Sao Paulo, Brazil. J Environ Med.1:95-102, 1999.

CETESB. Relatório Anual de Qualidade do Ar no Estado de São Paulo. Companhia de Tecnologia de Saneamento Ambiental. São Paulo, SP, 2004.

Derisio, J. C. Introdução ao controle de poluição ambiental. São Paulo: CETESB; 1992.

Dockery, D. W.; Pope III, C. A. Acute respiratory effects of particulate air pollution. Annu Rev Public Health 5:107-32, 1994.

Harikrishna, M. and Arun, C. “Stochastic analysis for vehicular emissions on urban roads — a case study of Chennai,” in Proceedings of the 3rd International Conference on Environmental and Health, M. J. Bunch, V. M. Suresh, and T. V. Kumaran, Eds., Chennai, India, December 2003.

Kan, H.-D. and Chen, B.-H. Statistical distributions of ambient air pollutants in Shanghai, China, Biomedical and Environmental Sciences, vol.17, no.3, pp.366–372, 2004

Neckel, V, J. Estatística I. Joinville: Sociesc, 2016.

Oguntunde, P.E; Odetunmibi, O.A. and Adejumo, A.O. A Study of Probability Models in Monitoring Environmental Pollution in Nigeria. Journal of Probability and Statistics. Volume 2014, Article ID 864965, 6 pages, 2014.

Pope, C. A. I.; Burnett, R. T.; Thun, M. J.; Calle, E. E.; Krewski, D.; Ito, K.; Thurston, G. D. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. J. Am. Med. Assoc. 287, 1132-1141, 2002

Press, W. H., Teukolsky, S. A., Vetterling, W. T.; Flannery, B. P. Numerical Recipes in C. 2nd Edition. Cambridge University Press. 1992.

Singh, V.P., 1998, Entropy-Based Parameter Estimation in Hydrology, Dordrecht: Springer. 368p.

Stephens, M. A. Use of the Kolmogorov-Smirnov, Cramer-von Mises and related statistics without extensive tables. Journal of the Royal Statistical Society, Series B 32:115-122, 1970.

Walck, C. Handbook on statistical distributions for experimentalists, Internal Report. Universitet Stockholm, 188p, 2007.