CHEMICAL ENGINEERING TRANSACTIONS
VOL. 43, 2015
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.aidic.it/cet
Chief Editors: Sauro Pierucci, Jiří J. Klemeš

ISBN 978-88-95608-34-1; **ISSN** 2283-9216
DOI: 10.3303/CET1543361

Using Ensembles of Artificial Neural Networks to Improve PM_{10} Forecasts

Romulo M. S. Souza, Guilherme P. Coelho*, Ana Estela A. da Silva, Simone A. Pozza

School of Technology (FT), University of Campinas (Unicamp), Rua Paschoal Marmo, 1888, Limeira – SP, 13484-332, Brazil.
guilherme@ft.unicamp.br

High concentrations of atmospheric pollutants provoke negative effects that range from respiratory problems in humans to altered growth in crops due to the reduction of solar radiation. In this context, the study of suspended particulate matter (PM) in the atmosphere is especially relevant. Several works in the literature are dedicated to evaluate PM impacts and to develop models to forecast PM concentrations. Among these models, artificial neural networks (ANNs) are often employed mainly due to the facts that they are capable of learning from a set of training data samples and that they are known to be universal function approximators. However, most ANN training algorithms are susceptible to initial conditions, so the resulting models of distinct training phases may present different accuracies for the same problem. It is known from the machine learning literature that the ensemble approach, which basically combines a set of slightly different high-accuracy predictors, tends to lead to more accurate forecasts. Therefore, in this paper an ensemble of ANNs is proposed to forecast the daily concentrations of PM_{10} ($\phi \leq 10$ μm) in the city of Piracicaba, Brazil. The ensemble was trained with daily samples collected from 07.2009 to 06.2013 and evaluated with one-day-ahead forecasts from 07.2013 to 06.2014. Experiments with distinct ANN configurations were made and an average reduction of 8.85 % was obtained in the Mean Squared Error. The ensembles were compared to individual ANNs that led to the best accuracy in the training dataset. It was also verified that, when compared to distinct single ANNs, the ensemble-based approach facilitated the generation of high accuracy models, as it increased the robustness of the development process. It is important to highlight that the proposed approach can be directly applied to other scenarios related to the prediction of PM concentrations, such as different atmospheric pollutants and meteorological data.

1. Introduction

Atmospheric pollutants may cause several negative impacts that range from health issues in humans to reduced growth in crops. In such scenario, suspended particulate matter (PM) is particularly important. From all different types of particulate matter, the focus of this work will be on PM_{10}, which are particles with aerodynamic diameter smaller than or equal to 10 μm. When particles with this size are inhaled, they can reach the lower respiratory tract, causing diseases and even death. Therefore, environmental agencies and public offices generally seek for tight control of PM_{10} concentration in urban areas. In order to previously act before PM_{10} concentrations rise above safety thresholds, techniques capable of forecasting these values are very important, and this work focuses on a specific tool: ensembles of artificial neural networks.
In the Southeast region of Brazil, particularly in the State of São Paulo, the sugarcane culture is one of the most important sources of PM_{10} (Uriarte et al., 2009), together with suspended soil dust and traffic. Among the main sugarcane production centres in São Paulo, the city of Piracicaba has a history of high concentrations of PM_{10} originated from a combination of sugarcane burning and vehicular and industrial emissions (Lara et al., 2005; CETESB, 2014). Therefore, PM_{10} concentration data, hourly collected from 01.07.2009 to 30.06.2014 in the city of Piracicaba, was considered in the study performed in this work.

Please cite this article as: Souza R., Coelho G., Silva A.E., Pozza S.A., 2015, Using ensembles of artificial neural networks to improve pm10 forecasts, Chemical Engineering Transactions, 43, 2161-2166 DOI: 10.3303/CET1543361

Datasets of particulate matter concentration can often be considered as time series, since they correspond to a collection of values evaluated periodically. Therefore, all techniques proposed in the literature for time series analysis and prediction (Prado and West, 2010) can be directly applied to PM data. Several authors have adopted this approach, such as Chen et al. (2012), who developed a two-stage spatiotemporal model to predict the daily concentration of $PM_{2.5}$ in the metropolitan area of Taipei, Goyal et al. (2006), who applied three statistical models (linear regression, ARIMA and a combination of both previous models) to forecast the daily average concentration of respirable suspended PM in urban areas of Delhi and Hong Kong, and Qin et al. (2014), who combined several time series forecast techniques to infer PM concentrations in four major cities of China.

One of the techniques adopted by Qin et al. (2014) were artificial neural networks (ANNs), more specifically back-propagation artificial neural networks. ANNs are a computational paradigm whose mechanisms are designed to model the way the human brain performs particular tasks (Haykin, 2009). This paradigm is constituted by several distinct techniques, developed to deal with a wide scope of tasks such as data clustering, classification, regression, associative memory and others. Given that, ANNs devoted to deal with regression problems are particularly relevant to the context of this work, as they are known to lead to good results in time series forecast problems. Therefore, a multitude of authors have applied ANNs to forecast PM concentrations in distinct scenarios, such as He et al. (2014), who combined Chaotic Particle Swarm optimization with ANNs to forecast street level PM, and Ordieres et al. (2005), who developed an ANN-based model to forecast $PM_{2.5}$ on the US-Mexico border.

The main advantage of the ANNs employed for time series forecast is that they are capable of learning from a set of examples (training data) or, in other words, of adjusting themselves in order to map the provided set of inputs into the corresponding set of outputs. Therefore, if the data samples of the training dataset properly represent the whole problem, such trained ANN may be able to generalize and correctly forecast unseen data samples. Besides, some types of ANNs (such as Multi-layer Perceptrons – MLPs) are known to be universal function approximators, which means that such networks, given an adequate configuration, are capable of approximating any arbitrary continuous function (Haykin, 2009). Although ANNs present these theoretical advantages, it is not always trivial to obtain an ANN that properly generalizes for a given problem (for further details, see Section 2.2). Therefore, several different approaches have been proposed to tackle this issue. Among them, the ensemble-based approach has been successfully adopted in many contexts (Hansen and Salamon, 1990).

An ensemble basically consists of a pool of high quality and diverse components (neural networks, in the context of this work), whose individual outputs are combined into a single output for each sample of the dataset. The general idea behind ensembles, which can be seen as a machine committee, is that distinct high-quality components may learn different aspects of the problem and, when combined, may be able to better generalize to previously unseen data samples. Such better generalization capability has been both empirically and theoretically demonstrated (Hansen and Salamon, 1990; Hashem et al. 1994).

Given the high quality results obtained with ensembles and reported in the machine learning literature, authors from the PM field have also adopted this technique in some contexts of the area. Zhou et al. (2014) proposed the use of a hybrid ensemble-based approach to decompose $PM_{2.5}$ concentration data to perform one-day-ahead forecasts. Djalalova et al. (2010) created several ensembles that incorporate different meteorological models, chemical mechanisms and emission inventories to forecast surface ozone and PM during the 2006 Texas Air Quality experiment. In Djalalova et al. (2010), the ensembles were built by different combination of the individual models. Chen et al. (2013) combined single stepwise regression and a mixture of wavelet transformation with stepwise regression in an ensemble to forecast PM_{10} concentrations in eastern China. Debry and Mallet (2014) proposed an ensemble technique to combine several machine learning algorithms to forecast ozone, nitrogen dioxide and PM_{10} concentration on the Prev'Air platform. The individual forecasts of each component of the ensemble in Debry and Mallet's work were then sequentially aggregated and combined by a weighted average approach. Siwek and Osowski (2012) combined several types of neural network predictors to forecast the daily average concentrations of PM_{10} in Warsaw, Poland. In their approach, the neural network predictors were combined with a wavelet decomposition strategy, which led to several individual prediction results that were later combined in an ensemble.

All previously discussed works obtained good results with the ensemble approach in the context of PM forecast. However, their focus was to obtain the best possible forecasts for their particular problems and, to do so, a multitude of different predictors and combination strategies were employed. The goal of this work was slightly different: to evaluate exactly what were the benefits (if any) of an ensemble-based approach to forecast PM_{10} concentrations. The best possible prediction accuracies were still aimed, but the focus was also in verifying whether ensembles introduce additional benefits to the process of PM forecast or even possible drawbacks. To do so, variations of Multilayer Perceptrons (MLPs) were chosen as the individual predictors to perform one-day-ahead forecast of PM_{10} for the city of Piracicaba, Brazil.

This paper is organized as follows. Section 2 presents further conceptual aspects about ensembles and the methodology adopted in this work. The experimental results are presented and discussed in Section 3, and the final conclusions and future steps of this research are given in Section 4.

2. Conceptual Aspects and Methodology

This section presents the theoretical concepts and the methodology adopted in this study.

2.1 Site Description and Data Collection

The city of Piracicaba (Figure 1) has an area of 1,378.50 km^2, approximately 365,000 inhabitants (IBGE, 2010) and a population density of 264,47 inhabitants/km^2. Although the city can be mainly characterized by agricultural activities such as sugarcane (Lara et al., 2005) and pasture, industrial activities, also mostly associated with agriculture, are increasing (Barretto et al., 2006). There are about 152,000 vehicles in circulation (CETESB, 2014). According to the Köppen classification, the climate in Piracicaba is classified as Cwa (humid subtropical, mild with hot summer and dry winter).
The climate has strong influence on air pollution problems, mainly in particulate matter (PM): Akyüz and Çabuk (2009) reported that ambient temperature and local wind conditions influence the concentration of PM_{10} and $PM_{2.5}$. It is also known from the literature that traffic (Penconek et al., 2013) and local resuspended soil dust represent a significant part of PM emission in several parts of the world, such as Northern Greece (Samara et al., 2003; Terzi et al., 2010) and Brazil (Pérez-Martínez et al., 2014). In Piracicaba, the same situation occurs: according to the 2013 Air Quality Report (CETESB, 2014), the main PM_{10} sources in the region may be related to emissions from industrial processes, vehicles and sugar/ethanol production. Therefore, Piracicaba can be considered as a representative of the midsized cities of Southeastern Brazil, which includes municipalities such as Ribeirão Preto, São José do Rio Preto and São Carlos, where rural industrial activity is equivalent to the urban industrial activity. This justifies the importance of monitoring the atmospheric quality and the choice of Piracicaba as the object of this study.
The PM_{10} data studied in this work was collected by CETESB's (the Environmental Agency of São Paulo, Brazil) automatic station in Piracicaba, and is available online at the QUALAR platform (QUALAR, 2014).

2.2 Ensembles of Artificial Neural Networks

As previously mentioned, ensembles of artificial neural networks, particularly of multilayer perceptrons (MLPs - Haykin, 2009), have been studied in this work, so that the advantages and disadvantages of this approach to predict the concentration of PM_{10} in urban areas could be evaluated.
In order to be effective and improve generalization, an ensemble must contain components that are diverse in their errors, which means that such components must error differently in their predictions. Figure 2 illustrates how diversity affects the generalization of an ensemble whose components are combined through a simple average. In Figure 2, both predictors (A and B) are correctly adjusted to the training data samples (used to adjust the predictors), but they fail to approximate function $f(x)$ for values of x not seen during the training phase. Therefore, it is possible to say that predictors A and B do not generalize well. It is also possible to notice that predictors A and B are diverse in their errors, as they result in distinct approximations of function $f(x)$. When these two predictors are combined into an ensemble, the resulting approximation of $f(x)$ (curve close to the dashed line in Figure 2) is significantly better than that of its individual components. Although Figure 2 depicts an illustrative example, it is clear in the literature that diversity is an important aspect of ensembles. Therefore, many techniques to stimulate diversity have been proposed. In this work, the widely adopted approach known as bagging was used (Breiman, 1996).
Bagging is a sampling technique intended to generate different training datasets for each predictor of the ensemble, so that each component can learn slightly different aspects of the problem and end up being different from the others. Therefore, considering a training dataset with N examples, bagging uniformly samples $M \leq N$ examples (with replacement) for each predictor.

2.3 Dataset Preparation and Ensemble Architecture

In this work, PM_{10} concentration data, hourly collected from 01.07.2009 to 30.06.2014 in the city of Piracicaba, was considered. To prepare this dataset for the experiments, the hourly samples were converted into daily averages and divided into five sets with 365 averages each (one for each year). The values in all five sets were normalized in [-1.0, +1.0] and organized so that each MLP could use seven consecutive delays of the time series as inputs to forecast the concentration value of the following day (one-day-ahead-forecasts).

Figure 1: Location of Piracicaba, Brazil

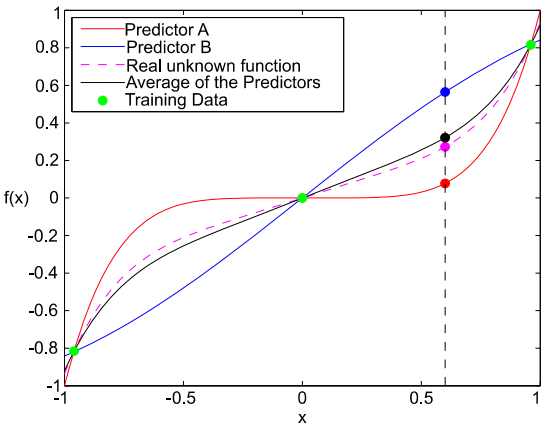

Figure 2: Illustration of the generalization capability of an ensemble with diverse components

The first four datasets were used as the full training dataset provided to the bagging module of our software (which sampled 80 % of the data for each component – MLP – of the ensemble) and the last dataset was used to evaluate the individual predictors and the ensemble itself (test dataset).

Three sets of experiments were performed here. In the first one, 10 MLPs with one hidden layer with 14 neurons were trained to perform one-day-ahead forecasts of PM_{10} concentration and combined into an ensemble with the same purpose. The second and third experimental scenarios were similar to the first one, except that the MLPs were configured to have 7 and 4 hidden neurons, respectively. All MLPs were configured with hyperbolic tangent activation function in the neurons of the hidden and output layers. The training algorithm was the Manhattan Propagation and the results were evaluated with respect to the Mean-Squared Errors (MSEs) both in training and test datasets.

3. Experimental Results

The MSEs obtained for the individual MLPs and for the ensembles (Table 1) indicate that the combination of high quality and diverse components into an ensemble leads to improvements when compared to the best component (MLP) of the training phase (represented in bold in Table 1). Considering the experimental rounds performed here, the ensemble-based approach led to an average improvement in MSE of 8.85 % over the best MLP, being the largest gain (18.83 %) obtained when MLPs with 4 neurons in the hidden layer were used. All these gains were obtained when the test dataset was considered.

Another important aspect that was verified in the experimental results is that the individual MLPs of each scenario presented a significant variation in MSE. Considering the test dataset, in the first scenario (14 hidden neurons) the difference between the best and worst MSEs was 87.39 (or 42.7 % of the best MSE), in the second scenario (7 hidden neurons) the difference was 29.13 (or 14.1 % of the best MSE) and in the third scenario (4 hidden neurons) 56.24 (or 27.9 % of the best MSE). Therefore, it is clear that the naive approach of training a single MLP and using it to forecast the concentration of MP_{10} for previously unseen data (mimicked here by the evaluation of the MLPs in the test dataset) is dangerous and may lead to high errors. A variation of this naive approach that may also be adopted by researchers is to repeat the training phase of MLPs a few times, select the best MLP obtained and use it to forecast previously unseen data. This approach is not adequate either, as the best MLPs of the training phases were never the best ones in the test phases (Table 1). In the third scenario (4 hidden neurons), for example, the best MLP in the training dataset was the worst one in the test dataset. Therefore, the combination of multiple MLPs was also capable of filtering the high variation of individual MLP performance, thus reducing the risk of ending up with a predictor that will not perform well when applied to make forecasts based on previously unseen data.

The daily average values of PM_{10} concentration collected in Piracicaba from 01.07.2013 to 30.06.2014 (test dataset) are represented in Figure 3, together with the forecasts provided by the ensemble built in the third experimental scenario. It is possible to observe several occurrences of PM_{10} concentration above the 24-h mean guideline defined by the WHO (2006), mostly in the dry season for the region (April to September). However, it is interesting to notice that, although the traditional rainy season in Southeast Brazil (October to March) has been anomalously dry in 2013/2014, the concentration of PM_{10} was reduced in the months of December to March, which indicates that the activities that emit PM_{10} were also reduced in this period. Figure 3 also indicates that the ensemble was able to properly forecast the PM_{10} concentration in the period,

including most of the peaks above the WHO guideline, which indicates that it can be considered a valuable decision support tool to help environmental agencies and public offices.

Table 1: Mean-squared errors obtained for the individual MLPs and ensembles for each experimental round (MLPs with 14, 7 and 4 neurons in the hidden layer). Gain represents the improvement of the ensemble's MSE over the best MLP (in bold, selected according to their MSE in the training dataset)

	14 Hidden Neurons		7 Hidden Neurons		4 Hidden Neurons	
	Training Data	Test Data	Training Data	Test Data	Training Data	Test Data
MLP-1	221.82	204.57	224.55	206.82	224.01	207.91
MLP-2	221.66	211.21	206.29	216.41	253.80	203.58
MLP-3	241.47	291.96	235.28	207.09	251.74	248.51
MLP-4	250.11	218.57	199.40	215.87	271.84	221.33
MLP-5	219.69	225.20	250.99	221.12	261.56	203.28
MLP-6	247.52	241.02	223.57	232.14	**203.26**	**257.27**
MLP-7	232.39	219.28	228.67	207.25	213.64	211.43
MLP-8	224.24	217.25	206.94	235.95	236.27	201.13
MLP-9	**218.72**	**216.66**	**192.89**	**220.01**	234.80	225.44
MLP-10	232.50	237.63	233.95	213.04	233.04	205.85
Ensemble	-	**209.45**	-	**210.34**	-	**208.81**
Gain	-	**3.33 %**	-	**4.39 %**	-	**18.83 %**

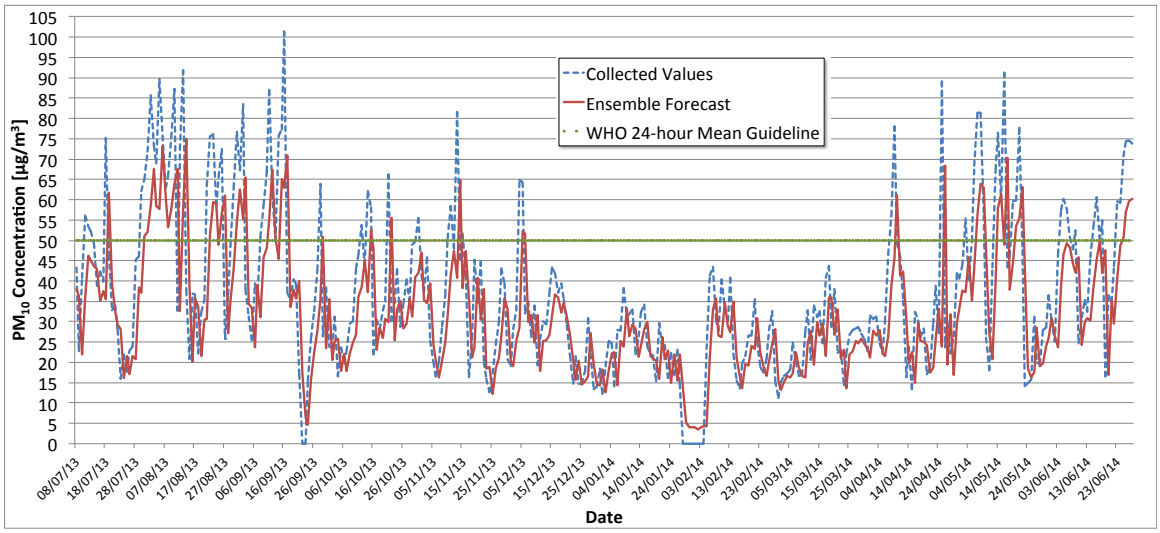

Figure 3 – Collected values of PM_{10} concentration in Piracicaba (from 07.2013 to 06.2014), together with the ensemble forecasts and WHO's guideline for 24-hour mean concentration of PM_{10} (WHO, 2006)

4. Conclusions

This paper presented a study meant to evaluate the benefits and drawbacks of ensemble-based approaches to forecast PM_{10} concentrations. To do so, three experimental scenarios in which ensembles of MLPs were built to forecast one-day-ahead concentrations of PM_{10} in the city of Piracicaba, Brazil were considered. The results have shown that ensembles not only lead to predictors with improved generalization capabilities but also filter the high variation of individual MLP performance that may occur between different training procedures. Although the ensemble approach requires an additional step to ensure diversity among the components, the results indicate that the final benefits easily overcome such extra burden. Therefore, the approach adopted here has shown to be an effective alternative to support actions to control PM_{10} concentrations in cities like Piracicaba, which presented several occurrences of PM_{10} concentrations above the maximum recommended values defined by the WHO in the considered time frame.

As future steps, the analysis performed here will be expanded to ensembles composed of distinct components, while the proposed methodology will be applied to forecast not only PM_{10} concentrations, but also other variables related to air quality monitoring. The evaluation of the results will also be extended to different regions of Brazil with environmental characteristics distinct from Piracicaba.

References

Akyüz M., Çabuk H., 2009, Meteorological variations of PM2.5/PM10 concentrations and particle-associated polycyclic aromatic hydrocarbons in the atmospheric environment of Zonguldak,Turkey, Journal of Hazardous Materials, 170, 13–21.

Barretto A.G.O.P., Sparovek G., Giannotti M., 2006, Piracicaba Rural Atlas <www.ipef.br/publicacoes/atlasrural/Atlas_Rural_de_Piracicaba_2006.pdf>, accessed in 18.10.2014 (in Portuguese).

Breiman L., 1996, Bagging predictors, Machine Learning 24(2), 123-140.

CETESB – São Paulo Environmental Agency, 2014, 2013 Air Quality Report <www.cetesb.sp.gov.br/ar/qualidade-do-ar/31-publicacoes-e-relatorios>, accessed in 08.11.2014 (in Portuguese).

Chen C-C., Wu C-F., Yu H-L., Chan C-C., Cheng T-J., 2012, Spatiotemporal modeling with temporal-invariant variogram subgroups to estimate fine particulate matter $PM_{2.5}$ concentrations, Atmospheric Environment 54, 1-8.

Chen Y., Shi R., Shu S., Gao W., 2013, Ensemble and enhanced PM_{10} concentration forecast model based on stepwise regression and wavelet analysis, Atmospheric Environment, 74, 346-359.

Debry E., Mallet V., 2014, Ensemble forecasting with machine learning algorithms for ozone, nitrogen dioxide and PM_{10} on the Prev'Air platform, Atmospheric Environment, 91, 71-84.

Djalalova I., Wilczak J., McKeen S., Grell G., Peckham S., Pagowski M., DelleMonache L., McQueen J., Tang Y., Lee P., McHenry J., Gong W., Bouchet V., Mathur R., 2010, Ensemble and bias-correction techniques for air quality model forecasts of surface O_3 and $PM_{2.5}$ during the TEXAQS-II experiment of 2006, Atmospheric Environment, 44, 455-467.

Goyal P., Chan A.T., Jayswal N., 2006, Statistical models for the prediction of respirable suspended particulate matter in urban cities, Atmospheric Environment, 40, 2068-2077.

Hansen L.K., Salamon P., 1990, Neural networks ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence, 12(10), 993–1001.

Hashem S., Schmeiser B., Yih Y., 1994, Optimal linear combinations of neural networks: An overview. Proc. of the IEEE Int. Conf. on Neural Networks, Orlando, USA.

Haykin S., 2009, Neural Networks and Learning Machines. Pearson/Prentice Hall, 3rd. Ed., Upper Saddle River, USA.

He H., Lu W-Z., Xue Y., 2014, Prediction of particulate matter at street level using artificial neural networks coupling with chaotic particle swarm optimization algorithm, Building and Environment, 78, 111-117.

IBGE – Brazilian Institute of Geography and Statistics, 2010, Cities <cod.ibge.gov.br/234XG>, accessed 18.10.2014 (in Portuguese).

Lara L.L., Artaxo P., Martinelli L.A., Camargo P.B., Victoria R.L., Ferraz E.S.B., 2005. Properties of aerosols from sugar-cane burning emissions in Southeastern Brazil. Atmospheric Environment, 39, 4627–4637.

Ordieres J.B., Vergara E.P., Capuz R.S., Salazar R.E., 2005, Neural network prediction model for fine particulate matter ($PM_{2.5}$) on the US-Mexico border in El Paso (Texas) and Ciudad Juárez (Chihuahua), Environmental Modelling & Software, 20, 547-559.

Penconek A., Zgiet B., Sosnowski T.R., Moskal A., 2013, Filtering of DEP (Diesel Exhaust Particles) in Fibrous Filters. Chemical Engineering Transactions, 32, 1987-1992.

Pérez-Martínez P.J., Miranda R.M., Nogueira T., Guardani M.L., Fornaro A., Ynoue R., Andrade M.F., 2014, Emission factors of air pollutants from vehicles measured inside road tunnels in São Paulo: Case Study Comparison, International Journal of Environmental Science and Technology, 11(8), 2155–2168.

Prado R., West M., 2010, Time Series: Modeling, Computation, and Inference. Chapman & Hall/CRC, Boca Raton, USA.

Qin S., Liu F., Wang J., Sun B., 2014, Analysis and forecasting of the particulate matter (PM) concentration levels over four major cities of China using hybrid models, Atmospheric Environment, 98, 665-675.

QUALAR, 2014, Air Quality Database < qualar.cetesb.sp.gov.br> accessed 01.07.2014 (in Portuguese).

Samara C., Kouimtzis T., Tsitouridou R., Kanias G., Simeonov V., 2003, Chemical mass balance source apportionment of PM_{10} in an industrialized urban area of Northern Greece. Atmospheric Environment, 37, 41–54.

Siwek K., Osowski S., 2012, Improving the accuracy of prediction of PM_{10} pollution by the wavelet transformation and an ensemble of neural predictors, Engineering Applications of Artificial Intelligence, 25, 1246–1258.

Terzi E., Argyropoulos G., Bougatioti A., Mihalopoulos N., Nikolaou K., Samara C., 2010, Chemical composition and mass closure of ambient PM_{10} at urban sites. Atmospheric Environment, 44, 2231-2239.

Uriarte M., Yackulic C. B., Cooper T., Flynn D., Cortes M., Crk T., Cullman G., McGinty M., Sircely J., 2009, Expansion of sugarcane production in São Paulo, Brazil: Implications for fire occurrence and respiratory health. Agriculture, Ecosystems and Environment, 132, 48–56.

WHO – World Health Organization, 2006, WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide – Global update 2005 – Summary of risk assessment <www.who.int/phe/health_topics/outdoorair/outdoorair_aqg/en/>, accessed in 14.11.2014.

Zhou Q., Jiang H., Wang J., Zhou J., 2014, A hybrid model for $PM_{2.5}$ forecasting based on ensemble empirical mode decomposition and a general regression neural network, Science of the Total Environment 496, 264–274.