Contents lists available at ScienceDirect

Science of the Total Environment

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

Intercomparison of air quality data using principal component analysis, and forecasting of PM_{10} and $PM_{2.5}$ concentrations using artificial neural networks, in Thessaloniki and Helsinki

Dimitris Voukantsis[a], Kostas Karatzas[a,*], Jaakko Kukkonen[b], Teemu Räsänen[c], Ari Karppinen[b], Mikko Kolehmainen[c]

[a] Department of Mechanical Engineering, Informatics Applications and Systems Group, Aristotle University, P.O. Box 483, GR-54124, Thessaloniki, Greece
[b] Finnish Meteorological Institute, P.O. Box 503, 00101, Helsinki, Finland
[c] Department of Environmental Science, University of Eastern Finland, P.O. Box 1627, 70211, Kuopio, Finland

ARTICLE INFO

Article history:
Received 2 February 2010
Received in revised form 11 August 2010
Accepted 30 December 2010
Available online 26 January 2011

Keywords:
Air pollution
Artificial neural networks
Atmospheric quality comparison and forecasting
Multi-layer perceptron
Particulate matter
Principal component analysis

ABSTRACT

In this paper we propose a methodology consisting of specific computational intelligence methods, i.e. principal component analysis and artificial neural networks, in order to inter-compare air quality and meteorological data, and to forecast the concentration levels for environmental parameters of interest (air pollutants). We demonstrate these methods to data monitored in the urban areas of Thessaloniki and Helsinki in Greece and Finland, respectively. For this purpose, we applied the principal component analysis method in order to inter-compare the patterns of air pollution in the two selected cities. Then, we proceeded with the development of air quality forecasting models for both studied areas. On this basis, we formulated and employed a novel hybrid scheme in the selection process of input variables for the forecasting models, involving a combination of linear regression and artificial neural networks (multi-layer perceptron) models. The latter ones were used for the forecasting of the daily mean concentrations of PM_{10} and $PM_{2.5}$ for the next day. Results demonstrated an index of agreement between measured and modelled daily averaged PM_{10} concentrations, between 0.80 and 0.85, while the kappa index for the forecasting of the daily averaged PM_{10} concentrations reached 60% for both cities. Compared with previous corresponding studies, these statistical parameters indicate an improved performance of air quality parameters forecasting. It was also found that the performance of the models for the forecasting of the daily mean concentrations of PM_{10} was not substantially different for both cities, despite the major differences of the two urban environments under consideration.

1. Introduction

The issue of air quality (AQ) is of major concern for many European citizens and one of the areas in which the European Union has been most active, in order to take preventive and regulatory measures. The Clean Air For Europe (CAFE) initiative has led to a thematic strategy setting out the objectives and measures for the next phase of European AQ policy. The resulting Directive (CAFE Directive — 2008/50/EC) underlines the need for a common framework of methods and criteria that will allow for a direct comparison of the AQ in different member states, as well as the forecasting and management of AQ. Moreover, the CAFE Directive includes mandates for the provision of specific information to the public, concerning concentrations of air pollutants, incidents of exceedances of atmospheric quality criteria, and their predictions for the next day(s). Given the aforementioned requirements, it is necessary to employ methods and tools that both analyze and model air pollution while in parallel they can extract knowledge in terms of similarities, differences and interdependencies of the studied AQ parameters.

On this basis, we propose a methodology consisting of Computational Intelligence (CI) methods that can be applied to AQ and meteorological data recorded in different cities, in order to identify and compare the air pollution profiles of the urban environment under consideration and to evaluate the potential for the forecasting of certain parameters of interest. We demonstrate the capabilities and the insight that the proposed methodology provides by considering the case of two European cities that share certain similarities and differences, i.e. Helsinki (in Finland) and Thessaloniki (in Greece).

The suitability of CI methods for the aforementioned tasks has been suggested by their inherent characteristics. Among others, CI methods: (i) are data-oriented, thus there is no need for strong assumptions during the modelling process, (ii) are computationally efficient, thus it is possible to handle huge amounts of data in little time (iii) can be applied for different tasks such as knowledge

* Correspondence author. Tel./fax: +30 2310994176.
E-mail address: kkara@eng.auth.gr (K. Karatzas).

0048-9697/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.scitotenv.2010.12.039

extraction, forecasting, etc. and (iv) can be easily integrated into existing information systems that could communicate useful information to citizens.

Earlier research work has already dealt with CI methods within the AQ domain, mainly for episode identification (Kolehmainen et al., 2000) and incident, i.e. threshold value exceedances, via the forecasting of target air pollutants such as O_3 (Bordignon et al., 2002; Barrero et al., 2006; Karatzas and Kaltsatos, 2007), NO_2 (Kalapanidas and Avouris, 2001; Nagendra and Khare, 2006), and PM_{10} (Slini et al., 2006; Niska et al., 2005; Kukkonen et al., 2003; Zickus et al., 2002). Additional applications of CI methods for knowledge extraction in the AQ domain have also appeared in literature (Ibarra-Berastegi et al., 2009; Oanh et al., 2010). For the city of Helsinki, Kukkonen et al. (2003) evaluated and inter-compared the performance of five neural network models, a linear statistical model and a deterministic modelling system for the prediction of urban NO_2 and PM_{10} concentrations, measured at two stations in central Helsinki. On the other hand, in Thessaloniki there have been studies such as Slini et al. (2006) that made use of linear regression (LIN-REG) models, classification and regression trees (CRTs) and artificial neural networks (ANNs) in order to forecast PM_{10} concentrations, while Tzima et al. (2007) have evaluated the performance of several classifiers in forecasting hourly PM_{10} concentration values.

The methods selected for the present study are principal component analysis (PCA) and artificial neural networks — multi-layer perceptron (ANN-MLP). Although these methods have been partially applied for similar purposes in Helsinki and Thessaloniki in previous studies, the objectives now are fundamentally different: (i) we consider these methods as parts of a structured framework that can be applied to analyze and inter-compare the statistical inter-dependencies of AQ and meteorological data measured in different cities, in order to reveal the similarities and differences at the air pollution profiles of the selected urban environments and associate the corresponding results with the prevailing air pollution sources and mechanisms of the cities. (ii) We apply a novel hybrid method in order to optimize the input variables of the forecasting models. This method is based on the evaluation of the actual performance (on the basis of certain statistical indices) of linear regression (LIN-REG) and ANN-MLP models. Such a method has not been previously presented in literature. (iii) We evaluate the performance of the resulting, optimized, ANN-based, AQ forecasting model mentioned above, both in Helsinki and Thessaloniki, by using the same methodology.

An additional novelty of this study is that ANNs have not previously been evaluated in different climatic and geographic regions, based on their AQ modelling performance, and by using a common framework and underlying methodology. We also aimed to find out whether the mathematical analysis, including ANNs, could result in new information concerning the source- and process-oriented patterns of air pollution and the meteorology in the studied cities. We addressed solely the forecasting of PM_{10} and $PM_{2.5}$ (where available) concentrations in this study, as those concentrations are frequently high in both cities, compared with the national guidelines and the EU limit values (Voutsa et al., 2002; Moussiopoulos et al., 2009, Kukkonen et al., 2005, Kauhaniemi et al., 2008).

2. Materials and methods

2.1. The selected cities and their surrounding regions

The Helsinki Metropolitan Area and its surroundings are situated on a fairly flat coastal region by the Baltic Sea at a latitude of 60.2°. The Helsinki Metropolitan Area comprises of four cities (Helsinki, Espoo, Vantaa and Kauniainen) with a total coverage of 743 km^2 and a population of approximately one million inhabitants. On the other hand, Thessaloniki is the second largest city of Greece and one of the most densely populated cities in Europe, accounting for approximately 16,000 inhabitants per km^2. Located at a latitude of 40.4° the city covers an area of approximately 18 km^2, while the urban web of the Greater Thessaloniki Area covers approximately 93 km^2. It is located in the inner part of the Thermaikos Gulf, surrounded in the northerly and north-easterly directions by the Hortiatis mountain, while numerous residential suburbs surround the city and an extended industrial zone is situated to the north-west of its outskirts. The studied areas as well as the location of the relevant AQ monitoring sites are presented in Fig. 1.

2.2. The common characteristics and differences of the selected cities

In this study, we have addressed the AQ for the cities of Helsinki and Thessaloniki, with the aim to investigate and inter-compare the AQ characteristics of a northern and a southern coastal European city. These cities have the following common characteristics: (a) they are both comparable in terms of the amount of population, which is of the order of magnitude of one million inhabitants for both urban regions; (b) both of them are coastal cities, having the seafront to their south; and (c) both of the selected urban regions represent a partly maritime-influenced and partly continental climate.

However, these urban regions also possess substantial differences: (a) the city of Thessaloniki is surrounded by mountainous areas that

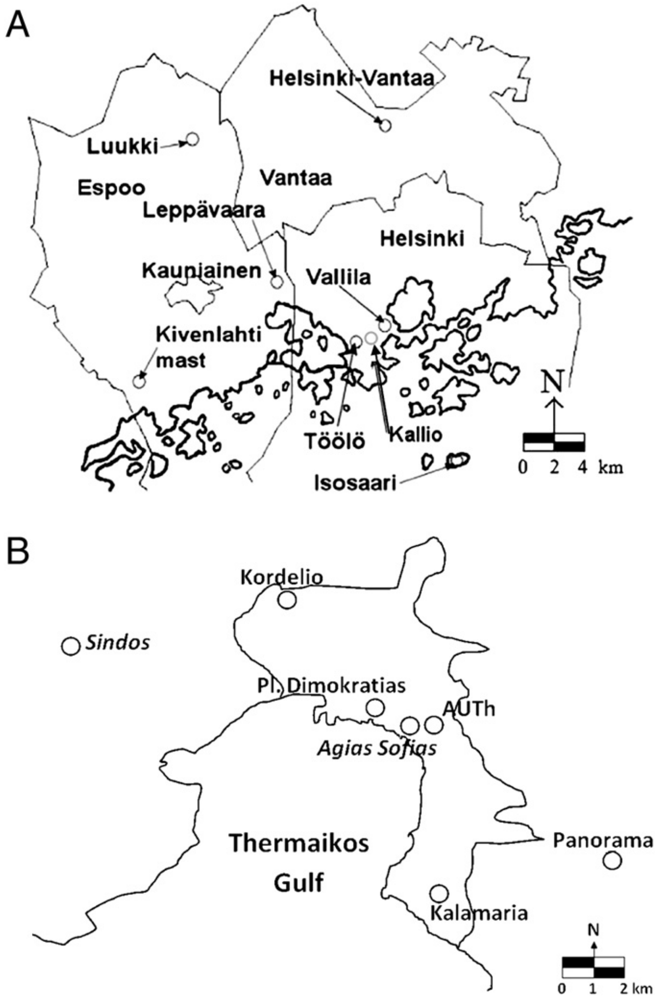

Fig. 1. Location of the AQ and meteorological measurement sites in the Helsinki Metropolitan Area (left) and in Thessaloniki (right). The sites of Helsinki–Vantaa, Isosaari and Kivenlahti mast are meteorological stations; the other ones are AQ stations. Helsinki, Vantaa, Espoo and Kauniainen are cities. Furthermore, the AQ/MET stations of Panorama, Kalamaria, AUTh, Kordelio and Sindos, as well as the AQ stations of Agias Sofias and Pl. Dimokratias.

influence the mesoscale atmospheric circulation and the regional characteristics of the climate. The Helsinki Metropolitan Area on the other hand is surrounded by a flat terrain, with only moderately high hills in the surrounding regions. (b) The climatic conditions for Helsinki are mostly continental with maritime influences, while relatively milder, compared to most other areas of the same latitude, mainly because of the Gulf Stream and the prevailing global atmospheric circulation. The climate of Thessaloniki is typically Mediterranean: mild winters, and atmospheric circulation commonly influenced by the sea breeze. (c) The density of population is substantially higher in the central areas of Thessaloniki, compared to those in Helsinki. Moreover, the percentage of open and green spaces is less than 5 m^2 per inhabitant for the Greater Thessaloniki Area (Moussiopoulos and Nikolaou, 2008), in comparison to 134 m^2 per inhabitant in Helsinki (City of Helsinki, 2007).

The aforementioned similarities and differences result in a distinct AQ situation for each one of the cities. Clearly, the climatic characteristics affect the prevailing air chemistry and atmospheric diffusion processes. A good example is ozone: whereas it is a prominent pollutant in Thessaloniki, it is of minor importance in Helsinki. However, both cities are substantially influenced by the local vehicular traffic, with the latter being responsible for high concentrations of Particulate Matter (PM) and NO_X that in some cases exceed the limit values set by the regulatory framework. For the city of Helsinki model computations suggest that approximately 80–95% of the ground level NO_X concentrations originated from traffic sources (Karppinen et al., 2000a,b), while for the city centre of Thessaloniki local vehicular traffic is identified to be the main source of PM pollution (Manoli et al., 2002). The contribution of the paved-road dust from multiple sources to the overall PM emissions in Thessaloniki was found to be of 28% for fine particles and 57% for coarse particles (Manoli et al., 2002), thus suggesting that re-suspension is an important contributor to high PM concentrations. On the other hand, predicted contribution from long-range transport to the $PM_{2.5}$ street level in Helsinki varied spatially from 40% in the most trafficked areas to nearly 100% in the outskirts (Kauhaniemi et al., 2008).

2.3. The AQ and meteorological monitoring stations

For Helsinki, measurement data from the stations of Kallio and Vallila have been used, while for Thessaloniki, data from the stations of Sindos and Agias Sofias were selected. The aim was to identify two stations per city, one influenced by traffic in the city centre (Vallila and Agias Sofias for Helsinki and Thessaloniki, respectively), and an urban background station (Kallio and Sindos for Helsinki and Thessaloniki, respectively). The urban background sites are representative of the average exposure of urban population to pollution, whereas the traffic sites represent the relatively more severely polluted urban environments.

The measurement height at both stations in Helsinki is 4.0 m. The station of Vallila is situated in a park at a distance of 14 m from the edge of the Hämeentie road. The average weekday traffic volume of Hämeentie was 13,000 vehicles/day in 2001. The heights of the buildings in the vicinity of the station, at the other side of the Hämeentie road and surrounding the park, range from 10 to 15 m. The Hämeentie road is fairly wide; there are four lanes for cars and additionally two lanes for trams.

The station of Kallio is located at the edge of a sports ground. The busiest streets in the vicinity of the station are Helsinginkatu at a distance of 80 m and Sturenkatu at a distance of 300 m. The average weekday traffic volume of Helsinginkatu was 7800 vehicles/day in 2001.

The PM_{10} and $PM_{2.5}$ concentrations in Helsinki were measured by continuous analyzers (Eberline FH 62 I-R) using β-attenuation at both AQ stations addressed in this study. Analyzers are manually calibrated twice a year while they are monitored with regular automatic span and zero check. This continuous method is not one of the official EU reference methods, and therefore, results of the continuous method have been compared with the results of Kleinfiltergerät that is one of the EU reference methods. The comparison was conducted at the station of Vallila from autumn 2000 to summer 2001 (Sillanpää et al., 2002) and indicated a good agreement of the results, thus no correction coefficients were required.

The Agias Sofias station in Thessaloniki is situated on a wide traffic island between two roads formulating the Ermou Street in the centre of Thessaloniki. The measuring height is 4 m, while the estimated daily traffic volume in the city centre (in the nearby Egnatia Str, 150 m from the monitoring site) is 60,000 vehicles/day (Assael et al., 2008). The station is within a street canyon, with a horizontal distance of 15 m from buildings that are approx. 30 m high. On the contrary, the station at Sindos is located in the premises of the higher technological educational institute of Thessaloniki, in an open area, approx. 200 m from a rural road, and 1.5 km away from a highway.

The PM_{10} concentrations in Thessaloniki were measured by analyzers (Eberline FH 62 I-R) that use β-attenuation at both AQ stations addressed in this study. All monitoring equipment is located inside temperature and humidity-controlled shelters. The air samples are analyzed on-line and in real-time and the data are collected automatically every 1 min and stored in the data logger of each station. Everyday an operator monitors the data flow from monitoring sites to a central data collecting point and takes care of data quality issues. In order to ensure that the data produced are accurate and reliable, strict maintenance, operational and quality assurance/control procedures are carried out every month. These are followed to a concrete protocol to allow comparability between monitoring sites.

In addition to AQ data, the meteorological data from the stations of Helsinki–Vantaa (about 15 km north of central Helsinki) and Helsinki–Isosaari (an island about 20 km south of central Helsinki) were taken into account for the Helsinki Metropolitan Area. For the Thessaloniki metropolitan area, meteorological data for the city centre were only available (and used) coming from the AUTh station (approx 900 m from Agias Sofias station), while the station of Sindos also monitors meteorological parameters, which were used in this study.

The monitoring stations in Helsinki are operated by the Helsinki Metropolitan Area Council (YTV). For a detailed description of their location, characteristics and instrumentation, the reader is referred to www.ytv.fi/eng/airquality. The stations in Thessaloniki are operated by the Prefecture of Central Macedonia, Environment Department. More details are available at www.rcm.gr (in Greek) and reported by Tzima et al. (2007).

Available atmospheric data included hourly measurements of concentrations and meteorological parameters for the time period 2001–2003 (three years). It should be noted that more recent, sufficiently comprehensive, properly quality assured and controlled datasets were not available meanwhile for both cities. However, for the main pollutants addressed (PM_{10} and $PM_{2.5}$), there have been no substantial trends since 2001 in terms of the percentages of exceedances of guidelines and limit values (in case of PM_{10}), or the average concentrations (in case of both PM_{10} and $PM_{2.5}$) for both cities according to the official web sites of their AQ information systems (www.airthess.gr for Thessaloniki and www.ytv.fi/eng/airquality for Helsinki).

Table 1 presents the AQ and meteorological parameters used in this study, as well as the percentage of incomplete data rows (defined as the sets of data, for which at least one value is missing) per year. The percentage of the latter ones has been overall low, with few exceptions, e.g., time-series of 2001 for the Vallila station; however, it should be noted that the minimum data capture criterion for PM_{10} data, i.e., more than 90% completeness within one year of data, was fulfilled for all the years and stations under consideration, thus fulfilling data quality criteria posed by the EU CAFE directive. In addition, the availability of PM_{10} and $PM_{2.5}$ for the case of Helsinki

Table 1
AQ and meteorological parameters available for each station at Helsinki and Thessaloniki. The column on the right-hand-side represents the percentages of the incomplete data rows per year (2001, 2002 and 2003, respectively). The incomplete data rows were defined as the sets of data, for which at least one value is missing from at least one parameter. The AQ parameters in the table within brackets were calculated on the basis of other parameters, i.e. CP = Coarse Particles = $PM_{10} - PM_{2.5}$, and NO = Nitrogen Monoxide = $NO_X - NO_2$.

Station	Air quality	Meteorological	% of incomplete data rows (per year)
Kallio (urb./back.)	NO_2, NO_X, (NO), O_3, PM_{10}, $PM_{2.5}$, (CP)	Temperature, relative humidity, wind direction, wind speed	5.79 3.74 2.87
Vallila (urb./traf.)	CO, NO_2, NO_X, (NO), PM_{10}, $PM_{2.5}$, SO_2, (CP)	Temperature, relative humidity, wind direction, wind speed	43.01 2.33 18.28
Sindos (urb./back.)	CO, NO_2, O_3, PM_{10}, SO_2	Temperature, relative humidity, wind direction, wind speed	19.27 4.06 20.91
Agias Sofias (urb./traf.)	CO, NO_2, O_3, PM_{10}, SO_2	Temperature, relative humidity, wind direction, wind speed	13.16 8.44 31.31

made possible the distinction between fine ($PM_{2.5}$) and coarse particles (PM_{10}–$PM_{2.5}$). This was appropriate, since these PM mass fractions are commonly associated with different emission source categories and atmospheric processes. It is therefore useful also in the modelling to separate these particulate matter measures.

Furthermore, the sine and cosine transformation were employed for the Wind Direction, while for temporal parameters, such as Month and Day of Week, similar transformations were used:

$$\sin x = \sin(2\pi(x - min(x) / (max(x) - min(x))$$

$$\cos x = \cos(2\pi(x - min(x) / (max(x) - min(x)).$$

In the above equations x is a temporal variable, and 'sinx' and 'cosx' the corresponding transformations (i.e. these are not equal to sin(x) and cos(x)). Finally, incomplete data rows were excluded from the data sets, as well as clearly erroneous values.

2.4. The principal component analysis (PCA) method

PCA is a CI method originating from multivariate statistical analysis, which allows for the identification of the major factors within a certain multidimensional dataset. It may also be applied for expressing the data in a way proper for highlighting their similarities and differences (Jolliffe, 2002). Although the capabilities of PCA are limited, as it represents a linear combination of parameters, it has been successfully applied for several tasks in the AQ domain during the last decades (Smeyers-Verbeke et al., 1984; Chavent et al., 2008), and is capable of identifying interrelations within the studied parameters. For the purposes of this study, PCA was considered as a tool capable of providing an overview of the interdependencies and variability of data, thus allowing for extracting information on the AQ mechanisms and for the comparison of the air pollution profiles of the cities under consideration.

On this basis, we applied the PCA method for the data described in Table 1, and resulted in a new set of uncorrelated variables, the principal components (PCs). By selecting the most significant PCs, it is possible to identify certain relationships between the parameters of the dataset under consideration for both studied areas, and reach to conclusions concerning AQ characteristics for each area, as well as the possible emission sources for some of the pollutants under consideration. PCA was combined with the parallel analysis criterion (Franklin et al., 1995) in order to identify the number of significant PCs, while the Varimax Rotation method (Lewis-Beck, 1994) was applied in order to allow for an easier interpretation of the results.

2.5. The artificial neural networks — multi-layer perceptron (ANN-MLP) models

ANN-MLP models are flexible capable of approximating any smooth differentiable function, thus being particularly applicable when modelling complex non-linear processes. ANN-MLP models have been utilized for several tasks within the AQ domain, such as forecasting, function approximation and pattern classification (e.g., Gardner and Dorling, 1999, Kukkonen et al., 2003).

2.5.1. Data preparation for the ANN-MLP training

In the current study, ANN-MLP models have been applied in order to forecast daily mean concentrations of particulate matter in the cities of Thessaloniki and Helsinki. Daily mean values are selected as forecasted variables instead of hourly values, due to the requirements issued by EU legislation (Directive 2008/50/EC). The dataset of daily resolution is derived by the corresponding ones of hourly resolution that was presented in Section 2.3. The variables evaluated as potential inputs for the ANN-MLP models have been suggested by previous studies (Kukkonen et al., 2003; Niska et al., 2005) and our understanding of atmospheric processes, whereas data availability limitations needed to be taken into account. On this basis, we have evaluated (i) the concentration values of pollutants being monitored: daily mean, minimum and maximum hourly averaged concentrations at each station and (ii) the mean, minimum and maximum hourly averaged meteorological values of temperature, relative humidity and wind speed, in a day. Two kinds of meteorological input values were also used: values measured during the day when the forecast is produced (denoted as T) and those on the next day (T + 1). The latter ones are used as proxy variables of the meteorological forecasts. In an actual application of this method, one would use meteorological data forecasted using a numerical weather prediction model (Niska et al., 2005).

Furthermore, days with less than 12 h of recordings and 4 continuous hours of missing data are excluded from the training and testing process of the ANNs. Then, the data are normalized by applying variance scaling and are split into a training set (2 years of data) and test set (1 year of data). In the next step, and for reasons of completeness, all 3 possible data combinations are evaluated (i) Data Set 1 — DS1: years 2002 and 2003 used for training and year 2001 used for testing; (ii) Data Set 2 — DS2: years 2001 and 2003 used for training and year 2002 used for testing; (iii) Data Set 3 — DS3: years 2001 and 2002 used for training and year 2003 used for testing. The ANN-MLP models were implemented by using the Neural Networks toolbox for Matlab, as well as the Mtools for Matlab provided by the Group of Environmental Informatics (Dept. of Environmental Science, University of Eastern Finland).

2.5.2. Selection of input variables for the ANN-MLP models

The selection of input variables for an ANN-MLP forecasting model is a key issue, as irrelevant or noisy variables may have negative effects on the training process, resulting to an unnecessarily complex model structure and poor generalization power. In earlier studies, several different input selection algorithms have been applied (Eleuteri et al., 2005; Kohavi and John, 1997; Niska et al., 2006). Moreover, sensitivity analysis, correlation analysis, multi-objective genetic algorithms and information geometric approaches have been applied in order to identify the optimal set of model inputs. However, these methods have their own advantages and disadvantages, therefore suitability should be carried out in case-specific analysis. In the current paper, the selection process of the input variables was based on the performance of actual regression models (linear and ANN regression models).

More specifically, the possible input variables of the models are the ones described in Section 2.5.1. The final selection is based on a novel hybrid optimization procedure followed in this study that involved a grid search algorithm and the actual performance (Index of Agreement, IA, see Appendix A) of LIN-REG and ANN-MLP models used as the objective function. During the selection process a large number of models are being developed, each one using different input variables, where all variable combinations are tested. The comparison between models performance is based on the 95% confidence intervals of the IA that results from 1000 repetitions of the calculation procedure via 1000 bootstrap samplings. Bootstrapping is a method which can be applied in order to determine whether we can trust the validity indicators, especially when using ANNs, where the output depends on the random initializations of weights. Furthermore, the bootstrapping is a non-parametric method and does not require any strong assumptions about the process or modelling problem (Efron and Tibshirani, 1993).

The selection process can be separated into two phases. During the first one, the objective function used for the model selection is derived by the performance of LIN-REG models. The search for potential input variables identifies the best possible predictors; however, due to the limited capability of an LIN-REG model to simulate complex processes the selection is probably not the optimum one. Therefore, during the second phase we make use of ANN-MLP models in order to evaluate the selection made during the first phase, and improve it if possible by further introducing input variables to the model.

On this basis, not only all available parameters were tested as candidates for becoming the input parameters of the ANN models, but this test was repeated for a large number of times in order to assure that their selection was not biased. Furthermore, we experimented with various ANN-MLP architectures by running the selection process with different number of neurons (from $N/2-2$ up to $N/2+2$, where N is the number of input variables), although we restricted the number of hidden layers to one. Finally, the selection process was repeated for all three datasets (DS1, DS2 and DS3), and the input variables resulting to best performing models were included at the final models.

This approach does not mathematically prove that the final ANN-MLP models constructed are the ones with the best performance, i.e. correspond to the global minimum of the error surface. Nevertheless, the developed hybrid method (being a mixture of linear and non-linear models), suggests an automated computational procedure, i.e. an algorithm, that certainly leads to quantifiable better results in terms of the performance evaluation of ANN_MLP models. This is supported by the fact that the performance of the resulting models is among the best ones reported in literature for daily averaged PM concentrations forecasting with the aid of CI methods.

2.5.3. ANN-MLP specifications and evaluation

Concerning the ANN-MLP model specifications and the way that their results are evaluated, it should be mentioned that the training process of the ANN-MLP models was based on the resilient back-propagation algorithm. Furthermore, the hyperbolic tangent sigmoid transfer function was used for the hidden layer and the linear function for the output layer. The evaluation of the final forecasting models was based on statistical indices, such as the IA, correlation coefficient (R), root mean square error (RMSE) and the Cohen's kappa (Cohen, 1960), also known as KI. For the latter index the forecasted concentrations (numerical) were transformed into a binary variable indicating exceedances (0: no, 1: yes) of the PM limit values. More details about the statistical indices are included in the Appendix A.

3. Results and discussion

3.1. Data presentation

The data used in this study corresponded to the time period 2001–2003. Table 2 presents basic statistics about the parameters studied at both cities.

The urban traffic stations demonstrated higher concentrations of pollutants, with the exception of O_3, for which concentrations were found to be higher at the urban background stations. Low concentrations of O_3 at the roadside stations were caused by the depletion of ozone in the oxidation of traffic-originated nitrogen oxides. The PM_{10} concentrations at the stations in Thessaloniki were much higher compared to those at the stations in Helsinki; mean values were found to be 2–3 times higher in Thessaloniki, while the ratio of the mean value to standard deviation was similar in both cities. Fig. 2 presents the hourly concentrations of PM_{10} at the selected urban traffic stations in Helsinki and Thessaloniki, i.e. Vallila and Agias Sofias, respectively.

Table 2
Basic statistics for the AQ and meteorological parameters available for each station at Helsinki and Thessaloniki.

Helsinki, Kallio 2001–2003 (data rows: 26,233)					Thessaloniki, Ag. Sofias 2001–2003 (data rows: 26,269)				
Pollutant	Min	Max	Mean	Std	Pollutant	Min	Max	Mean	Std
NO_2 (μg/m^3)	0	143.0	24.3	15.7	NO_2 (μg/m^3)	0	232.0	48.7	28.5
NO_X (μg/m^3)	0	397.0	34.7	32.6	CO (mg/m^3)	0	12.3	1.6	1.1
O3 (μg/m^3)	0	156.0	46.8	23.8	O_3 (μg/m^3)	0	200.0	40.9	32.8
PM_{10} (μg/m^3)	0	157.2	16.3	12.8	PM_{10} (μg/m^3)	1.0	431.0	65.9	40.3
$PM_{2.5}$ (μg/m^3)	0	81.3	8.5	6.9	SO_2 (μg/m^3)	0	234.0	25.5	24.3
Temp (°C)	−31.0	30.5	5.6	10.2	Temp (°C)	−6.2	37.8	16.9	8.4
RH (%)	17.0	100.0	76.0	17.8	RH (%)	16.0	103.0	61.8	17.4
WS (m/s)	1.0	15.4	4.4	2.1	WS (m/s)	0	9.8	2.0	1.4
Helsinki, Vallila 2001–2003 (data rows: 26,228)					Thessaloniki, Sindos 2001–2003 (data rows: 26,243)				
Pollutant	Min	Max	Mean	Std	Pollutant	Min	Max	Mean	Std
CO (mg/m^3)	0	1.7	0.3	0.2	CO (mg/m^3)	0	2.2	0.4	0.2
NO_2 (μg/m^3)	0	137.1	28.1	17.5	NO_2 (μg/m^3)	0	163.0	18.5	15.5
NO_X (μg/m^3)	0.8	618.0	50.6	50.5	O_3 (μg/m^3)	0	173.0	59.5	39.3
PM_{10} (μg/m^3)	0	191.0	20.0	16.0	PM_{10} (μg/m^3)	1.1	379.0	50.0	32.0
$PM_{2.5}$ (μg/m^3)	0	91.6	9.6	7.8	SO_2 (μg/m^3)	0	187.0	11.6	16.0
SO_2 (μg/m^3)	0	59.0	4.1	5.3	Temp (°C)	−12.1	36.3	15.4	8.6
Temp (°C)	−31.0	30.5	5.6	10.2	RH (%)	15.0	105.0	74.7	23.1
RH (%)	17.0	100.0	76.0	17.8	WS (m/s)	0	11.5	2.3	1.7
WS (m/s)	1.0	15.4	4.4	2.1	–	–	–	–	–

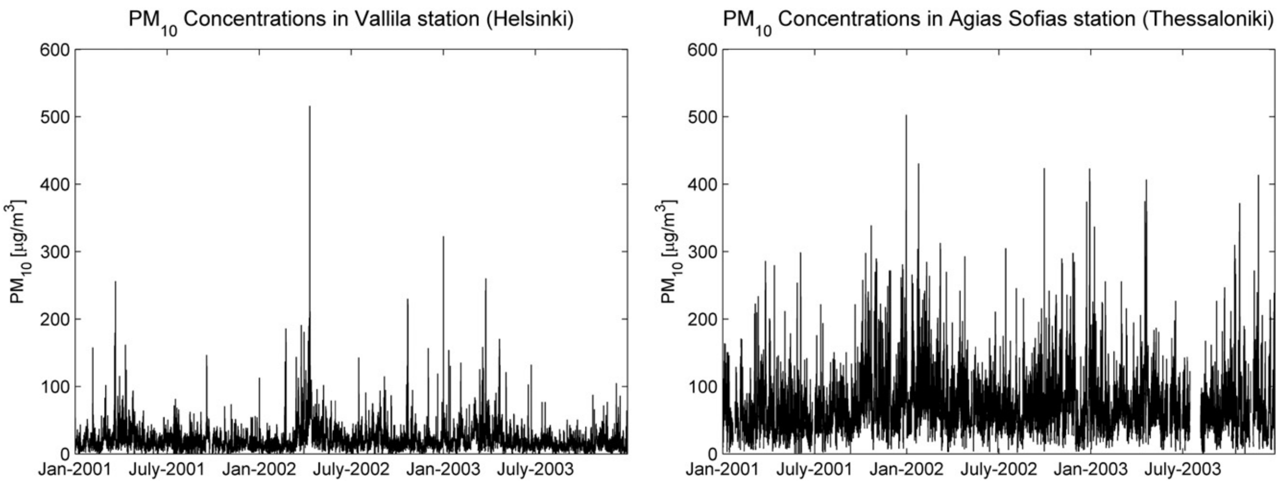

Fig. 2. Hourly PM_{10} concentrations (μg/m^3) for the traffic stations of Vallila in Helsinki (left) and Agias Sofias in Thessaloniki (right) during the time period 2001–2003.

3.2. Inter-comparison of the air pollution for Helsinki and Thessaloniki

PCA was applied to the data recorded at each one of the stations under consideration. In all cases, 4 PCs were identified by the parallel analysis criterion to correspond to non-random variation. The results of PCA are summarized in Table 3, where only the most significant PC contributions (PC coefficients > 0.2) are presented. The contribution of a variable to a PC can be either positive or negative, depending on the sign of the corresponding PC coefficient.

The first PC (PC1) is similar for all stations, corresponding to 20%–26% of the overall data variations. PC-1 consists of positive contributions of AQ parameters associated with traffic-originated emissions (CO, NO_2, PM_{10}, $PM_{2.5}$ and SO_2), and negative contributions of O_3 (which is consumed during the oxidation of traffic-originated NO) and wind speed. Furthermore, the percentage of the overall data variation explained by this PC is higher for the traffic stations (Vallila and Ag. Sofias) and smaller for the urban background stations (Kallio and Sindos), indicating that the data variations expressed by this PC can be attributed mainly to Local Traffic, which may thus be identified as the major local pollution source category in both cities.

The second PC is characterized mostly by positive contributions of O_3, coarse particles (for the stations in Helsinki) and temperature, as well as negative contributions of relative humidity and SO_2 (for the stations in Thessaloniki). The percentages of data variation explained by this PC ranges from 15% for traffic stations up to 20% for urban background stations. The contributions of the variables on the particular PC, indicate the influence of the O_3 formation mechanism (mostly for Thessaloniki), and the importance of re-suspended coarse particles (for Helsinki), while the contribution of SO_2 concentrations can be attributed to central heating (for Thessaloniki). The high concentrations of re-suspended coarse particles from street and road surfaces in Helsinki occur mainly in spring and summer; this explains the correlations with temperature (at both stations) and partly with O_3 concentrations that are commonly higher in spring and summer. Thus, this PC expresses the specific seasonal characteristics within the data.

Although, the 3rd and 4th PCs indicated some similarities, they cannot be directly associated to a particular atmospheric mechanism or emission source category, with the exception of the 4th PC for the Sindos station, where SO_2 concentrations are associated with certain

Table 3
PCA results for the stations in Helsinki (Kallio and Vallila) and Thessaloniki (Sindos and Ag. Sofias). Contributions (positive or negative) are indicated for each PC and station, as well as the overall data variation percentage explained by the particular PC. The percentages within the brackets correspond to the yearly variations of the aforementioned percentage (available only for the PCs that remained unchanged for the seasonal subsets).

		PC1	PC2	PC3	PC4
Kallio	%	20.39 (21.09–25.70)	19.80 (13.24–19.76)	14.69	13.90
	(+)	NO_2, CP, $PM_{2.5}$	CP,O_3, Temp	$PM_{2.5}$, O_3, Temp	NO_2, $PM_{2.5}$, CP, sinWD, WS
	(−)	O_3, WS	RH	cosWD	Temp
Vallila	%	24.12 (21.26–28.84)	14.68 (13.20–19.06)	13.14	12.27
	(+)	CO, NO_2, CP, $PM_{2.5}$, SO_2	CP, Temp	CO, CP, cosWD, WS	$PM_{2.5}$, SO_2, sinWD
	(−)	WS	RH	$PM_{2.5}$, Temp	
Sindos	%	21.58 (16.78–23.73)	19.13 (15.72–31.25)	22.48	11.02
	(+)	CO, NO_2, PM_{10}, SO_2	O_3, Temp	CO, NO_2, RH	SO_2, sinWD
	(−)	O_3, WS	SO_2, RH, cosWD	O_3, Temp, WS	
Ag. Sofias	%	26.14 (24.02–31.66)	16.73 (13.18–23.31)	16.12	18.32
	(+)	CO, NO_2, PM_{10}, SO_2	O_3, Temp	O_3, Temp	O_3, WS
	(−)	O_3, WS	SO_2, RH, WS, cosWD	sinWD, cosWD	RH, sinWD
Comments on PCs		Positive contribution of all air quality parameters correlated with local traffic and negative contributions from O_3 and wind speed	Positive contributions of CP, O_3 and Temperature — negative contributions SO_2 and wind directions		

wind direction, a fact that can be attributed to emissions originating from the industrial area located to the North of Sindos, thus indicating pollution transportation at local scale.

Additionally, PCA was applied to subsets of the initial data separated by season of the year (spring, summer, autumn and winter). Throughout this process it became evident that the first two PCs could be identified in all subsets, indicating the same contributions (and underlying atmospheric and emission mechanisms) of air pollution and meteorological parameters for all seasons of the year. Table 3 presents the variation of the percentage explained by each PC (indicated within brackets). The first PC, related to local traffic, indicated the minimum percentage during summers and the maximum during winters, a fact that was expected due to (i) the more unfavourable dispersion conditions in winter and (ii) overall slightly smaller amount of the traffic during the summer periods, compared to winter ones. During the summer period, the frequency of long-range PM transported episodes (that tend to statistically reduce the percentage contribution of local pollution) is higher both in Greece, due to long-range transported dust and wild-land fire plumes (Lazaridis et al., 2008; Kaskaoutisa et al., 2008), as well as in Finland, mainly due to wild-land fire plumes (Sofiev et al., 2009).

The second PC (PC-2), characterized by specific seasonal characteristics, demonstrated a minimum percentage during the winter period and a maximum percentage during summer, a fact that can be attributed to two seasonal AQ profiles: the increased O_3 formation during the summer period (mostly for Thessaloniki) and the release of re-suspended coarse particles especially in spring (for Helsinki). The yearly variations of percentages explained by each PC reinforce the initial characterization of the PCs as Local Traffic (PC-1) and Specific Seasonal Characteristics (PC-2).

The overall percentage of the data variations that PCA identified as non-random, ranged from 65% to 75%, expressed by 4 PCs. Local Traffic was found to be responsible for approximately 25% of the data variations for both cities, while 15–20% was characterized as Specific Seasonal Characteristics, consisting of different mechanisms and sources for each city.

However, a significant percentage of the data variations (PC-3 and PC-4) could not be associated to certain air emission source categories or air pollution processes, while 25%–35% of the data variations were identified by PCA as random. These findings suggest that a significant amount of the data variations requires additional analysis, in order to reveal the importance and relative influence of each parameter to the AQ characteristics of the studied areas.

A major limitation of the PCA method is that it is not capable of resolving strongly non-linear relationships. Clearly, some physical interpretation of the PCs is also required, and it is not in all cases a straightforward task to associate these reliably with specific source categories or processes. We have therefore continued the analysis of the AQ and meteorological datasets by applying ANN models, with the additional aim of developing efficient operational AQ forecasting models for PM_{10} and $PM_{2.5}$ (wherever the latter being available).

3.3. Forecasting of PM_{10} and $PM_{2.5}$ using the ANN-MLP models

3.3.1. Selection of input variables for the ANN-MLP models

On the basis of the procedure presented in Section 2.5.2, the most appropriate input variables, i.e., the ones leading to the best performing ANN-MLP forecasting models, were selected. The number of the input variables during the first phase (LIN-REG) of the selection

Table 4
Sets of input variables that were selected for the ANN-MLP models for the forecasting of PM concentrations at Helsinki and Thessaloniki. The mean, min and max refer to the mean, minimum and maximum hourly averaged values in a day, while sin and cos refer to the sine and cosine transformations of the corresponding variables (Section 2.1). Two kinds of meteorological input values are used: values measured during the day, when the forecast is produced (denoted as T) and those of the next day (T + 1).

Station	Pollutant	Input variables selected				Total
		TIME	AQ	MET (T)	MET (T + 1)	
Kallio	PM10	Month (sin, cos) Weekday (sin, cos)	NO_2 (min) NO (mean, min, max) O_3 (max) PM_{10} (mean, max)	RH (min) WS (mean, max) WD (sin, cos)	RH (T + 1) WS (T + 1)	18
Kallio	PM2.5		NO_2 (min) O_3 (mean, min) PM_{10} (min) $PM_{2.5}$ (mean)	Temp (mean) RH (mean, max) WS (max) WD (cos)	Temp (T + 1) WS(T + 1)	12
Kallio	CP	Month (sin, cos) Weekday (sin)	NO_2 (mean) O_3 (max) $PM_{2.5}$ (min) CP (mean, min, max)	RH (mean, max)	Temp (T + 1) RH(T + 1)	13
Vallila	PM10	Month (cos) Weekday (sin, cos)	CO (mean) NO_2 (min) NO (mean) PM_{10} (mean)	Temp (max) RH (min) WS (max)	RH (T + 1) WS (T + 1)t	12
Vallila	PM2.5	Month (sin)	CO (mean) NO_2 (mean, max) NO (max) PM_{10} (min) $PM_{2.5}$ (mean, min) SO_2 (min, max)	WD (cos)	WS (T + 1)	12
Vallila	CP	Month (sin, cos) Weekday (sin, cos)	CO (min, max) NO_2 (max) NO (mean, max) CP (mean)	Temp (mean) RH (max) WD (sin)	RH (T + 1) WS(T + 1)	15
Sindos	PM10	Month (cos) Weekday (sin, cos)	CO (mean, max) NO_2 (min, max) O_3 (mean) PM_{10} (mean, min) SO_2 (mean, max)	Temp (mean) RH (min) WS (max)	Temp (T + 1) RH(T + 1) WS (T + 1)	18
Ag. Sofias	PM10	Month (cos) Weekday (sin)	NO_2 (min, max) O_3 (max) PM_{10} (mean)	Temp (min) RH (mean) WS (min, max)	Temp (T + 1) WS (T + 1)	12

process ranged from 5 to 7, whereas during the second phase the input variables were increased to 12–18 (ANN-MLP), depending on the station and PM mass fraction to be forecasted. These results suggest that the use of ANN-MLP models in the selection process, improved the performance of the models, by introducing further input variables. The latter ones had been evaluated by LIN-REG models without any improvement in terms of performance. The results are presented in Table 4.

The differences at the input variables provide with further insight concerning the mechanisms responsible for PM concentrations. In particular, the number of input variables required for the forecasting of daily mean PM_{10} concentrations at the urban background stations (Kallio and Sindos) is larger than the ones required for urban traffic stations (Vallila and Agias Sofias), while there are six common input variables in all four cases: NO_2min (minimum hourly averaged NO_2 concentration in a day), PM_{10}mean (mean PM_{10} concentration in a day), WSmax (maximum hourly averaged wind speed in a day), WSmean(T + 1) (mean daily wind speed during the day for which the forecasting is produced), cosMonth, sinWeekday. It is evident that this suggests the forecasting of PM_{10} being closely associated with local vehicular traffic. This mechanism is more evident for the traffic stations, where PM_{10} are produced, compared to the urban background stations, and could explain the difference at the required input variables.

Furthermore, for the stations in Helsinki for which separate measurements of fine ($PM_{2.5}$) and coarse (PM_{10}–$PM_{2.5}$) particles were available, it became evident that the forecasting of coarse particles (CPs) depends on input variables associated with the time of the year and week (sinMonth, cosMonth, sinWeekday and cosWeekday), while this is not the case for $PM_{2.5}$ forecasting. A physical interpretation for this is that most of the $PM_{2.5}$ concentrations are of a long-range transported origin in the Helsinki Metropolitan Area (Kauhaniemi et al., 2008), while coarse PM is mostly originated from local vehicular traffic, and to the particle re-suspension of car tyre origination. The highest concentrations of coarse particles (and PM_{10}) in Finland commonly occur in spring, mainly due to the suspension of particulate matter from road and street surfaces, after snow has melted and roads have dried up (e.g., Kukkonen et al., 2005).

3.3.2. Models for the forecasting of PM_{10} and $PM_{2.5}$

The performance details of the final ANN-MLP models (IA and KI) are presented in Table 5. Furthermore, the performance of LIN-REG models, i.e. the reference models that make use of the same input variables as the corresponding ANN-MLP ones, are presented in Table 6. It should be noted that the IA is a parameter that varies from 0.0 (theoretical minimum) to 1.0 (perfect agreement between the observed and predicted values). However, Karppinen et al. (2000b) evaluated for comparison purposes the values of the IA assuming that the predicted values correspond to a random number distribution, with the same mean and the same overall variability as those of the measured data. In such a case, the IA varied from 0.39 to 0.41. This implies that for an extremely poor model (that predicts almost totally random values) the IA is not equal to 0.0, but approximately 0.4.

The results presented in Tables 5 and 6 indicate that ANN-MLP models outperform the corresponding LIN-REG models. However, it is evident that the performance of the models, in both cases, also depends on the dataset used, a fact that may be partly attributed to the quality of the data, especially to the percentage of missing values of the test set (see Table 1). Furthermore, the IA for models corresponding to the urban background stations is not substantially better compared to those of the traffic stations. In addition, the model performance was not found to be systematically better for either city. However, models for the forecasting of $PM_{2.5}$ concentrations indicated better performances compared to models for the forecasting of coarse particles at the station of Kallio; nevertheless, there was no substantial difference in this respect at the station of Vallila. On the other hand, the KI demonstrated a better model performance for Thessaloniki in comparison to Helsinki; it should be mentioned, however, that the value for the KI for Helsinki is not statistically reliable due to the small number of exceedances (1 and 2 during the years 2001 and 2003 respectively).

The results presented in Tables 5 and 6 indicated that the ANN-MLP models can be more successful for the forecasting of PM_{10} and $PM_{2.5}$, compared to LIN-REG models, since the statistical indices of the first ones were systematically better compared to the ones of the reference models. Furthermore, the performance of the ANN-MLP models is very satisfactory and thus they can be considered for operational use.

Concerning Thessaloniki, the models developed by Slini et al. (2006) for the simulation of daily PM10 concentrations at Thessaloniki city centre (urban traffic station), indicated an IA equal to 0.574 for classification and regression tree models, and equal to 0.515 for an ANN model, while in our paper the IA of the ANN model and for the specific station reached 0.877. Concerning Helsinki, the model performance parameters of the present study cannot be directly compared with those achieved by Kukkonen et al. (2005). There are several reasons for this: Firstly, the model evaluation work by Kukkonen et al. (2005) considered the agreement of the sequential hourly concentration time series of PM_{10}, and not daily averaged PM_{10} concentrations as in the present study. Secondly, the input data used for the ANN models was different: solely the meteorological data for the next day (T + 1) were used as input, and the set of meteorological variables was substantially more extensive, compared to the present study. Similar reasons also prohibit a direct comparison with the results of Niska et al. (2005).

Coming to the values of the KI in AQ forecasting, it should be noted that they are usually below 50% (Athanasiadis et al., 2005). On this basis, the results obtained in the current paper indicate a better overall performance than those in most of the previous published studies. In particular, the comparison of the results obtained here with

Table 5
Index of agreement and kappa index for the best performing MLP models for the forecasting of mean daily concentrations of particulate matter.

Model	Index of agreement			Kappa index		
	DS1	DS2	DS3	DS1	DS2	DS3
Kallio (PM_{10})	0.818	0.837	0.870	0.000	0.564	1
Kallio ($PM_{2.5}$)	0.837	0.820	0.870	0.000[a]/0.355[b]	0.418[a]/0.577[b]	0.279[a]/0.500[b]
Kallio (C.P.)	0.800	0.810	0.850			
Vallila (PM_{10})	0.782	0.853	0.891	0.665	0.495	0.614
Vallila ($PM_{2.5}$)	0.781	0.808	0.865	0.000[a]/0.560[b]	0.685[a]/0.526[b]	0.454[a]/0.711[b]
Vallila (C.P.)	0.737	0.840	0.866			
Sindos (PM_{10})	0.856	0.882	0.817	0.579	0.621	0.590
Ag. Sofias (PM_{10})	0.876	0.877	0.792	0.509	0.541	0.528

[a] Based on the current limit value (25 μg/m^3).
[b] Based on the new limit value (20 μg/m^3).

Table 6
Index of agreement and kappa index for LIN-REG models (reference models) for the forecasting of mean daily concentrations of particulate matter, using the same input variables as the corresponding ones of Table 5.

Model	Index of agreement			Kappa index		
	DS1	DS2	DS3	DS1	DS2	DS3
Kallio (PM_{10})	0.808	0.816	0.858	0.000	0.000	0.000
Kallio ($PM_{2.5}$)	0.809	0.801	0.849	0.000[a]/0.000[b]	0.453[a]/0.577[b]	0.000[a]/0.324[b]
Kallio (C.P.)	0.782	0.771	0.815	–	–	–
Vallila (PM_{10})	0.756	0.832	0.865	0.665	0.426	0.319
Vallila ($PM_{2.5}$)	0.764	0.794	0.851	0.000[a]/0.000[b]	0.460[a]/0.449[b]	0.451[a]/0.711
Vallila (C.P.)	0.729	0.812	0.833	–	–	–
Sindos (PM_{10})	0.805	0.816	0.858	0.504	0.501	0.480
Ag. Sofias (PM_{10})	0.829	0.842	0.747	0.348	0.474	0.465

[a] Based on the current limit value (25 μg/m^3).
[b] Based on the new limit value (20 μg/m^3).

those of other models applied for the same areas (Tzima et al., 2009, for Thessaloniki and Zickus et al., 2002, for Helsinki), demonstrated that the KI achieved is the highest among the above mentioned studies. This is expected to be mainly attributed to the novel input parameter selection procedure.

Thus, for Thessaloniki, Tzima et al. (2009), applied a number of CI and Machine Learning methods, including ANN-MLP, for the estimation of mean daily PM_{10} concentration levels. Their analysis covered a different location and a longer time interval, in comparison to the current study; however, it addressed a data set with a similar number of exceedances per year as in our paper. The highest performances, in terms of the KI, were obtained with ANN-MLP and Support Vector Machine models (KI: 0.49 and 0.51, respectively), whereas in the present study the best performance of the ANN-MLP model reached a

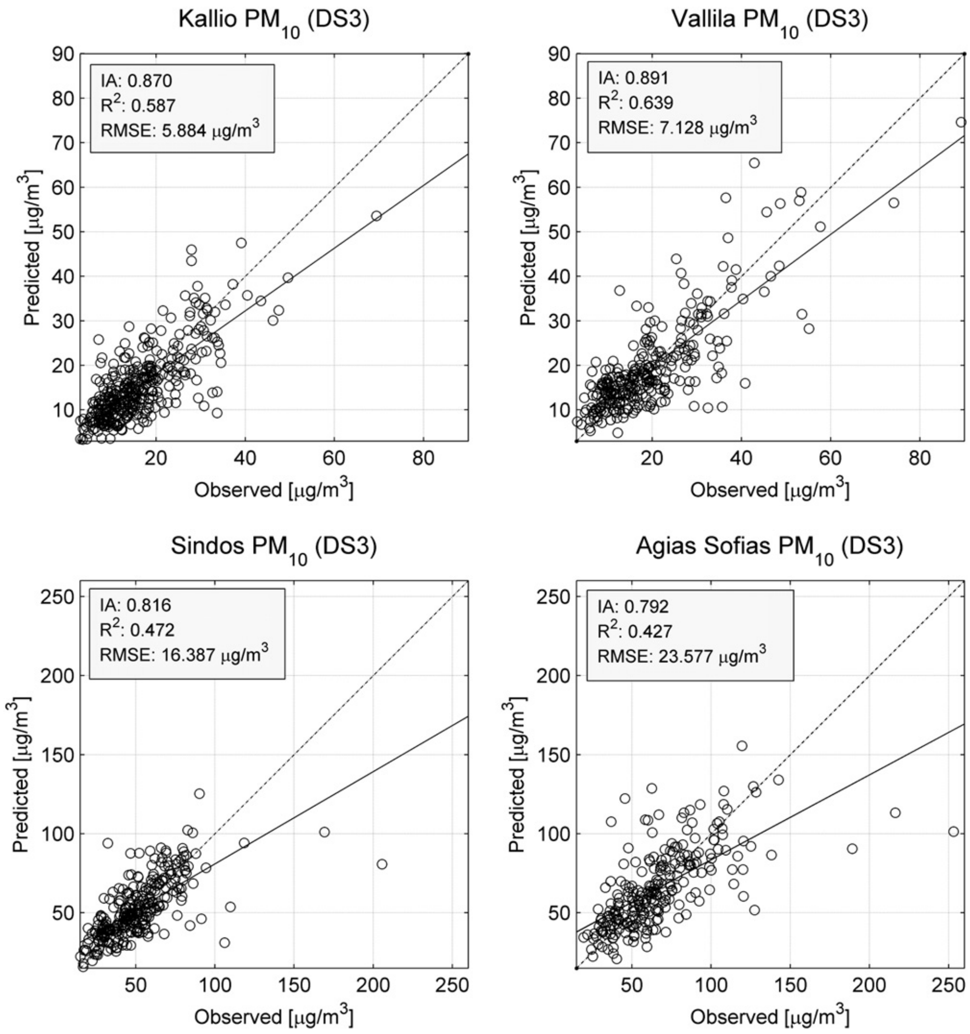

Fig. 3. ANN-predicted versus observed values of PM_{10} (μg/m^3) for DS3, i.e. forecasting of the data for the year 2003. The dashed line corresponds to perfect fit (y = x), while the solid line to the least-square fit.

KI equal to 0.62, although this result was achieved on the basis of a data set that was smaller, and thus there were less exceedances, in comparison to the data used by Tzima et al. (2009). The latter authors improved the performance of their models by applying a weighted KI scheme, thus suggesting a direction of future research.

Concerning Helsinki, Zickus et al. (2002), estimated the performance of four Machine Learning Algorithms in the Helsinki area for the estimation of mean daily PM_{10} concentration levels. Their analysis covered different locations and time intervals in comparison to the current study, and addressed a data set with a larger number of exceedances (10) in comparison to the ones addressed in the present paper (3). Their estimated equivalent of the KI (called in their paper the Index of Success), ranged between 0.38 and 0.43 for ANN models, while the best performance that they were able to achieve was an index equal to 0.47 corresponding to the multiple adaptive regression splines method (MARS). This performance was equivalent to the one achieved in the current study, although the latter one was obtained with the data that contained a substantially lower number of actual exceedances. This indicates that the methodology applied in this study results in a relatively better performance.

Fig. 3 presents the predicted versus the observed concentration values for the PM_{10} models for the year 2003. In the case of Helsinki, the PM_{10} values are much lower compared to those of Thessaloniki; this also contributes to overall lower values of RMSE in case of Helsinki. The forecasting models corresponding to the urban/traffic stations (Vallila and Agias Sofias) of the two cities indicate wider distribution of the predicted PM_{10} concentrations. The concentrations of certain episodic days for the case of Thessaloniki are not properly forecasted. These days correspond to high PM_{10} concentrations (>180 μg/m^3) in the Metropolitan Area of Thessaloniki, in April 2003, probably due to long range transport, as the specific pollutant demonstrates high values in other monitoring sites of the area. The data-driven model used in this case is not capable of properly forecasting long range transported episodes, as the latter require additional parameters, not available in the frame of our study.

4. Conclusions

On the basis of the requirements issued by the European directive (CAFE — 2008/50/EC), concerning the need for a common framework of methods and criteria that will allow for a direct comparison of the AQ in different member states, we have proposed a data-driven methodology consisting of CI methods in order to inter-compare and forecast AQ parameters. We have applied this method for the case of two European cities located in the northern and southern parts of Europe, i.e., Helsinki, Finland and Thessaloniki, Greece, respectively. The CI methods chosen for the aforementioned tasks were PCA and ANNs. Additionally, we have implemented a novel hybrid method for selecting the input variables of the ANN-MLP models. The latter one was based on the evaluation of actual performance of ANN-MLP models and contrary to some previous published studies, not on any empirical criteria, that could possibly not be valid, e.g., due to data quality issues.

The inter-comparison part of our paper was carried out by PCA, and resulted in the calculation of two significant PCs and their association with certain air emission source categories and air pollution processes at each studied area. The first PC for both Helsinki and Thessaloniki was identified to express data variations that indicated mainly the influence of local vehicular traffic, similarly for both cities, thus explaining approximately 25% of the overall data variability. On the other hand, the second PC indicated major differences between the two cities, in terms of air pollution mechanisms and emission source categories affecting the quality of breathed air. These mechanisms were mostly related with seasonal effects, such as the formation and depletion of ozone during the summer period in Thessaloniki, and the re-suspension of coarse particles especially in spring in Helsinki. With the aid of this second PC, half of the overall data variability could be satisfactorily explained in both cities. The rest of the data variability could not be addressed using this method, due to the complicated nature of the atmospheric phenomena and their non-linear characteristics. A further research step should therefore include tools capable of modelling more complex and strongly non-linear processes.

The forecasting of distinct PM mass fractions was carried out by ANN-MLP models. The input variables for the latter ones were selected with the aid of a novel hybrid scheme-method, involving LINREG and ANN-MLP models, aiming at the optimization of statistical indices of the model performance, such as the IA and the KI. This process resulted to optimized models indicating satisfactory performances. The IA ranged from 0.73 to 0.89, depending mainly on the quality and completeness of the datasets, while the KI for PM_{10} in both cities reached 60%, indicating an outstanding operational level of AQ forecasting. The latter result outperforms the corresponding results obtained by some previous studies for the same areas, even if those previous studies were based on more extensive data sets.

Although the mechanisms responsible for high PM concentrations are different for the two cities under consideration, the performance of the models was unexpectedly similar and can be considered to be good, based on the examination of various statistical model performance measures. This is probably attributed mainly to the fact that the hybrid scheme applied for the feature selection of the ANN model led to an improved performance. Moreover, the application of a non linear algorithm such as the ANN-MLP, and the multi-fold training and validation scheme adopted, improved the accuracy of the forecasting model, by covering a substantial part of the non-linear mechanisms and factors influencing air pollution. As the next step, it may be even more efficient to employ non-linear and self-trained methods for data investigation, that may improve the results received with PCA, whereas for the forecasting part it may be advantageous to investigate a hybrid, multi-algorithm approach, that learns and adapts to the data set of the AQ observations.

Acknowledgements

This study has been conducted in the frame of the scientific collaboration action COST ES0602 "Towards a European Network on Chemical Weather Forecasting and Information Systems" and the EU-funded project "Transport related Air Pollution and Health impacts — Integrated Methodologies for Assessing Particulate Matter" (TRANSPHORM).

Appendix A. Statistical indices

Model validation and performance is based on the following statistical measures. p_i refers to predicted values and a_i to actual (observed) ones, while with $\overline{p}$ and $\overline{a}$ are denoted the average of the predicted and observed data, respectively.

Correlation coefficient: is a dimensionless indicator ranging from −1 to 1, indicating linear correlation between the observed and predicted values. The correlation coefficient is calculated by the following equation

$$r = \frac{S_{PA}}{\sqrt{S_P S_A}}$$

where $S_{PA} = \frac{\sum_i (p_i - \overline{p})(a_i - \overline{a})}{n-1}$, $S_P = \frac{\sum_i (p_i - \overline{p})^2}{n-1}$ and $S_A = \frac{\sum_i (a_i - \overline{a})^2}{n-1}$.

The square of the correlation coefficient, also referred as coefficient of determination (R^2), is a statistical indicator usually used to maintain compatibility with other studies. It is limited to the range [0,1].

Root mean squared error (RMSE): is among the most commonly used indicators when evaluating the performance of ANNs. It is calculated by the following equation

$$\mathrm{RMSE} = \sqrt{\frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{n}}.$$

The RMSE preserves the units of the target variable; however extreme error values have more effect than small errors, due to the exponentiation.

Index of agreement (IA): is a dimensionless statistical indicator calculated by the following equation

$$\mathrm{IA} = 1 - \frac{\sum_i |p_i - a_i|^2}{\sum_i (|p_i - \overline{a}| + |a_i - \overline{a}|)^2}.$$

The IA expresses the difference between the predicted and observed values. It is limited to the range 0...1, with high values indicating good agreement between observations and predictions.

Kappa index (Cohen's kappa): is a statistical measure of inter-rater agreement for qualitative (categorical) items, calculated by the equation

$$\mathrm{k} = (\mathrm{P_o} - \mathrm{P_e}) / (1 - \mathrm{P_e})$$

where

$$\mathrm{P_o} = \mathrm{P_{FF}} + \mathrm{P_{TT}}$$

$$\mathrm{P_e} = (\mathrm{P_{FF}} + \mathrm{P_{TF}}) \cdot (\mathrm{P_{FF}} + \mathrm{P_{FT}}) + (\mathrm{P_{FT}} + \mathrm{P_{TT}}) \cdot (\mathrm{P_{TF}} + \mathrm{P_{TT}})$$

$\mathrm{P_{TT}}$ the probability of predicting an episode, while there is an episode.

$\mathrm{P_{FF}}$ the probability not predicting an episode, while there is no episode.

$\mathrm{P_{FT}}$ the probability of predicting an episode, while there is no episode.

$\mathrm{P_{TF}}$ the probability of not predicting an episode, while there is an episode.

Values of Cohen's kappa closer to 1 indicate good agreement, between predicted and observed episodes.

References

Assael MJ, Delaki M, Kakosimos KE. Applying the OSPM model to the calculation of PM_{10} concentration levels in the historical centre of the city of Thessaloniki. Atmos Environ 2008;42:65–77.

Athanasiadis I, Karatzas K, Mitkas P. Contemporary air quality forecasting methods: a comparative analysis between statistical methods and classification algorithms. In 5th int'l conference on urban air quality measurement, modelling and management, Valencia, Spain; 2005.

Barrero MA, Grimalt JO, Canto'n L. Prediction of daily ozone concentration maxima in the urban atmosphere. Chemom Intell Lab Syst 2006;80:67–76.

Bordignon S, Gaetan C, Lisi F. Nonlinear models for groundlevel ozone forecasting. Stat Meth Appl 2002;11:227–46.

Chavent M, Guιgan H, Kuentz V, Patouille B, Saracco J. PCA- and PMF-based methodology for air pollution sources identification and apportionment. Environmetrics 2008;20(8):928–42.

City of Helsinki. Achievements and challenges of sustainable development in Helsinki; 2007 www.hel.fi. last visited 9 August 2010.

Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20(1):37–46.

Efron B, Tibshirani R. An introduction to the bootstrap. New York: Chapman and Hall; 1993.

Eleuteri A, Tagliaferri R, Milano L. A novel information geometric approach to variable selection in MLP networks. Neural Netw 2005;18(10):1309–18.

Franklin SB, Gibson DJ, Robertson PA, Pohlmann JT, Fralish JS. Parallel analysis: a method for determining significant principal components. J Veg Sci 1995;6:99-106.

Gardner MW, Dorling SR. Neural network modelling and prediction of hourly NOx an NO2 concentrations in urban air in London. Atmos Environ 1999;31:709–19.

Ibarra-Berastegi G, Sáenz J, Ezcurra A, Ganzedo U, de Argandoña JD, Errasti I, et al. Assessing spatial variability of SO2 field as detected by an air quality network using self-organizing maps, cluster, and principal component analysis. Atmos Environ 2009;43(25):3829–36.

Jolliffe IT. Principal component analysis. 2nd ed. New York: Springer; 2002.

Kalapanidas E, Avouris N. Short-term air quality prediction using a case-based classifier. Environ Modell Softw 2001;16:263–72.

Karatzas K, Kaltsatos S. Air pollution modelling with the aid of computational intelligence methods in Thessaloniki, Greece. Simul Model Pract Theory 2007;15 (10):1310–9.

Karppinen A, Kukkonen J, Elolähde T, Konttinen M, Koskentalo T, Rantakrans E. A modelling system for predicting urban air pollution, model description and applications in the Helsinki metropolitan area. Atmos Environ 2000a;34(22): 3723–33.

Karppinen A, Kukkonen J, Elolähde T, Konttinen M, Koskentalo T. A modelling system for predicting urban air pollution, comparison of model predictions with the data of an urban measurement network. Atmos Environ 2000b;34(22):3735–43.

Kaskaoutisa DG, Kambezidisa HD, Nastos PT, Kosmopoulos PG. Study on an intense dust storm over Greece. Atmos Environ 2008;42(29):6884–96.

Kauhaniemi M, Karppinen A, Härkönen J, Kousa A, Alaviippola B, Koskentalo T, et al. Evaluation of a modelling system for predicting the concentrations of PM2.5 in an urban area. Atmos Environ 2008;42(19):4517–29.

Kohavi R, John GH. Wrappers for feature subset selection. Artif Intell 1997;97:273–324.

Kolehmainen M, Martikainen H, Hiltunen T, Ruuskanen J. Forecasting air quality parameters using hybrid neural network modeling. Environ Monit Assess 2000;65: 277–86.

Kukkonen J, Partanen L, Karppinen A, Ruuskanen J, Junninen H, Kolehmainen M, et al. Extensive evaluation of neural network models for the prediction of NO_2 and PM_{10} concentrations, compared with a deterministic modelling system and measurements in central Helsinki. Atmos Environ 2003;37(32):4539–50.

Kukkonen J, Pohjola M, Sokhi SR, Luhana L, Kitwiroon N, Fragkou L, et al. Analysis and evaluation of selected local-scale PM_{10} air pollution episodes in four European cities: Helsinki, London, Milan and Oslo. Atmos Environ 2005;39(15):2759–73.

Lazaridis M, Latos M, Aleksandropoulou V, Hov Ø, Papayannis A, Tørseth K. Contribution of forest fire emissions to atmospheric pollution in Greece. Air Qual Atmos Health 2008;1(3):143–58.

Lewis-Beck MS. Factor analysis and related techniques. London: Sage; 1994.

Manoli E, Voutsa D, Samara C. Chemical characterization and source identification/ apportionment of fine and coarse air particles in Thessaloniki, Greece. Atmos Environ 2002;36:949–61.

Moussiopoulos N, Nikolaou K. Environment and sustainability indicators for Thessaloniki, Organisation for the Master Plan and Environmental Protection of Thessaloniki978-960-98642-0-6; 2008.

Moussiopoulos N, Vlachokostas C, Tsilingiridis G, Douros I, Hourdakis E, Naneris C, et al. Air quality status in Greater Thessaloniki Area and the emission reductions needed for attaining the EU air quality legislation. Sci Total Environ 2009;407(4):1268–85.

Nagendra SM, Khare M. Artificial neural network approach for modelling nitrogen dioxide dispersion from vehicular exhaust emissions. Ecol Modell 2006;190: 99-115.

Niska H, Rantamäki M, Hiltunen T, Karppinen A, Kukkonen J, Ruuskanen J, et al. Evaluation of an integrated modelling system containing a multi-layer perceptron model and the numerical weather prediction model HIRLAM for the forecasting of urban airborne pollutant concentrations. Atmos Environ 2005;39:6524–36.

Niska H, Heikkinen M, Kolehmainen M. Genetic algorithms and sensivity analysis applied to select inputs of a multi-layer perceptron for the prediction of air pollutant time-series. Lect Notes Comput Sci 2006;4224:224–31.

Oanh NTK, Thiansathit W, Bond T, Subramanian R, Winijkul E. Compositional characterization of PM2.5 emitted from in-use diesel vehicles. Atmos Environ 2010;44(1):15–22.

Sillanpää M, Saarikoski S, Koskentalo T, Hillamo R, Kerminen V-M. PM10 monitoring and inter-comparison with the reference sampler in Helsinki, report 2002. 14+11 pp., available at Finnish Meteorological Institute and Helsinki Metropolitan Area Council; 2002 www.fmi.fi/kuvat/FINal_PM_Report.pdf.

Slini T, Kaprara A, Karatzas K, Moussiopoulos N. PM10 forecasting for Thessaloniki, Greece. Environ Modell Softw 2006;21:559–65.

Smeyers-Verbeke J, Den Hartog JC, Dehker WH, Coomans D, Buydens L, Massart DL. The use of principal components analysis for the investigation of an organic air pollutants data set. Atmos Environ 1984;18(11):2471–8.

Sofiev M, Vankevich R, Lotjonen M, Prank M, Petukhov V, Ermakova T, et al. An operational system for the assimilation of satellite information on wild-land fires for the needs of air quality modelling and forecasting. Atmos Chem Phys 2009;9: 6483–513.

Tzima F, Karatzas K, Mitkas P, Karathanasis S. Using data-mining techniques for PM10 forecasting in the metropolitan area of Thessaloniki, Greece. Proc of the 20th int joint conf on neural networks, Orlando; 2007. p. 2752–7.

Tzima F, Niska H, Kolehmainen M, Karatzas K, Mitkas P. An experimental evaluation of ZCS-DM for the prediction of urban air quality, in information technologies in environmental engineering. Proc of the 4th international ICSC symposium on information technologies in environmental engineering, Thessaloniki; 2009. p. 291–304.

Voutsa D, Samara C, Kouimtzis T, Ochsenkuhn K. Elemental composition of airborne particulate matter in the multi-impacted urban area of Thessaloniki, Greece. Atmos Environ 2002;36:4453–62.

Zickus M, Greig AJ, Niranjan M. Comparison of four machine learning methods for predicting PM10 concentration in Helsinki, Finland. Water Air Soil Pollut 2002;2: 717–29.