20th International Congress on Modelling and Simulation, Adelaide, Australia, 1–6 December 2013
www.mssanz.org.au/modsim2013

Air Quality Forecasting in Europe using Statistical Persistence

D.S. Zachary [a], **B. Chiera, J. Boland**[b]

[a]*CRP Henri Tudor 29, Avenue J.F. Kennedy L-1855 Luxembourg*
[b]*, University of South Australia, Adelaide, Australia*
Email: dan.zachary@tudor.lu

Abstract: We present an application of a single-indicator forecasting model to predict short-term urban air pollution levels in Europe. Prior knowledge of hourly, daily, and weekly levels of pollution can assist urban planners in policy and management procedures. Moreover, predictions using large-scale pollution forecasting systems, such as the Monitoring Atmospheric Composition & Climate program and the obsAIRve project, can be supported at the local scale by a statistical model using historical data from ground stations. An example of the use of historical data from ground stations is provided in the previous work of Chiera *et al.,* [2010] in which forecasts of the El Niño Southern Oscillation, represented by a two-state digital signal of the atmospheric pressure, were provided. Key to these forecasts was the concept of persistence, introduced to capture the observed behaviour of the digital signal remaining in one state for a prolonged time, before switching to the other state. This persistent behaviour was captured using Bayesian statistics to yield the *CFZG Model*, that is an adaptive Bayesian single-indicator forecasting model of a quasi-stochastic climate process.

In this paper we augment the methodology of Chiera *et al.,* [2010] to produce a multi-indicator model to predict pollution levels at measuring stations located in 36 European nations, based on observed persistent behaviour in air quality. An attractive feature of the adapted usage of the CFZG model is that it can be applied to multiple pollutant signals including all of the primary European pollutants such as Nitrous Oxides, Particulate Matter, Volatile Organic Compounds and Ozone. Unlike the single-indicator CFZG model, which used measuring stations from two locations only, we use measuring stations for pollutant signals which are geographically disparate, located in both rural and urban sites across 36 countries, all of which are registered with the European Environmental Agency. We present examples of typical nitrous oxide and ozone levels across selected sites and forecasting results for our chosen case study — rural Bosnia-Herzegovina — and compare the forecast against a control test that uses a random signal.

Keywords: *Air Quality, Statistics, Persistence, Modelling, Forecasting, Bayesian*

1 INTRODUCTION

The quantification of atmospheric pollution is an important societal concern. In past decades, European legislation has specified the need for member states to forecast and detail air composition, both in terms of gaseous and particles species. Two types of modelling forecast systems exist, both based on the use of meteorological data and chemistry models: statistical and deterministic. Historically, statistical models have played an important role although have generally been replaced since the 1980s by deterministic models. The latter have been developed using the well established European measuring networks, Menut and Bessagnet [2010], Balk *et al.,* [2011] and have been very successful for forecasting on the continental scale.

A major weakness present in both modern simulation and deterministic air quality models is the uncertainty linked to emissions. An important database for European emissions is the European Monitoring and Evaluation Programme (EMEP [2013]) which has, for example, a typical resolution of 50 km and therefore cannot provide details for accurate local pollution forecasting. Data is often available only by downscaling and then used as input, for example, to the chemistry transport (deterministic) models that require smooth emission data at hourly, daily, weekly, and seasonal levels. Moreover, pollution profiles developed by these models require the use of averaged meteorology and human activity changes (Menut and Bessagnet [2010]). A combination of statistical and deterministic models can however be advantageous, where on small scales, data may be limited or missing. Refined knowledge on these smaller scales can assist urban planners in policy and management procedures, while short-term air quality forecasting conditioned on anthropogenic hypotheses, such as emissions maps, can enhance the historical knowledge of the air pollution signal.

An important stream of statistical methodology is to the application of modelling quasi-stochastic signals in climate. The class of such signals is large and contains a number of domains ranging from water and soil modelling, through to the El Niño Southern Oscillation as well as air pollution. The underlying dynamics of many quasi-stochastic signals are still not fully understood although active work in this area continues (Liu [2010], Stewart [2010], De-Zheng [2007]) and approaches range from synthetic multi-model forecasting (Na *et al.,* [2011]) to capturing statistical trends (L'Heureux *et al.,* [2012]). A recently introduced novel statistical approach to capturing single-indicator quasi-stochastic processes was the *CFZG Model*, an adaptive Bayesian model, used to capture time series data of the El-Niño Southern Oscillation as a binary signal, yielding competitive forecasting with relatively little computational overhead (Chiera *et al.,* [2010]). The success of the CFZG model has opened up the prospect of other applications in the natural and social sciences where the underlying process is not understood or cannot be captured in a single framework.

A fundamental concept exploited in the CFZG model was that of *persistence*, that is, the degree of continuity in remaining in the current physical state before a phase transition is effected, due to the natural mechanisms of the physical process. The notion of persistence in quasi-stochastic signals can be generalised to a number of physical and environmental processes. Moreover, the CFZG model explores the full history of the signal and captures any delayed processes by learning from all available past information. The method can be used to detect natural driving forces in the signal, particularly in the presence of unexpected patterns and as such the CFZG model quantifies and combines events that are both near and far in time from a forecasting perspective.

In this paper we will utilise the general construction of the CFZG model to forecast short-term air quality levels and demonstrate an extended usage of the CFZG model to a European case study for which 36 nations have provided data to the European Environmental Agency (EEA) air quality database (EEA [2013]). The hundreds of stations represented in the EEA database provide a detailed view of urban and rural locations throughout Europe via air quality monitored data and information as well as multi-annual time series of measurement and meta-information for several pollutants. The applicability of the enhanced CFZG model will be to help fill in the gaps for local-scale forecasting, complementing large ensemble environmental models, such as the Monitoring Atmospheric Composition & Climate project MACC [2012] or in the obsAIRve project ObsAIRve [2013] using remote sensing information from the project GMES [2012].

This paper is outlined as follows. In Section 2 we present a generalised version of the CFZG model as well as the general air pollutant forecasting tool, using the data available from the EEA database. In Section 3 we describe the scope of air quality monitoring framework and the main details needed for this case study, as well as assimilation. In Section 4 we present the European case study and give our conclusions in Section 5.

2 THE CFZG MODEL

The CFZG model (Chiera *et al.,* [2010]) is an adaptive Bayesian model for forecasting a binary digitised time series data representing a physical process. In the original application, a series of $+1$ and -1 values were used

to correspond to positive and negative values in the physical process, with the baseline to determine the change between these positive and negative values, selected as 0.

Here we introduce the nomenclature *Above* and *Below* to represent meaningful events in the air pollution cycle. Specifically, *Above* refers to air pollution being above a suitably derived baseline for the pollutant of interest, while *Below* captures instances where air pollution is below this same baseline. Persistence is the scenario in which air pollution is consistently in an *Above* or *Below* phase. We introduce the use of three baselines for two distinct circumstances: (1) in the case of cyclic pollution, such as seasonal ozone, running averages and sinusoidal fits of the seasonal data; and (2) during 'regular' conditions, the use of a simple average.

We denote by $\mathcal{A}$ and $\mathcal{B}$ the *Above* and *Below* phases respectively. We can set aside a portion of the data to provide historical information and count the *Above* and *Below* events to compute probabilistic persistence of the signal. We begin with a counting scheme over the historical data for the number of *Above* episodes

$$\begin{aligned} \mathcal{A}_1 &= \text{Number of Above episodes of length 1} \\ \mathcal{A}_2 &= \text{Number of Above episodes of length 2} \\ \vdots &\quad \vdots \\ \mathcal{A}_{\mathbf{a}} &= \text{Number of Above episodes of length } \mathbf{a} \end{aligned}$$

where $\mathbf{a}$ is the length of the longest recorded $\mathcal{A}$-episode. We can similarly define $\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_{\mathbf{b}}$ for *Below* episodes where $\mathbf{b}$ is the length of the longest recorded $\mathcal{B}$ episode. We can then determine the total number of *Above* and *Below* episodes as

$$\mathcal{A}_T = \sum_{i=1}^{\mathbf{a}} \mathcal{A}_i \quad \text{and} \quad \mathcal{B}_T = \sum_{i=1}^{\mathbf{b}} \mathcal{B}_i$$

respectively, which naturally partitions the historical portion of the data series into a total of $\mathcal{A}_T + \mathcal{B}_T$ episodes. Using these counts we can compute the conditional probabilities

$$\alpha_j = P(\mathcal{A}_{\mathrm{j}+1} | \mathcal{A}_{\mathrm{j}}), \quad \mathrm{j} = 1, 2, \ldots, \mathbf{a} - 1$$

to yield the probability of observing a $(j+1)^{st}$ *Above* event given j consecutive *Above* events have already been observed. Similarly the conditional probability

$$\beta_j = P(\mathcal{B}_{\mathrm{j}+1} | \mathcal{B}_{\mathrm{j}}), \quad \mathrm{j} = 1, 2, \ldots, \mathbf{b} - 1$$

to give the probability of observing a $(j+1)^{st}$ *Below* event given j consecutive *Below* events have already been observed. Both of these conditional probabilities have natural estimators

$$\hat{\alpha}_j = \frac{\sum_{i=j+1}^{\mathbf{a}} \mathcal{A}_{\mathrm{i}}}{\sum_{i=j}^{\mathbf{a}} \mathcal{A}_{\mathrm{i}}}, \quad \text{and} \quad \hat{\beta}_j = \frac{\sum_{i=j+1}^{\mathbf{b}} \mathcal{B}_{\mathrm{i}}}{\sum_{i=j}^{\mathbf{b}} \mathcal{B}_{\mathrm{i}}}.$$

Due to the Bayesian approach of the CFZG method, a j^{th} positive event given $j-1$ consecutive positive previous events. This probability will almost certainly be different to the probability of observing a $j+1^{st}$ positive event directly following j consecutive positive events with the difference in these distributions exploited when forecasting. Figure 1 describes the future possible binary trajectories of the digitised signal based on these conditional probabilities. In this example a sequence of j consecutive 1s, $\mathcal{A}_j$, has been observed and the probabilistic trajectories of the data are shown in the tree, with the relevant probabilities on the edges of the tree resulting in the sequence shown at each node.

3 CASE STUDY: AIR POLLUTION ACROSS 36 EUROPEAN NATIONS

The data for this case study is comprised of air pollutant signals from several hundred air quality stations from 36 nations and is compiled in the European Environmental Agency (EEA) database where continuous measurements of several pollutant species are regularly tabulated. Pre-modelling analysis of the database

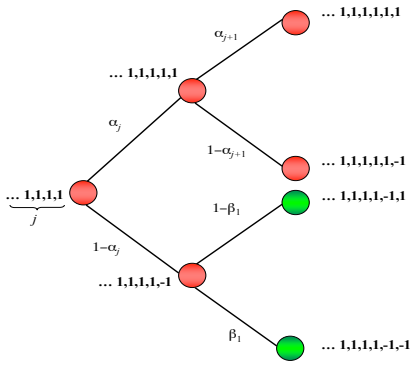


Figure 1: The probabilistic evolution when an $\mathcal{A}_j$ sequence has been observed Chiera *et al.,* [2010].

included the download of data, checking for data consistency and filtering stations that had insufficient and/or unreliable datasets. The subset of the stations that have completed initial pre-testing are shown in Figure 2 along with details of the measurement frequency. The station map (Figure 2) shows approximately 80% of the total number of stations with extensive coverage of Western Europe. Noteworthy exceptions where coverage is lacking includes France, Spain and Portugal.

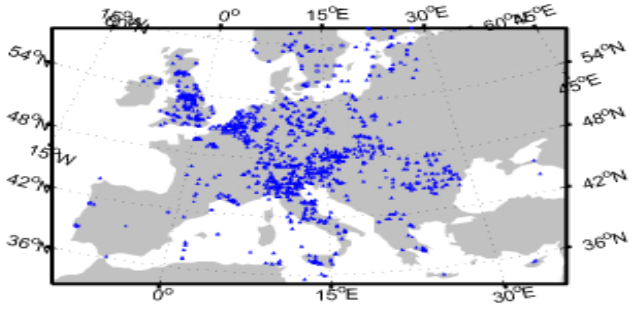


Figure 2: Stations that have currently been studied in the pre-analysis stage. Data has been taken at different frequencies: ■ (weekly), • (daily), ▲ (daily max), ▼ (hourly), ♦ (hourly - eight hour running average)

The EEA database pollutants include gases, hydrocarbons and suspended particles (Table 1) and most stations measure multi-pollutants. Measurements are further classified according to: (1) station type (Traffic, Industrial, Background); and (2) zone type (Urban, Suburban, Rural) which, for readability considerations, are not shown in Figure 2. Each station is designated with a precise location using latitude and longitude. The participant nations that contribute to the EEA database are given in Table 2.

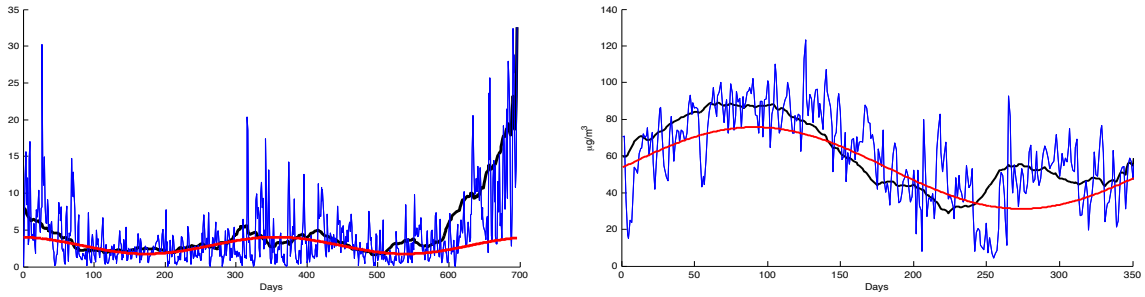
Table 1: EEA database pollutants.

Gases	NO, NO_2, O_3, PM10, SO_2, CO
Hydrocarbons	3-Trimethylbenzene, i-Pentane (2-methylbutane), C_6H_6, Toluene
Suspended particles	PM10, Cd (Cadmium), Lead, BaP (Benzo-pyrene), Black Smoke

Table 2: Participant nations

AU - Austria	ES - Estonia	IT - Italy	NO - Norway
BA - Bosnia & Herzegovina	FI - Finland	LI - Liechtenstein	PL - Poland
BE - Belgium	FR - France	LT - Lithuania	PT - Portugal
CH - Switzerland	GB - Great Britain	LU - Luxembourg	RO - Romania
CY - Cyprus	GR - Greece	LV - Latvia	RS - Serbia
CZ - Czech Republic	HR - Croatia	MK - Macedonia	SE - Sweden
DE - Germany	HU - Hungary	MT - Malta	SI - Slovenia
DK - Denmark	IE - Ireland	ME - Montenegro	SK - Slovakia
EE - Spain	IS - Iceland	NL - Netherlands	TR - Turkey

Figure 3 shows an example of a pre-analysis data check that also serves to indicate an appropriate baseline for the pollutant signals. In this case, a two and one year time series for NO_2 and O_3 are shown for the period 1 January 2006 to 21 December 2006 and 2007, respectively, at a background (rural) station in Bosnia-Herzegovina, about 50km west of Sarajevo and 10km north of the town Bradina and an urban measurement station in Sarajevo. The O_3 signal, and to a lesser extent, the NO_2 signal, show seasonal features and in this case an appropriate baseline would be either a running average or sinusoidal fit of the signal, from which *Above* and *Below* phases can be captured. An example of the mean for a 40-day running average for NO_2 and O_3 are given (black lines), along with a sine-curve fit (red lines).



(a) NO_2 time series (b) O_3 time series

Figure 3: An example of daily (a) NO_2 and (b) O_3 times series data for Bosnia-Herzegovina from 1 January 2006 to 31 December 2007 (blue lines). Days where measurements were not available were omitted. The red lines indicate sinusoidal baselines while 40-day running average baselines are indicated in black.

4 RESULTS AND DISCUSSION

An initial review of the European data has revealed that a large set of European station data is available, both in rural and urban settings, for different sectoral emissions (traffic, industry, residential areas), multiple pollutants (Table 1) and over different measurement frequencies, namely hourly, daily, and weekly (Figure 2). Approximately 80% of the total number of stations have been reviewed to date.

Systematic testing has allowed for a first look at the data set for Bosnia-Herzegovina and a comparison of rural and urban sites. We focused on a section of the original air quality signal that displayed the qualities of remaining in both the Above and Below states for a prolonged period of time (Figure 4 (a) and (b) for rural sites, and (c) and (d) for urban sites). In this preliminary test we considered conditional forecasting in which we assumed the current state is prolonged, as well as unconditional forecasting. For conditional forecasting, the current state of Above or Below was determined by the value of the air quality measurement relative to the baseline (purple dashed line $--$, Figure 4 (a) and (b)). For example, in Figure 4 (a), Day 5 commences above the baseline and a forecast conditioned on being in the Above phase is produced ($-\nabla$). On Day 14 the

signal is below the baseline and thus the model is conditioned on the Below state ($- \cdot \triangle$). We also consider the unconditioned forecast (⧫) which uses only the data, without specifically expecting to remain in the Above or Below state, instead allowing the data itself and Bayesian nature of the model to forecast air quality.

A visual comparison reveals that for the forecasts produced by the conditioned model, although the amplitude of air quality was over- and/or under-estimated, overall, the CFZG model was able to follow the trend of air quality for both NO_2 and O_3. The unconditioned forecast produced similar results. A test of forecast validity (not shown here due to space considerations) was conducted by comparing the model against a control signal, defined at each time step by $A \cdot x$, where x is a flat random value defined on (-1,1), of amplitude A, chosen to be equivalent to the data signal amplitude. The standard deviations were calculated for the difference between the data and model, and then for the model and control signal. For both the NO_2 and O_3 data for the rural and urban stations, the control signal gave standard deviations ranging from 10 - 20% larger than that produced by the model. The improvement of the model over the control signal indicates a promising use of the method for systematic and regular forecasting at stations. Finally, we note that these results are preliminary and are subject to further testing for more pollutant signals across the 36 European nations for a more complete comparison of urban and rural air quality signals. However, the initial results are encouraging for the future applicability of the CFZG model to air quality modelling.

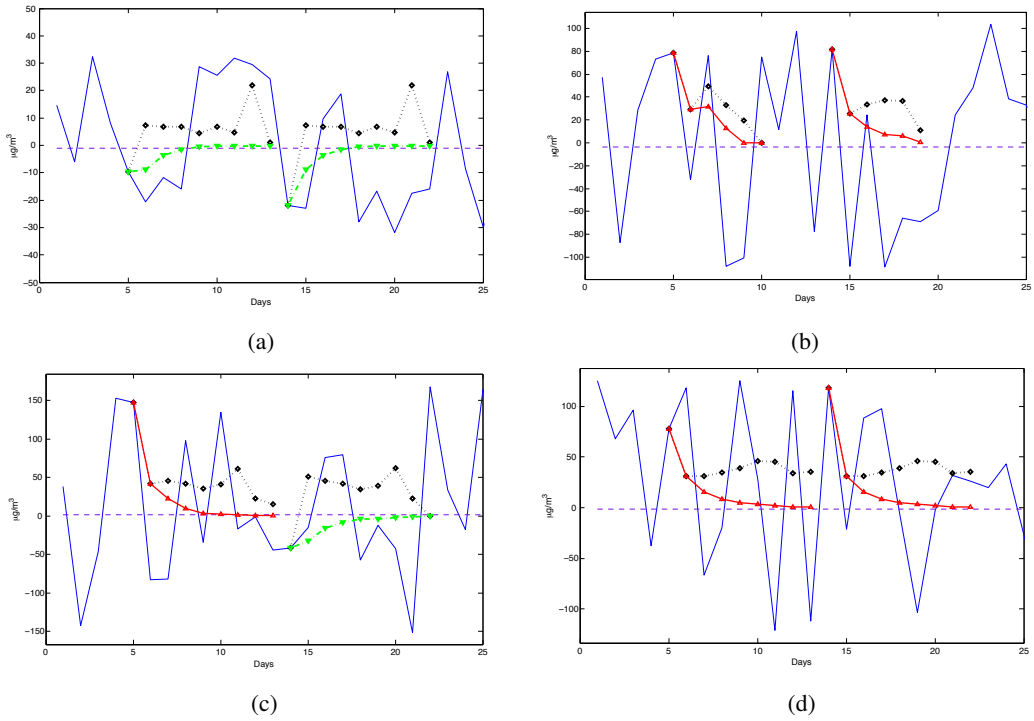


Figure 4: Forecasts for NO_2 and O_3 are given for a rural area ((a),(b)) and an urban area in Sarajevo ((c) and (d)). The original signal (–, blue) is compared with the forecast conditioned on Above (–∇ red) , Below ($- \cdot \triangle$ green) and the unconditioned model (black ⧫). The baseline also shown (purple, $--$).

5 CONCLUSIONS

The single-indicator CFZG forecast model has been extended to the multi-indicator problem of short-term air pollution forecasting of gases, hydrocarbons and suspended particles signals measured at a collection of urban and rural sites across 36 European nations. Appropriate baselines to indicate an air quality signal being *Above* or *Below* a meaningful measuring point included an average, a running average and sinusoidal signals to allow for seasonal behaviour of air pollution. This study focused on a rural station west of Sarajevo and initial validations indicated that both the conditioned and unconditioned models for Above and Below these baselines performed in a more meaningful fashion than when applied to the control signal, defined as a random signal with a similar amplitude as the data. These initial results are encouraging for further exploration.

Future work will consider a selection of case study areas that capture a salient distribution of rural and urban air quality information. The model will be further developed to include an integrated signal indicator for a better understanding of signal strength, as well as the integration of information from other relevant physical conditions such as geographical, orographical or meteorological, that could affect the signal, with an ultimate aim of a forecast map. It is intended that the results from this work will be used to mitigate air pollution risk by better anticipating days of poor air quality at specific locations.

ACKNOWLEDGEMENT

The authors would like to acknowledge support from the Ministry of Research and Higher Education (Luxembourg) for their financial support in contributing to this research.

REFERENCES

Bureau of Meteorology (The), Commonwealth of Australia: Elnino Statistics, `http://www.bom.gov.au/climate/glossary/elnino/elnino.html`, cited 2008.

Balk T., Kukkonen Jaakko., KaratzasK., Bassoukos T., Epitropou V. (2011), *A European open access chemical weather forecasting portal*, Atmospheric Environment 45, 6917-6922.

Chiera B., Filar J., Zachary D.S., Gordon A.H., (2010) *Comparative Forecasting and a test for persistence in the El Niño Southern Oscillation*, Decision Making under Uncertainty, J.A. Filar, A. Haurie, International Series in Operations Research and Management Science,Series Ed.: Hillier, Frederick S, 2010 (CFZG).

De-Zheng S. (2007), *Nonlinear Dynamics in Geosciences: The Role of El Niño Southern Oscillation in Regulating its Background State*, Springer. doi:10.1007/978-0-387-34918-3. ISBN 978-0-387-34917-6, Retrieved 2009-07-24.

EEA (2013), The European Environmental Agency, Data and maps, last accessed, June 2013, `http://www.eea.europa.eu`.

EMEP (2013), The European Monitoring and Evaluation Programme, last accessed, July 2013, `http://http://www.emep.int`.

GMES (2012), Global Monitoring for Environment and Security, the European contribution to the Global Earth Observation System of Systems (GEOSS) initiative, Retrieved on 2012-12-21, `http://www.fp7-space.eu`.

L'Heureux, M., Collins, D., & Hu, Z.-Z. (2012), *Linear trends in sea surface temperature of the tropical Pacific Ocean and implications for the El Niño-Southern Oscillation*, Climate Dynamics, 114. doi:10.1007/s00382-012-1331-2.

Liu T. (2005). *El Niño Watch from Space*, National Aeronautics and Space Administration. Retrieved 2010-05-31.

MACC (2013), Monitoring Atmospheric composition and climate, 3 day forecast for European cities, application of the GMES - Global Monitoring for Environment and Security, Retrieved 2012-12-21, `http://macc-raq.gmes-atmosphere.eu`.

Menut L. and Bessagnet B. (2010), *Atmospheric composition forecasting in Europe*, Ann. Geophy, 28, 61-74, 2010.

Na H., Jang B.G., Choi W-M., Kim K.-Y. (2011), *Statistical simulations of the future 50-year statistics of cold-tongue El Niño and warm-pool El Niño.*, Asia-Pacific J. Atmos. Sci., Vol 47, Num 3, 223-233.

ObsAIRve (2013), Europe-wide air quality monitoring and forecasting, bridging the gap between FP project GMES.

Stewart R. (2009), *El Niño and Tropical Heat - Our Ocean Planet: Oceanography in the 21st Century*, Department of Oceanography, Texas A&M University, Retrieved 2013-03-06.

Voukantsis D., Karatzas K., Kukkonen J., Rasanen T., Karppinen A., Kolehmainen M., (2011), *Intercomparison of air quality data using principal component analysis, and forecasting of PM10 and PM2.5 concentrations using artificial neural networks, in Thessaloniki and Helsinki*, Science of the Total Environment 409 (2011) 12661276.