MODELLING STUDY FOR CHARACTERIZING AND PREDICTING URBAN AIR POLLUTION

Gregorio Andria, Giuseppe Cavone, Anna M. L. Lanzolla, Alessandro Rubino

Department of Electrics and Electronics (DEE) – Politecnico di Bari
Viale del Turismo 8, 74100 Taranto

Keywords: environmental monitoring, missing data, mathematical models, Kalman filter, Kriging.

Abstract – This work proposes the development of an air pollution model based on a joined application of Kalman filter and Kriging technique. The use of modelling techniques in data environmental analysis allows to characterize the pollutants behaviour, in order to validate the measured data and to predict the values of contaminant substances emissions; so it results a very useful analysis tool, especially when there are numerous missing or erroneous data.
The joint applications of both Kalman filter and Kriging algorithms allows taking the main advantages of two different methods, in order to improve the performance of the developed model and to reduce its uncertainty too.

1. Introduction

The alarming growing of atmospheric pollution has led several countries in the world to establish severe lows and regulations defining the air quality and the required emission standard levels. In this context, lots of public administrations have installed several monitoring systems able to carry out, in real time, qualitative and quantitative information about the characteristics of urban centres environment. In particular, a considerable importance is turned to both atmospheric pollution and air quality evaluation.
As well known, the air quality depends on different factors, including the population density (residential, productive), the volumes of traffic, the energy demand and the physical characteristics of the territory (i.e. geographic conformation, orography) [1].
The availability of measured environmental data useful to the determination of the air pollution level, is often not enough to describe the temporal and spatial behaviour of all pollutant emissions; for this reason it is very important developing suitable mathematical model allowing to estimate eventual missing and/or non correct data, based on both the analysis of past trends and the correlations with available quantities.
In this context, the use of forecast algorithms assumes a remarkable importance. They must assure high accuracy, considerable versatility and facility in using. These characteristics are easily verifiable in Kalman filter [5] and in Kriging [6] algorithms that become useful tools to improve the treatment of the measured data obtained from monitoring environmental systems.

2. Proposed method

In our study, we propose the analysis of environmental data recoded in several monitoring stations installed in Taranto area (South Italy). This town is characterized by different sources that can be classified in two main classes: (i) sources producing daily averaged values of pollutants practically constant in the time, as the big industrial centre (with continuous production cycle) for the manufacture of the steel, for the refinement of the oil and for the production of the cement; (ii) sources producing emissions with concentration values varying during the day of the year as small industries, autoveicolare traffic, heating and cooling systems and so on. All the so identified sources contribute to reduce the air quality introducing smog and gas that spread in the urban centre and compromise the life quality of people.
The examined environmental network consists of 7 automatic acquisition and recoding stations that are able to measure both chemical substances and meteorological quantity. These stations are placed on various typologies of urban areas (parks, residential areas, sub-urban areas), according to prescriptions of the national laws.
Our main attention is addressed on stations characterised by high road traffic and located near busy traffic junctions. The most important pollutants analysed in our study are CO, benzene and toluene, since they represent the main contributors to air pollution caused to road traffic [9].
In previous works [3], [4], the authors identified and validated a mathematical model, based on interpolation techniques that permits to describe the time-varying behaviour of analysed substances and to highlight the multiple correlations between

them. The use of these techniques needs a continuous and numerous sets of values, otherwise it is necessary developing suitable reconstruction methods to estimate missing and/or incorrect data [2].
In this work the authors propose the application of Kalman filter to the analysis of environmental data in order to try overcoming the problem relevant to the eventual presence of not complete time series of measured values.
The Kalman filter is a recursive procedure that allows the data filtering and provides the estimate of the analysed quantities [5]. It allows the study of data in recursive way, carrying out the best estimate of required parameters, even if the characteristics of the observed phenomena are unknown. Thanks to recursive approach of the filter, it is possible to analyse a considerable set of values and extrapolate the information contained in every single data, reducing its noise contribution.
The Kalman filter is frequently used for its forecasting capabilities of the system variables in real time. In this case the model parameters are estimated from one initial set of measurements; subsequently, the filter model is applied, with the estimated parameters, to another set of data in order to calculate the forecasting performance of the same system variables.
By applying the Kalman filter to environmental analysis, we have characterized the time-behaviour of analysed substances by means of suitable mathematical models and we have identified some temporal relationships between different pollutants. In particular, the daily averaged concentrations of the benzene have been estimated by using the averaged concentration values of other considered pollutants as carbon monoxide (CO) and toluene that present similar characteristic with respect to benzene, by the following simple expression [3]:

$$\hat{BE}_i = \left(c_3 \cdot CO_i + c_2 \cdot TO_i + c_1\right) \cdot k\left(v_i\right) \tag{1}$$

where CO_i e TO_i represent the *i-th* daily averaged values of CO and toluene respectively, $k(v_i)$ is the corrective coefficient that takes into account of wind conditions defined in a previous article [3], and c_3, c_2, c_1 are the coefficients of the linear regression obtained by applying the Kalman filter to the measured data. As already sentenced, Kalman filter needs an initial estimate to get started. In our case we have imposed the null initial condition, so the coefficients have been calculated in iterative way by means of the following expression:

$$\begin{bmatrix} \hat{c}_{3i} \\ \hat{c}_{2i} \\ \hat{c}_{1i} \end{bmatrix} = \begin{bmatrix} \hat{c}_{3(i-1)} \\ \hat{c}_{2(i-1)} \\ \hat{c}_{1(i-1)} \end{bmatrix} + K_i \cdot \left\{ BE_i - \begin{bmatrix} CO_i & TO_i & 1 \end{bmatrix} \cdot \begin{bmatrix} \hat{c}_{3(i-1)} \\ \hat{c}_{2(i-1)} \\ \hat{c}_{1(i-1)} \end{bmatrix} \right\} \tag{2}$$

where $\hat{c}_{3i}$, $\hat{c}_{2i}$, $\hat{c}_{1i}$ and $\hat{c}_{3(i-1)}$, $\hat{c}_{2(i-1)}$, $\hat{c}_{1(i-1)}$ are the estimates of coefficients at *i-th* and *(i-1)-th* step respectively, K_i is the *Kalman gain* at *i-th* step, BE_i, CO_i e TO_i represent the *i-th* daily averaged values of benzene, CO and toluene respectively.
Thanks to the iterative structure of Kalman filter it is possible to improve the estimate of vector relevant to analysed quantities until the data are available or the difference between two subsequent estimated values is negligible enough.
By comparing the coefficient values obtained with the application of Kalman filter and with the use of enough multiple correlation it is possible to observe that the results are very similar for both methods. This confirms the effectiveness of developed model. Besides, the application of Kalman filter allows a small reduction of the estimate error from 1.73% to 1.60% and offers the typical advantages of recursive methods to estimate the state of a system from measurements that contain missing or erroneous data.
Figure 1 shows the behaviour of daily averaged values of benzene and its estimate obtained by applying Kalman filter.
The developed model examines environmental data recorded in only one monitoring station. To improve the model performances it is possible to take into account the values measured in the other stations of monitoring network and to identify the spatial relationships between the analysed substances. For this aim the Kriging technique has been applied. It is a geostatistical interpolation technique that considers both the distance and the degree of variation between the quantity values in known data points in order to estimate quantity values in unknown area [6]. This method uses a function called *variogram* to express the spatial variation; then, it minimizes the error of predicted values that are estimated by spatial distribution. The estimate is obtained by means of a weighted linear combination of the known sample values

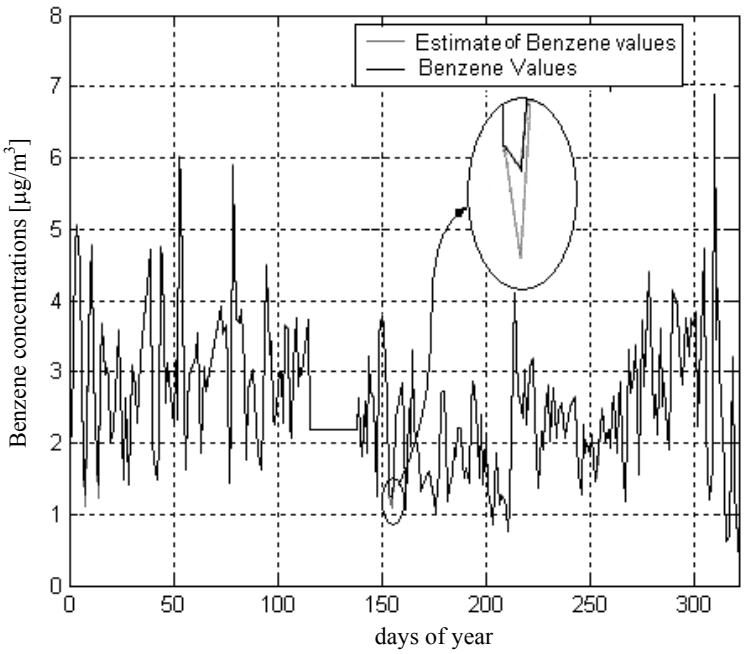

Fig. 1: daily average Benzene concentrations relevant year 2001 (dark line) and estimate values obtained by applying Kalman filter (grey line).

around the point to be estimated. Kriging algorithm possesses three major advantages with respect to other interpolation methods: (i) its interpolations are made by weights that do not depend upon data values, (ii) it provides an estimate of the interpolation error, (iii) it is an exact interpolation since the interpolation at any observation point is the observation itself.

The effectiveness of Kriging depends on both the correct specification of several parameters that describe the variogram and the stationary characteristics of analysed variables.

In order to characterize properly the analysed territory a classification relevant to pollutant spreading characteristics, has been carried out. Particularly, the monitoring stations have been separated in two different classes: *isotropy* (the pollution dispersion varies in the same way in all directions) and *anisotropy* (the pollution dispersion varies differently in different directions). Besides, the stations have been classified on the basis of different typologies of area and emissive sources near there. So, we have indicated as ***A*** the stations placed in park or residential areas, ***B*** the stations placed in very high road traffic areas, and ***C*** the stations placed near industrial centres.

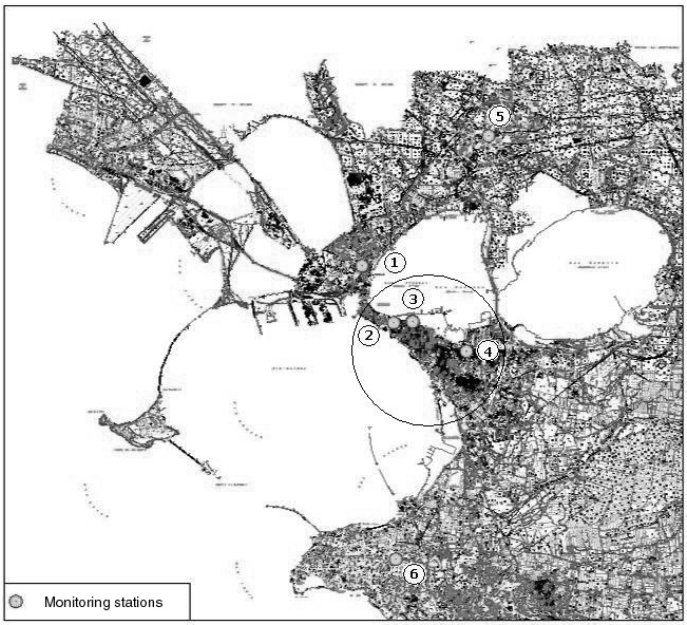

Fig. 2: Plant of Taranto area with the network of monitoring stations. The circle bounds the three stations analysed in Kriging technique.

By analysing the data relevant to the network of monitoring stations in Taranto, it is possible to observe that the concentrations of pollutants vary against several factors as the territorial characteristics, the typology of emissive sources and the meteorological conditions.

In our analysis, only three stations have been examined (two of them are of typology *C* and the other one is of typology *A*), that have the reciprocal distance less than 2 km. In this way, it is possible to assume that the local conditions are very similar in all analysed areas. The figure 2 shows the plant of Taranto where the three selected stations (indicated as 2, 3 and 4) are contained in the circled area. In particular, the stations 2 and 4 have similar characteristics because they are both located in roads with very high vehicular traffic. The station, 3 although placed in a monumental park, is affected by the pollution emissions caused by the near road with heavy traffic. These remarks led us to choose the three stations that bound a homogeneous area. Besides the three stations are not characterized by a preferential direction for pollution dispersion (caused by a different territorial conformation or by a strong prevalent wind in a particular direction, as compared to the other ones), so they have been considered as isotropy.

When the interest area is defined, the function describing in the best way the behaviour of pollutant concentrations in the spatial domain must be identified. For this aim, an experimental variogram, based on the correlation between the known samples acquired in the analysed stations, has been carried out. Particularly, the station number 3 has been considered as point of reference because it has a most central position in the analysed area. So the daily averaged values of benzene concentrations relevant to year 2001 have been represented as function of the distances of the stations 2 and 4 with respect to the station 3.

It is important to underline that the Kriging technique is based on the hypothesis of stationary data, but the behaviour of daily averaged values of benzene concentrations is a parabolic curve during the days of year. To overcome this limit, a division of year in twelve different classes (representing the twelve months of year) has been carried out. In this way, it is plausible to presume that the data are quite stationary in each class, so it is possible to calculate the experimental variogram defined as:

$$\gamma_k^*(h) = \frac{1}{2N_k} \cdot \sum_{j=1}^{N_k} \left(Z_{kj}(0) - Z_{kj}(h) \right)^2 \tag{3}$$

where N_k is the number of measurements contained in class k ($N_k \cong 30$), $Z_{kj}(0)$ *e* $Z_{kj}(h)$ represent the *j-th* values of daily averaged of benzene relevant to class k, acquired in the station pairs with reciprocal distance h.

By analysing the different experimental variogram defined for each class, we have tried to identify an analytical function describing in an effective way the relationship between the spatial points. The function that best fit the calculated experimental variogram is the exponential one, defined as follows:

$$\gamma_k(h) = c_k \left[1 - \exp\left(-\frac{3h}{a_k} \right) \right] \tag{4}$$

where c_k and a_k represent the *sill* (the value reached for great distances) and *range* (the maximum distance over which the variogram vary very slowly) for class *k,* respectively. It allows to estimate the benzene concentrations, in each class *k,* in every point at distance *h* from the station of reference.

Then, for each class, a weighed linear combination of the known sample values has been calculated, in order to estimate concentration data of substances acquired in a particular station under analysis. The model weights take into account spatial distances between the analysed stations.

The daily averaged values of benzene concentrations acquired in the station 2 have been estimated as function of the measurements relevant to the stations 3 and 4, by means of the following relationship:

$$\hat{BE}_{2i} = (\lambda_{1_k} BE_{3i} + \lambda_{2_k} BE_{4i}) \tag{5}$$

where BE_{3i} and BE_{4i} are the *i-th* daily averaged values of benzene (contained in class *k*) relevant to the station 3 and 4, respectively, and λ_{1_k}, λ_{2_k} represent the weights of the linear combination for benzene values included in class *k*. These weights have been calculated by solving the following system:

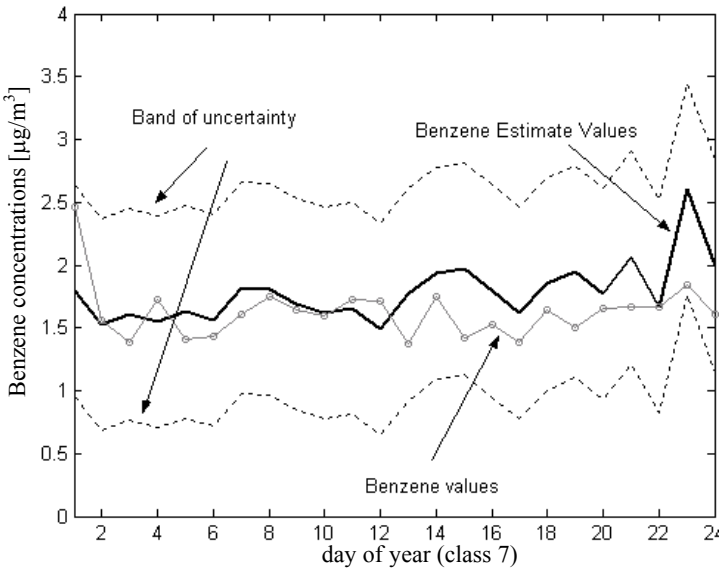

Fig 3: Behaviors of daily averaged values of benzene concentrations (continuous bold line) of its estimate obtained using Kriging model (line with circle marker), and of uncertainty bands (dashed lines).

$$\begin{bmatrix} \gamma_k(h_{33}) & \gamma_k(h_{34}) & 1 \\ \gamma_k(h_{43}) & \gamma_k(h_{44}) & 1 \\ 1 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} \lambda_{1_k} \\ \lambda_{2_k} \\ \mu_k \end{bmatrix} = \begin{bmatrix} \gamma_k(h_{32}) \\ \gamma_k(h_{42}) \\ 1 \end{bmatrix}; \qquad \sum_{i=1}^{n} \lambda_{1_k} = 1 \tag{6}$$

where the generic term $\gamma_k(h_{ij})$ represents the variogram relevant to the class *k* defined in eq. (4), calculated for the distance h_{ij} between the stations *i* and *j*. Besides, eq. (6) allows to calculate μ_k, that represents the Lagrange parameter used to calculate the minimized variance of Benzene estimate in the class *k* ($\sigma_{BE_{2k}}$) by means of the following expression:

$$\sigma_{BE_{2k}} = \sqrt{\lambda_{1_k}\gamma_k(h_{32}) + \lambda_{2_k}\gamma_k(h_{42}) + \mu_k} \tag{7}$$

To improve the estimate of benzene concentrations a corrective coefficient taking into account the different locations of the analysed stations has been introduced. It has been empirically calculated as the mean of ratio between the benzene concentrations and their estimates by means of eq. (5), for each class.

Figure 3 shows the results of benzene estimate by applying the Kriging method. By tacking a glance to this figure, it is possible to highlight the estimated values are always included within the uncertainty band.

The same technique has been applied to values of CO concentration, so similar results have been obtained too. This confirms the effectiveness of the developed model.

Finally, after checking the performances of the two analysed models, we have applied a hybrid model based on the joint application of both Kalman filter and Kriging techniques. In particular we have initially reconstructed the behaviour of daily averaged values of CO concentrations relevant to the monitoring station 2 by using the Kriging model applied to the data of stations 3 and 4. After, we have estimated the daily averaged values of benzene concentrations in the station 2 by means of the eq.(1). In this way, we have obtained an estimate error percentage

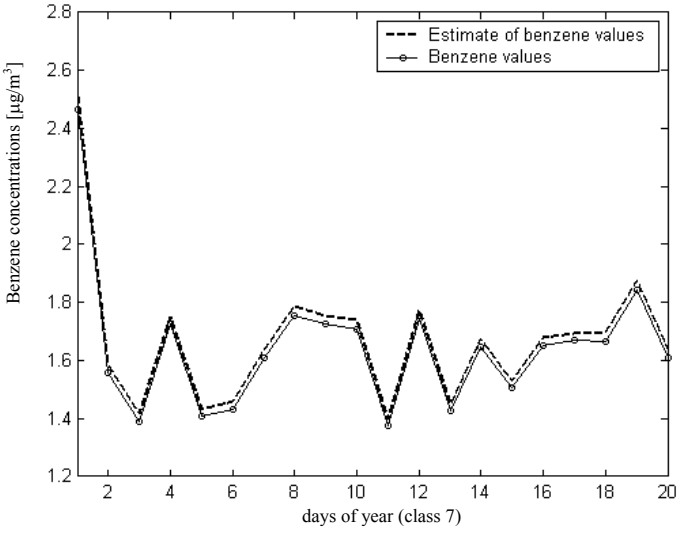

Fig. 4: Daily averaged of benzene concentrations relevant year 2001 (line with circle marker) and their estimate values obtained by applying the combined algorithms (dashed line).

equal about 3%.
Figure 4 shows the results of benzene estimate obtained by applying the so described hybrid model.
The same technique has been applied to values of Benzene concentration acquired in 2000 and 2002 years and similar results have been obtained. This confirms the effectiveness of the developed model.

3. Conclusions

The approach suggested in the paper allows the environmental engineers to analyse and to characterize concentration measurements of air pollutants, relevant to very high road-traffic areas. In particular, the joined use of the two described algorithms allows to carry out a good estimate of analysed substances and to reconstruct the eventual missing data in time and space domains. By using these techniques, it is possible to work with a continuous and valid set of data, with the further aim of reducing the measurement uncertainty.

References

[1] P. Zanetti: “Air pollution modelling, theories, computational methods and available software”, Van Nostrand Reinhold, New York 1990.

[2] S. Brandini, D. Bogni, S. Manzoni: “Knowledge based environmental data validation”, IEMSS 2002 Integrate Assessment and Decision Support, pp. 330-335, Lugano, June 2002.

[3] G. Andria, G Cavone, V, Di Lecce, A.M.L. Lanzolla: “Measurement and Characterization of Environmental Pollutants via data Correlation of Sensor outputs”, Proc. the IMTC/03 Instr. & Meas. Techn.Conf., Vail-CO, USA, Maggio 2003.

[4] G. Andria, G Cavone, V, Di Lecce, A.M.L. Lanzolla “Mathematic model for measurement and characterization of air pollution in areas with high road-traffic level,” Proc. of the XVII IMEKO World Congress, Dubrovnik, June 2003.

[5] H. W. Soreson: “Kalman Filtering: Theory and Application”, IEEE Press, New York, 1985.

[6] P. Brooker: “Kriging”, Engineering and Mining Journal, Vol. 180, No. 9,1979.

[7] S. Bendat, A.G. Piersol, “Analysis and measurements procedures”, J. Wiley & Sons, New York, 1971.

[8] J.R. Taylor, "An introduction to error analysis", University Science Books, Sausalito, CA, 1997

[9] A. Karppinen, J. Kukkonen, T.elolahde, T. Koskentalo: “A modelling system for predicting air pollution: comparison of model predictions with the data of an urban measurement network in Helsinki”, Atmospheric Environment 34-2000 pp. 3735-3743