IL NUOVO CIMENTO Vol. 26 C, N. 6 Novembre-Dicembre 2003
DOI 10.1393/ncc/i2004-10001-7

Non-linear atmospheric stability indices by neural-network modelling

A. Pasini(*), C. Perrino and A. Žujić(**)

CNR-Istituto sull'Inquinamento Atmosferico - via Salaria Km. 29,300
I-00016 Monterotondo Stazione (Roma), Italy

(ricevuto il 20 Gennaio 2004; revisionato l'1 Marzo 2004; approvato il 3 Marzo 2004)

Summary. — New atmospheric stability indices have been recently developed for the evaluation of primary pollution and the application results show their ability to grasp the physical features of the boundary layer. They are based on radon progeny measurements and multiple linear correlations with benzene. Here, neural networks are used in order to catch non-linearities in the boundary layer and to build non-linear indices. Their application to the modelling of benzene behaviour shows better prognostic results if compared with those coming from linear indices.

PACS 92.60.Fm – Boundary layer structure and processes.
PACS 07.05.Mh – Neural networks, fuzzy logic, artificial intelligence.

1. – Introduction

The relevance of radon progeny measurements in order to estimate (and possibly forecast) the diffusive properties of the atmospheric boundary layer has been recently established (see, for instance, [1,2] and references therein for diagnostic characterisations, and [3,4] for two attempts at forecasting).

In a previous paper [5], data from these detections were correlated (*via* multiple linear correlations) with diurnal and nocturnal average concentrations of benzene: the final result was the development of indices useful to describe the stability properties of the low atmospheric layers and to evaluate primary pollution episodes in an urban environment. In particular, these stability indices allow us to uncouple the role of the meteorological environment and of the emission rate in determining urban pollution events.

In that previous paper these indices showed their reliability as diagnostic tools in the boundary layer. Here, we change perspective, investigate the non-linearities hidden

(*) E-mail: pasini@iia.cnr.it
(**) On leave of absence from "Vinča" Institute, Belgrade, Jugoslavia

in the system and apply a fully non-linear method in order to obtain a more accurate characterisation of the boundary layer *via* radioactivity data.

2. – Experimental and previous results

Natural radioactivity and benzene detections have been performed at an urban background station located inside the park of Villa Ada, in the centre of Rome, during a 1-year monitoring activity. The station is placed inside a wide green area, some hundreds meters away from the nearest road, and it is not directly influenced by traffic emissions. In particular, radon progeny measurements come from the use of a "stability monitor" (SM200, OPSIS AB, Furulund, Sweden) that consists of a particulate matter sampler equipped with a Geiger-Müller detector for determining the total beta radioactivity of the short-lived radon decay products. This instrument, widely described elsewhere [5], allowed us to obtain twelve measurements per day, with a sampling time set to 2 hours. Benzene measurements, carried out by a continuous analyser, have been provided by the local monitoring network.

The atmospheric concentration of a whichever pollutant depends on the mixing properties of the boundary layer, the emission flux and the chemico-physical transformations. In the case of the so-called "primary pollutants", defined as pollutants which are found in the atmosphere in the same form in which they were emitted, the chemico-physical transformations can be neglected and an estimate of the mixing properties of the lower atmosphere can be a very useful tool to uncouple the role of the dilution factor and the emission factor in determining the air concentrations of these compounds. A good estimate of the mixing properties of the boundary layer was obtained [5] by developing the Atmospheric Stability Index (ASI), which has the purpose to characterise each day in terms of meteorological predisposition to the occurrence of a primary pollution event.

The ASI is made of two scalars, referring to morning and evening hours, respectively. Each scalar has been calculated on the basis of a multiple linear regression analysis of the values of natural radioactivity and of its time derivatives against benzene concentration at the urban background station. For the calculation of the ASI the year has been divided in three periods, according to different duration and intensity of the solar radiation (winter period: from October to February; intermediate period: March, April, August and September; summer period: from May to July).

The correlation between the ASI values and the concentration of primary pollutants is generally good, but it must be highlighted that the ASI takes into account only one of the two parameters which are the driving forces in determining primary pollutants concentration and that a one-to-one correlation could be possible only in the theoretical case of a constant emission flux. Conversely, the study of the differences between the modelled ASI value and the primary pollutants concentration is of help in identifying days when the atmospheric pollution is heavier (or lighter) than predictable on the basis of the atmospheric mixing, and thus to evaluate, for example, the real effect of traffic restrictions measures.

Also, the ASI can be a very reliable tool to compare the long-term trend of pollutants concentration: for example, to determine if a decrease in the air concentration of a given species from one year to another, or from one period to another, is due to a real improvement of the air quality or, instead, only to a lighter meteorological situation. Whichever variation in the concentration of pollutants due to modifications in the emission fluxes, in fact, can be easily masked by high variations which are simply due to meteorological parameters.

3. – Application of a neural-network model and new results

As stated above, the aim in using the ASI is not the very accurate reconstruction of benzene values in every situation, since this index takes only the dilution factor of the lower atmosphere into account and neglects the contribution of traffic emission variations to benzene concentration. Anyway, with a change of perspective, we could ask if the meteorological contribution to benzene behaviour is correctly modelled: for example, can the amount of variance explained by this linear model be increased? Here, we analyse the following heuristic hypothesis: the linear nature of the indices previously used is not able to fully catch the complex non-linear relationships among different variables in the boundary layer. In fact, boundary layer is a highly non-linear system, as clearly recognised by many studies (see, for instance, [6, 7] for general references).

In this conceptual framework, we take the same data set as in [5], analyse statistically the non-linearities hidden in these data and apply a fully non-linear neural-network model in order to obtain non-linear improved indices, which can represent good correlation laws even when tested on new data.

A preliminary statistical analysis has been performed in order to explore if linear and non-linear correlations among the variables of interest have the same magnitude: in this way we assess if a fully non-linear method can lead to differences (hopefully, improvements) in the description of the system. If we begin to use a neural-network jargon, we can call "inputs" the variables (2-hour radioactivity values and their time derivatives) used in order to model the behaviour of benzene concentration (which is our "target"). Now we can compare the relative importance of inputs by means of linear and non-linear bivariate analyses against the target. Here we use the standard linear correlation coefficient R and its non-linear generalisation R_{nl}, the so-called correlation ratio, whose square can be written as [8, 9]

$$R^2_{\mathrm{nl}} = \frac{\sum_i q_i \left(\bar{y}_i - \bar{y}\right)^2}{\sum_i \sum_\alpha \left(y_{i\alpha} - \bar{y}\right)^2}. \tag{1}$$

We can consider the target (benzene concentration, in our study) as the dependent variable and one input at a time as the independent variable. Then we group the values of the input into classes. Here R_{nl} is defined in terms of the average of the target for every specific i-th class of the chosen input: in fact, q_i is the sample size for the i-th class of the input, $\bar{y}_i = (1/q_i) \sum_\alpha^{q_i} y_{i\alpha}$ is the average of target for the i-th class of the input, $\bar{y} = (1/M) \sum_i \sum_\alpha^{q_i} y_{i\alpha}$ is the average of target on all the classes of the input and $M = \sum_i q_i$ is the total size of the set here considered.

We performed this analysis for the three periods of the year previously defined, both in diurnal and nocturnal cases. In general, we found differences between linear and non-linear correlation values for the same input-target sets; in particular, we discovered that inputs having low linear correlation with the target often assume high non-linear correlation values. In fig. 1 a comparison of R and R_{nl} values is shown for diurnal situations during the summer period (six 2-hour radioactivity values and five time derivatives are arbitrarily numbered).

Even if the correlation ratio does not measure all types of non-linearity, calculations of R_{nl} on our problem allow us to understand that some non-linearities are hidden in the relationships among the variables considered here. Furthermore, as a consequence of these bivariate analyses, the inputs to be considered for an optimal non-linear regression could be generally different from the variables chosen for an optimal linear regression.

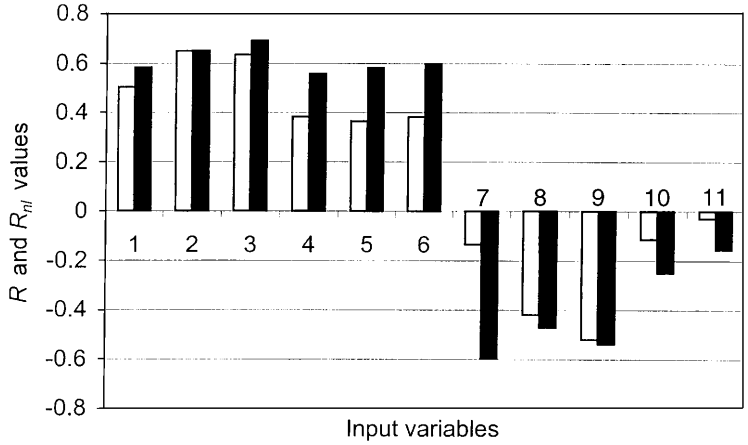

Fig. 1. – Example of a bivariate analysis of the inputs against the target in linear (R —white) and non-linear ($R_{\rm nl}$ —black) cases.

In order to perform a multivariate non-linear regression, we use a neural-network model which has already shown its good characteristic features in dealing with similar problems in the boundary layer (see, for instance, [10, 3, 9, 4]). The structure of this model has been described in [10] and, even more extensively, in [9]: in this second paper, however, a more complicated cost function was chosen, due to the particular problem to deal with. In the present paper we stress only that we use feed-forward networks and a backpropagation training endowed with gradient descent and momentum terms (see, for instance, [11] for an introduction to this kind of networks); a method of early stopping is also included, in order to avoid overfitting, and the cost function is a simple quadratic function, like in [10].

On the basis of the described method, we have now the goal of finding a reliable relationship which correctly links input and target variables even in cases when benzene concentration (our target) is *a priori* unknown. In order to achieve this goal, we trained the model on a set of coupled examples using detected values of radioactivity and benzene. Because of the quite few data available, we adopt little networks with 4 inputs, 5 neurons in a single hidden layer and one output (diurnal or nocturnal benzene concentration). Even if the values of the correlation ratio give us some first information about the relative importance of the inputs (2-hour radioactive data and their time derivatives) in a non-linear multivariate regression, many empirical proofs have been done in order to establish the best inputs for the 6 cases analysed here (diurnal and nocturnal situations for the three periods mentioned in the Experimental).

Each period is now divided in three sets: the first months in each diurnal and nocturnal interval represent the training set and include also the validation set (useful in order to establish the threshold for the early stopping of the neural model and chosen as 15 random days inside these first months), while the last month of each period is our test set. Of course, for the linear model we do not need a validation set and all the months but the last one represent the set on which the coefficients of the linear regression are calculated. In this way we properly test the generalisation performance of our linear and non-linear models, that is to say the ability of the relationships to correctly represent data coming from another sample (the test set) of the same population.

TABLE I. – *Values of the linear correlation coefficient (detected vs. modelled values) on the* 6 *test sets in the two cases of linear and non-linear models.*

Period	Linear model	Neural model
Winter - morning	0.786	0.809 ± 0.038
Winter - evening	0.740	0.812 ± 0.020
Summer - morning	0.558	0.576 ± 0.037
Summer - evening	0.639	0.725 ± 0.031
Intermediate - morning	0.664	0.877 ± 0.012
Intermediate - evening	0.698	0.795 ± 0.038

In order to evaluate the modelling performance by comparing model outputs and targets on the test sets, we used several indices: linear correlation coefficient R, mean square error, mean absolute value and variance. In table I results for the six indices (average benzene concentration values) are presented in terms of R (detected *vs.* modelled). Note that: 1) the error bars associated with the results by the neural model are derived from 60 different runs of the model (10 for each case) with different random initial weights, so that the model is able to widely explore the landscape of the cost function; they indicate ± 2 standard deviations; 2) the values of other indices of performance (here not shown) are in general agreement with the results presented in table I.

As one can see, the majority of improvements obtained by the application of the fully non-linear neural-network model is statistically significant. Furthermore, the calculation of the bias, that is to say the systematic error done by the two models, allows us to appreciate that, even in the few cases when the statistical significance of performance improvement is not sure, the values of the bias shown by the neural model are lower than the corresponding values derived by the application of the linear model, thus giving us more reliable results. An example of such a case is given in fig. 2, where the perfor-

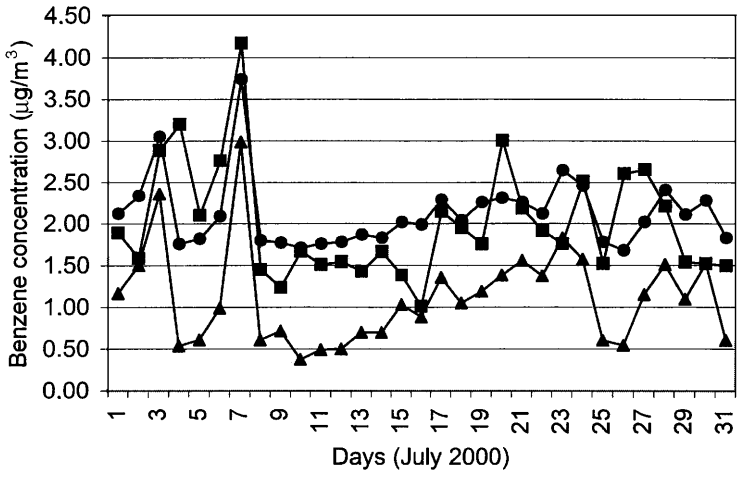

Fig. 2. – Performance of linear and neural models in describing benzene concentration during summer daytime periods (test set): detected data (square), linear-model estimation (triangle), neural-model estimation (circle).

mances of the two models are compared with respect to the detected values of benzene concentration on the test set of summer diurnal situations.

4. – Conclusions

The application of a neural-network model allows us to improve modelling performance of benzene behaviour starting from radioactivity detections. In our opinion this is mainly due to the application of a fully non-linear method in a complex environment like the boundary layer. These results are relevant for obtaining a better characterisation of the dilution properties of the lowest atmospheric layers.

* * *

One of the authors (AŽ) undertook this work with the support of the ICTP Programme for Training and Research in Italian Laboratories, Trieste, Italy.

REFERENCES

[1] Allegrini I., Febo A., Pasini A. and Schiarini S., *J. Geophys. Res.*, **99** (D9) (1994) 18765.
[2] Sesana L., Barbieri L., Facchini U. and Marcazzan G. M., *Radiat. Prot. Dosim.*, **78** (1998) 65.
[3] Pasini A., Ameli F. and Pelino V., *Nuovo Cimento C*, **24** (2001) 331.
[4] Pasini A. and Ameli F., *Geophys. Res. Lett.*, **30** (2003) 1386, doi:10.1029/2002GL016726.
[5] Perrino C., Pietrodangelo A. and Febo A., *Atmos. Environ.*, **35** (2001) 5235.
[6] Stull R. B., *An Introduction to Boundary Layer Meteorology* (Kluwer, Dordrecht) 1988.
[7] Mahrt L., *Theor. Comput. Fluid Dynamics*, **11** (1998) 263.
[8] Marzban C., Mitchell E. D. and Stumpf G. J., *Weather Forecast.*, **14** (1999) 1007.
[9] Pasini A., Pelino V. and Potestà S., *J. Geophys. Res.*, **106** (D14) (2001) 14951.
[10] Pasini A. and Potestà S., *Nuovo Cimento C*, **18** (1995) 505.
[11] Hertz J., Krogh A. and Palmer R. G., *Introduction to the Theory of Neural Computation* (Addison-Wesley, New York) 1991.