ELSEVIER

Ecological Modelling 120 (1999) 157–165

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecomodel

Neural network models to study relationships between lead concentration in grasses and permanent urban descriptors in Athens city (Greece)

Ioannis Dimopoulos [a,*], J. Chronopoulos [b], A. Chronopoulou-Sereli [a], Sovan Lek [c]

[a] *Laboratory of Physics and Agricultural Meteorology, Department of General Sciences, Agricultural University of Athens, Iera Odos 75, 118 55 Athens, Greece*

[b] *Laboratory of Floricultural and Landscape Architecture, Department of General Sciences, Agricultural University of Athens, Iera Odos 75, 118 55 Athens, Greece*

[c] *CESAC UMR 5576, CNRS-Univ. Paul Sabatier, 118 route de Narbonne, 31062 Toulouse cedex, France*

Abstract

The aim of the present work is to propose a model for the estimation of lead concentration in grasses using urban descriptors easily accessible and to study the specific effect of each descriptor on lead concentration. Six descriptors were considered: the density of vegetation, the vegetation height, wind velocity, height of building, distance of adjacent street, traffic volume. Lead concentrations were determined in one grass species, *Cynodon dactylon* (L.) Pers, (Bermuda grass), collected from 30 different locations in Athens city. The proposed model is a multilayer perceptron (MLP) trained by backpropagation. The predictive quality of the model was judged by two cross-validation methods. The generalization ability of the model is confirmed by a determination coefficient higher than 0.91. The study of the first partial derivatives of the output of the MLP with respect to each input is used to identify of the factors influencing the lead concentration and the mode of action of each factor. Results allow to classify the environmental descriptors by their decreasing influence on lead concentration: distance of adjacent street, traffic volume, density of vegetation, wind velocity, height of building and vegetation height.

Keywords: Urban pollution; Heavy metal; Modelling; Backpropagation; Multiple regression; Sensitivity analysis

1. Introduction

In the city of Athens the constant increase of the population over the last decades has resulted in high traffic volumes and consequently high automobile emissions. Compounded by the narrowness of the roads this has caused discomfort (due to environmental conditions) to the city residents. Consequently, the air, plants and the soil are contaminated by various contaminants such as lead (Pb) (Ndiokwere, 1984; Ho and Tai, 1988; Mielke, 1991; Francek, 1992).

* Corresponding author. Fax: +30-152-94233.
E-mail address: gphy2hrk@auadec.aua.gr (I. Dimopoulos)

0304-3800/99/$ - see front matter
PII: S0304-3800(99)00099-X

In a city environment the main sources of Pb pollution are car exhausts, fumes and tyre wear, if there are no smelting sites, heavy industry or other sources of Pb contamination nearby (Akhter and Madany, 1993). In addition to the automobile emissions, the high density of large buildings amplifies pollution of plants because dispersion of the pollutants over wider areas is prevented (Capannesi et al., 1988).

Regarding the dispersion of pollutants in parks, studies suggest that the pollution burden is greater in the peripheral than in the central zones of the open areas (Shao-Lian et al., 1989; Grodzinska et al., 1990). In a previous study, the authors (Chronopoulos et al., 1997) examined the impact of traffic conditions on the vegetation and soil of two major parks in Athens and concluded that the density and composition of the peripheral vegetation has a remarkable effect on the dispersion of Pb and Cd towards the inner sites of the parks.

The concentration of pollutants in the different parts of the plants is strongly dependent on the plant species. Plant species as well as the design patterns of parks can also affect the distribution of Pb concentration in plants. A limited number of plant species that tolerate and colonize environments polluted with heavy metals are selected and used in the composition of city parks and avenue median dividers. Several plant species were studied to evaluate Pb contamination in city environments. *Cynodon dactylon* (L.) Pers. is one of the most frequently studied plant species for this purpose (Ho and Tai, 1988; Sukkop, 1990).

In order to establish realistic simulation models of Pb deposition and accumulation by plant species several inter-dependent models of environmental processes have to be linked together. Direct measurements of deposition rates using micrometeorological methods have advanced the knowledge of deposition processes. However, routine implementation of these methods for monitoring deposition rates is difficult and pollutant dispersion models for urban and industrial regions are only just beginning to be developed.

The purpose of our study is the evaluation of Pb levels in vegetation in an urban environment, using environmental parameters that are easily accessible and that strongly influence the diffusion of the pollutants which are mainly the result of high traffic (Preer, 1977; Wong, 1996). At the present study, we use and compare the predictive capacity of two statistical methods: Multiple Linear Regression (MLR) and Neural Networks (NN). Model-predicted and observed values are compared by different statistical parameters. For the NN model we propose a new simple method to study the relationship between the Pb concentrations estimated by the model and each influencing variable.

2. Materials and methods

2.1. Study area and environmental descriptors

The city centre of Athens is characterized by the presence of high densities of tall buildings and very infrequent sites covered by vegetation, such as parks. National Garden and Areos Park are the two major parks in the city centre they occupy relatively large areas of 15.8 and 24.0 hectares, respectively. These two parks are surrounded by avenues and streets, with different traffic volumes and an orientation that inhibits air circulation and dispersion of pollutants.

Squares of considerable size, covered with vegetation and able to provide comfortable environmental conditions for the citizens, are almost absent from the city of Athens. The great majority of city squares (approximately 92%) are less than 1.0 ha in size.

Samples were collected during the summer of 1995, from the plant species *Cynodon dactylon*, at 30 different locations (three public squares 1.0 ha in size, three public squares 5.0 ha in size, three public squares 10 ha in size, two parks and 14 traffic islands, Fig. 1). At each site three samples of *Cynodon* were bulked together to give a composite sample of about 5 g. A total of 140 plant samples were studied. *Cynodon dactylon* was selected for monitoring Pb contamination since it was found at all studied parks, squares and traffic islands. All plant samples were oven-dried at 70–80°C and ground to a fine powder by a microhammer mill to pass through a 1 mm mesh screen.

From each powder sample three subsamples of 1 g were weighed and metals were extracted by digestion with a 2:1 $HClO_4/HNO_3$ solution. Then the samples were filtered and diluted with deionised water to the final volume for Pb determination. Lead concentration was determined in the extracted solutions by atomic absorption spectrometry (GBC 908 FBT). The detection limit was 100 ppb for Pb with an accuracy of 1% RSD.

Every sample was described by a set of permanent descriptors (discrete and continuous).

- DENS: mean density of vegetation between the sample point and the nearest adjacent street (the values of DENS varied over the range 0–90%).
- GRAD: mean vegetation height between the sample point and the nearest adjacent street (the value of GRAD varied over the range 0–2 m).
- AIR: Wind velocity recordings were carried out with a digital measurement device at a network of 140 selected points. The measurement points were located at the plant sampling sites. The measurements were made at a height of 2.0 m above ground using a cap anemometer. A total of 38 measurement trips were conducted. After processing the data obtained, the average wind velocity was determined for the selected points. The reduction of the wind velocity for each measurement point compared with the maximum mean wind velocity was determined. Then a variable, AIR, was introduced to take into account the reduction of the wind velocity and the degree of ventilation at the measurement points. When reduction did not exceed 20%, ventilation was considered good (AIR = 3). At the points where the reduction varied between 20 and 40% ventilation was considered moderate (AIR = 2) and whenever the reduction exceeded 40% ventilation was considered poor (AIR = 1).
- BUILD: mean height of the adjacent buildings (the value of BUILD varied over the range 2–8 floors).
- DIST: distance between the sample point and the nearest adjacent street (the value of DIST varied from 0–66 m)
- TRAF: Traffic volume as expressed from the number of traffic lanes (the value of TRAF varied from 2–8 lines).

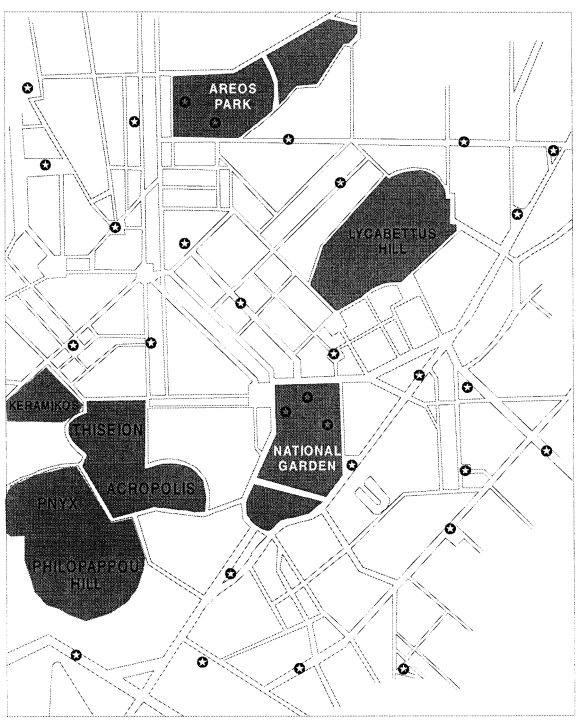

Fig. 1. Locations of measurement points in Athens city centre.

2.2. Modelling techniques

The techniques of multiple linear regression and stepwise multiple linear regression (Weisberg, 1980; Tomassone et al., 1983) were used. Calculations were done using SPSS software.

Multilayer Perceptrons (MLP), the most commonly used artificial neural networks, are general purpose, flexible, nonlinear models, $f: R^n \rightarrow R^m$, of the general form:

$$f(x) = \phi_n[W_n \phi_{n-1}[W_{n-1}\phi_{n-2}[\ldots \phi_1[W_1 x]]]] \quad (1)$$

$$f^j(x) = \phi_i^L\left[\sum_{j=0}^{J_{L-1}} w_{ji}^L \phi_j^{L-1}\left[\sum_{u=0}^{J_{L-2}} w_{uj}^{L-1}\phi_u^{L-2}\left[\ldots \phi_l^1\left[\sum_{e=0}^{J_0} w_{el}^1 x_e\right]\right]\right]\right], \; j = 1,\ldots,m \quad (2)$$

where W_i stands for the parameter matrix or weight matrix and ϕ_i stands for diagonal nonlinear operators; the elements of which are the so-called activation functions. MLP's with a nonlinear activation function are genuinely nonlinear and it has been proved (Cybenko, 1989) that, under some weak assumptions, any function can be approximated with an arbitrary accuracy by an MLP. Estimation of W is called training, learning or adaptation of the weights and regression via MLP is called supervised learning. The backpropagation algorithm is the most frequently used for training (Rumelhart et al., 1986).

A major problem in the use of MLP for model building is the determination of the optimal architecture of the network (number L of layers and J_j, $j=1\ldots L$, where J_j is the number of node for layer j). Usually, the *trial-and-error* method is applied to test various alternative model architectures and choose the one with the optimal generalisation capability. Generalisation is defined by the ability of a model to predict data other than those on which it has been trained. A model with too many free parameters will fit the training data arbitrarily closely, but will not necessarily lead to optimal generalisation (overfitting).

Two classes of generalisation criteria are usually used for model architecture selection and model testing. The first class contains criteria based on the fitting errors (e.g. Akaike information criterion, Akaike, 1974). The second class of criteria is based on the principle of cross-validation (CV), according to which, the decisions on the model structure and predictive capacity are made on samples of data different than the sample used to estimate the parameters of the model. Usually overfitting is controlled by using a subset of the data, the validation set. This subset is not used for the computation of the weight matrix but for stopping the training process and taking decisions on the architecture parameters. The generalization ability is estimated by using another subset of the data, the test set, which neither participated in the weight estimation, nor in the architecture optimization, but only for the ultimate evaluation of the model. Separation of the data into the subsets is not straight-forward. Several questions arise concerning this method, they are discussed in Weigend et al. (1992).

One of the most efficient methods is k-fold cross-validation. The data set is divided into k approximately equal parts, and each part is used in turn as the test set for the network trained on the remainder, and the observed error rates on the k parts are averaged.

The error of a network, as a function of the weights that define it, is filled with hills and valleys. A trivial change in the training data can change the weights. Even with exactly the same training set, different random starting weights can result in dramatically different final results. Therefore, we do not dare assert that a network trained with all of the known data is essentially identical to networks trained with subsets of the data. To take into account this problem Moody and Utans (1991) propose a modification of the above cross-validation method: nonlinear k-fold cross-validation (NL K–f CV). In this work we use the two alternative kinds of CV: (1) CV with training, validation and test data sets and (2) NL K–f CV.

2.3. Preparation of data

The input data had very different orders of magnitude according to the variables. To standardize the scales of measurement, the values of the variables were converted by the relationship:

$$Z_s = \frac{X_o - \bar{X}}{\sigma_x} \tag{3}$$

with Z_s: standardized values, X_o: original values, $\bar{X}$ and σ_x the mean and standard deviation of the variable. The dependant variable Pb was also centred, reduced and converted over the interval [0…1] because the logistic function used for the NN output neuron modulates the response to values between 0 and 1.

2.4. Study of the influencing factors

In multiple linear regression, the influence of each variable can be roughly assessed by checking the final values of the regression coefficients. In mathematical terms, each coefficient of a linear model is the partial derivative of the response of the model with respect to the variable of that coefficient. The MLR partial coefficients therefore

generally give an indication of environmental reality, although it is not possible for this type of model to represent a nonlinear relationship such as that which probably exists between Pb levels and some influencing factors. On the other hand the neural network is a 'black box' type model and does not clarify the participation of each of the explanatory variables (descriptors). In this study we use a simple method based on the use of the partial derivatives of the network

response with respect to each descriptor. The link between the modification of inputs, x_j, and the variation of outputs, $y_j = f(x_j)$, is the Jacobian matrix $\mathrm{d}y/\mathrm{d}x^t = [\partial y/\partial x]_{\mathrm{mxn}}$. It represents the sensitivity of the network outputs according to small input perturbations. For a network with n inputs, one hidden layer with ni nodes, and one output (i.e. $m = 1$), the gradient vector of y_j with respect to x_j is $d_j = [d_{j1},\ldots,d_{je},\ldots,d_{jn}]^T$ (Dimopoulos et al., 1995), with:

$$d_{je} = s_j \sum_{i=1}^{ni} w_{is} I_{ij} (1 - I_{ij}) w_{ei} \quad (4)$$

(under the assumption that a logistic sigmoid function is used for the activation. When s_j is the derivative of the output node with respect to its input, I_{ij} is the output of the ith hidden node for the input x_j, the scalars w_{is} and w_{ei} are the weights between the output node and the ith hidden node, and between the eth input node and the ith hidden node).

The sensitivity of the *MLP* output for the data set with respect to input x_e is:

$$\mathrm{SSD}_e = \sum_{i=1}^{ni} (d_{je})^2 \quad (5)$$

and the derivative can be efficiently computed as a minor extension to the backpropagation algorithm used for training.

3. Results and discussion

3.1. Performance of the models

3.1.1. Multiple linear regression modelling

3.1.1.1. Complete model. With all the eight variables, the equation of the MLR model and determination coefficient became:

Pb =	−0.374DENS	+0.156GRAD	−0.033AIR	+0.097BUILD	−0.724DIST	+0.092TRAF
(*t*	−3.728	1.739	−0.617	1.646	−12.125	1.247)
(*Sig.*	0.000	0.024	0.538	0.102	0.000	0.214)

$$R^2 = 0.703 \quad (6)$$

3.1.1.2. Stepwise model. Only three independent variables were retained by the model:

Pb =	−0.228DENS	+0.139 BUILD	−0.7 DIST
(*t*	−4.155	2.914	−12.650)
(Sig.	0.000	0.004	0.000)

$$R^2 = 0.696 \quad (7)$$

The study of Fig. 2 shows several problems of the MLR model (Eq. (7)): an underestimation of the low values (Fig. 2a), the residuals (differences between observed and estimated values) tend to increase with estimated values (Fig. 2b). The residual distribution is far from normality (Fig. 2c).

3.1.2. Neural network

With the cross-validation approach, a good predictive model can be obtained using a network with three neurons in the hidden layer and sigmoid as activation function. In Table 1, the performance of the MLP model estimated by two CV methods is shown (MSE = Mean square error). The high value of the determination coefficient

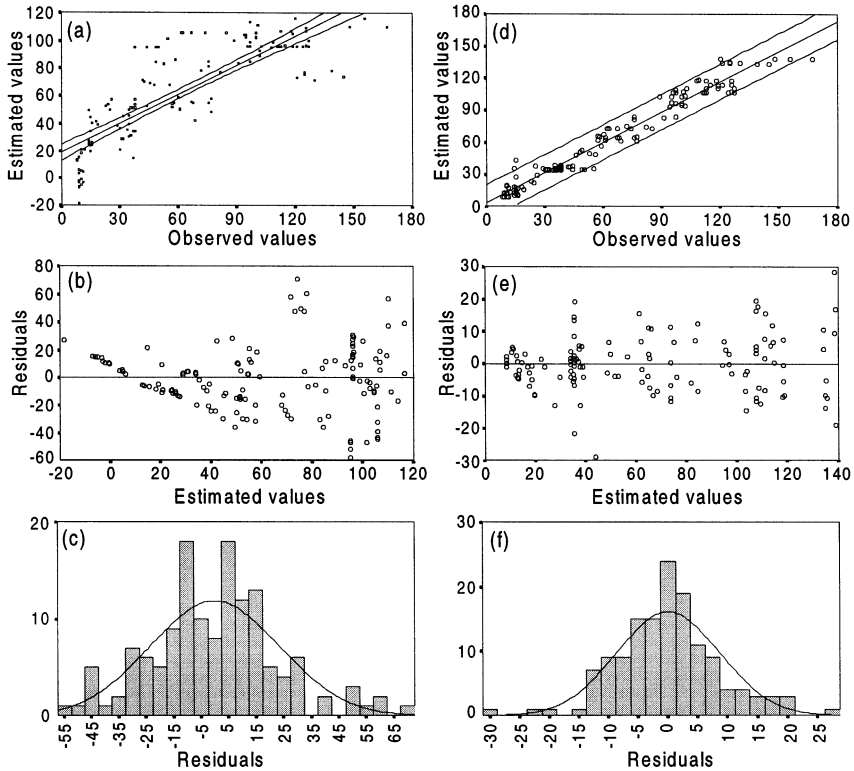

Fig. 2. Relationship between observed and estimated values of Pb; (a) MLR and (d). MLP Relationship between the residuals and the estimated values of Pb; (b) MLR and (e) MLP. Distribution of residuals (observed values-estimated values of Pb); (c) MLR and (f) MLP.

demonstrates the predictive capacity of the model (R^2 higher than 0.9). The fact that MLP provide a good predictive model was highlighted by the independence of the residuals from the variable to be predicted (Fig. 2e) and their normality (Fig. 2f). The distribution of residuals is better balanced with MLP than with MLR. Values that exceed the limits of the normal approximation are rather scarce.

Table 1
Mean square error (MSE) and determination coefficient R^2 for the NN model estimated by two alternative kinds of CV method

		MSE	R^2
CV	Training (80)	54.024	0.953
	Validation (30)	79.247	0.956
	Test (30)	107.779	0.938
NL 10–*f* CV	Training	50.37	0.972
	Test	88.547	0.911

3.2. Influence of factors

The study of MLR model (7) leads to the conclusion that the most significant factors affecting Pb diffusion are in decreasing order significance DIST, DENS and BUILD. Pb concentration decreased with DIST and DENS and increased with BUILD. The rest of the factors are either not very important or they are correlated to the three more significant factors.

The study of the MLP model, according to the method presented in Section 2.4, led to the layout of Fig. 3. Thus, for instance every point of DDENS versus DENS (Fig. 3a) resulted from Eq. (4) with $j = 1,\ldots,140$. Eq. (5) allows the variables to be classified according to their increasing influ-

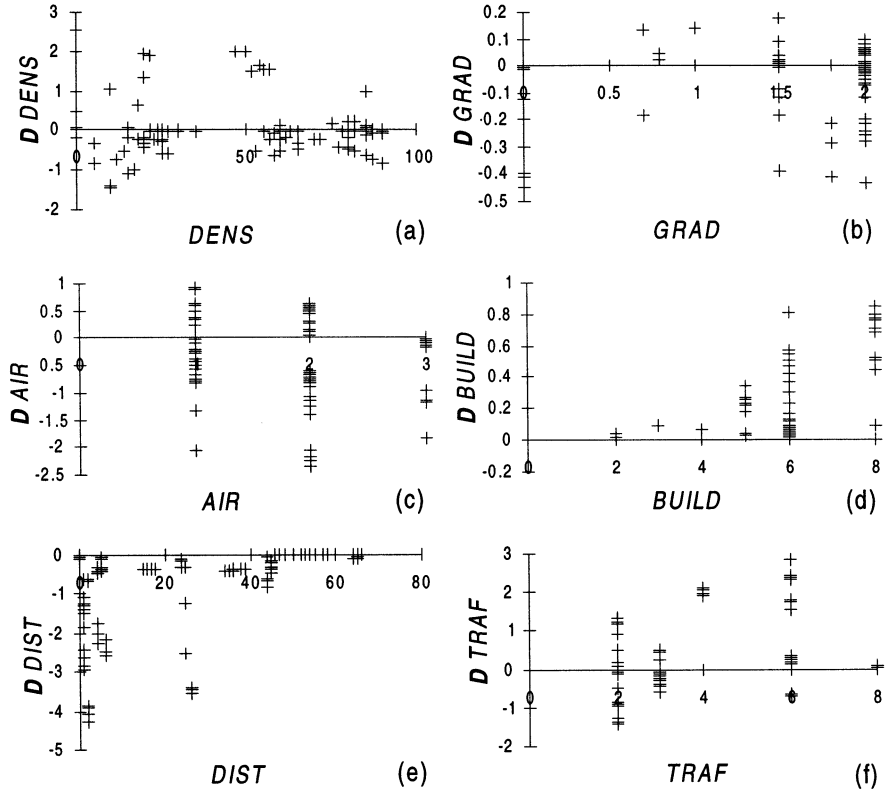

Fig. 3. Partial derivatives of the *NN* model reponse with respect to each descriptor.

ence on Pb concentration: DIST ($SSD_{DIST} = 339.95$), TRAF ($SSD_{TRAF} = 137.83$), DENS ($SSD_{DENS} = 100.29$), AIR($SSD_{AIR} = 97.78$), BUILD ($SSD_{BUILD} = 12.99$), GRAD ($SSD_{GRAD} = 2.99$). The study of Fig. 3 leads to the following remarks:

- The influence of density (DENS) on the Pb concentration is rather complicated and non-linear (Fig. 3a). The negative values of partial derivatives (DDENS) for the majority of the values of DENS show that the increase of the density contributes to the reduction of Pb concentrations.
- The influence of the height of vegetation on the reduction of Pb diffusion is shown in Fig. 3b. The negative values of partial derivatives (DGRAD) show that the height of vegetation contributes to the reduction of Pb concentration. This reduction increases with height. These results for the contribution of the density and the height of vegetation are in agreement with the remarks of Horbert et al. (1988) that plant density and structure provide an intensive decline in contamination in the central area of the parks.
- Concerning the factor AIR, two hypotheses may be made:
  1. 'Good' ventilation allows better Pb diffusion, reducing its high concentrations at points close to the emission source.
  2. 'Poor' ventilation does not facilitate Pb diffusion to distant points and can thus explain the reduction of concentrations in the centre of parks where ventilation is poor.
- The increase of the negative derivatives DAIR with AIR (Fig. 3c) shows that the first hypothesis is more positive. In a previous study (Chronopoulos et al., 1997), it has been pointed out that the dispersion of Pb depends significantly on the facility of the movement of

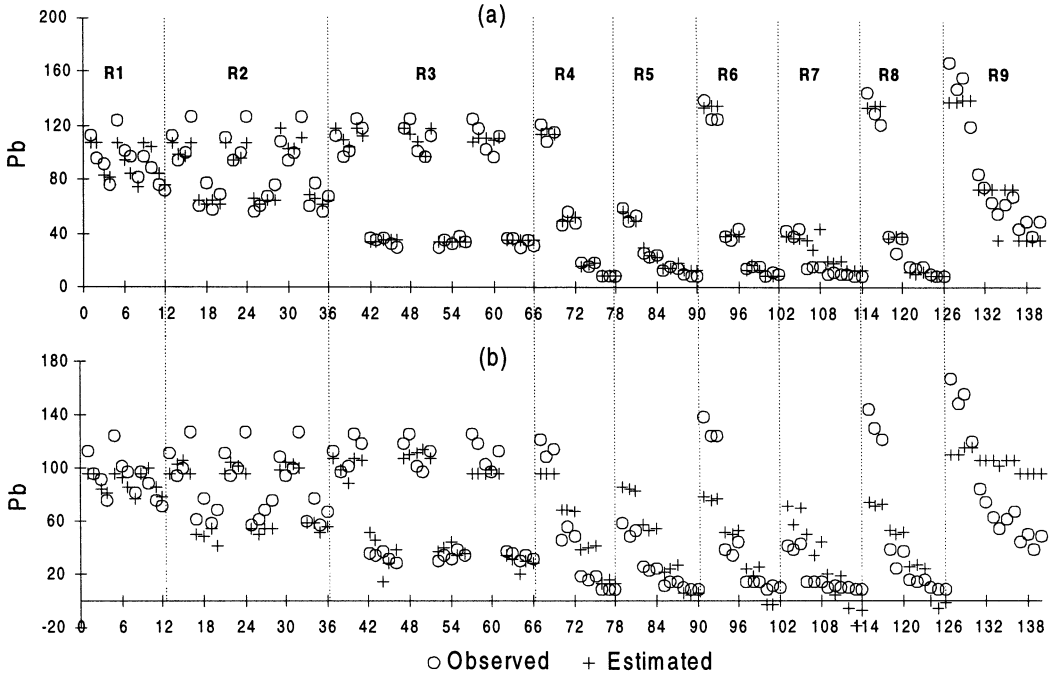

Fig. 4. Pb levels observed and estimated by the MLP model (a) and the MLR model (b) in the different regions of the study area: **R1**: three public squares (1.0 ha in size), **R2**: three public squares (5.0 ha in size), **R3**: three public squares (10.0 ha in size), **R4**: Areos Park (24.0 ha in size), **R5**: Areos Park-Mavromateon Street, **R6**: National Garden (24.0 ha in size)-Vas. Sofias Avenue, **R7**: National Garden-Irodou Attikou Street, **R8**: National Garden-Amalias, **R9**: traffic islands.

air masses, which prohibits or inhibits the dispersion of pollutants. The increase of the negative derivatives DAIR with AIR (Fig. 3c) shows that the first hypothesis is positive.

- The increase of the number of the floors of the adjacent buildings supports the increase of Pb concentrations (Fig. 3d).
- The decrease of Pb concentration with the increase of DIST is evident and nonlinear (Fig. 3e). The reduction of Pb is very intense near the emission points and becomes negligible when the distance is greater than 45 m. The MLR model without taking into consideration the traffic factor as expressed by TRAF is unable to properly estimate the Pb concentrations on traffic islands (Fig. 4b, R9). The slight reduction of the estimated values from the MLR model in that case is due to the fact that the values of the factor BUILD decrease at those points. The MLP, taking into account the factor TRAF gives much better estimations of Pb concentrations.
- The increase of the positive derivatives DTRAF with TRAF shows that Pb concentrations increase with traffic volume as expressed by the number of traffic lanes.

4. Conclusions

The basic idea behind the approach proposed here is the simulation of the system by a statistical model and the use of the resulting model to evaluate the contribution of each explanatory variable to the response of the explained variable. The comparison between the response of the model to the environmental variables on the one hand, and results from field observations on the other hand, shows similarities and indicates neural network modelling can be trusted. MLP adjusts the result of the estimations to the values actually measured. The result can be considered satisfactory since the model built up from a set of

'training data' can predict concentrations for another set of data obtained in the same geographic area.

The advantage of MLP over MLR models seems to arise from the ability of MLP to directly take into account any nonlinear relationships between the Pb concentrations and each explanatory factor. The approach proposed here can be extended to other applications in which non-linear relationships are observed.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. on Automatic Control 19, 716–723.

Akhter, M.S., Madany, I.N., 1993. Heavy metals in street and house dust in Bahrain., Water. Air and Soil Pollution 66, 112–119.

Capannesi, E., Caroli, S., Rosada, A., 1988. Evergreen oak leaves as natural monitor in environmental pollution. J. Radioanal. Nucl. Chem. 123, 713–729.

Chronopoulos, J., Haidouti, C., Chronopoulou-Sereli, A., Massas, I., 1997. Variations in plant and soil lead and cadmium content in urban parks in Athens, Greece. Sci. Total Environ. 196, 91–98.

Cybenko, G., 1989. Approximations by superpositions of a sigmoidal function, Math. Control. Signals and Syst. 2, 303–314.

Dimopoulos, Y., Bourret, P., Lek, S., 1995. Use of some sensitivity criteria for choosing networks with good generalization ability. Neural Process. Lett. 2 (6), 1–4.

Francek, M.A., 1992. Soil lead levels in a small town environment: a case study from Mt Pleasant, Michigan. Environ. Pollut. 76, 251–257.

Grodzinska, K., Szarek, G., Godzik, B., 1990. Heavy metal deposition in Polish National Parks -Changes during 10 years., Water. Air and Soil Pollution 49, 409–419.

Ho, Y.B., Tai, K.M., 1988. Elevated levels of lead and other metals in roadside soil and grass and their use to monitor aerial metal depositions in Hong Kong. Environ. Pollution 49, 37–51.

Horbert, M., Kirchgeorg, A., Chronopoulou-Sereli, A., Chronopoulos, J., 1988. Impact of Green on the Urban Atmosphere in Athens, Scientific Series of the International Bureau KERNFORSCHUNGSANLAGE, 181 pp.

Ndiokwere, C.L., 1984. A study of heavy metal pollution from motor vehicle emissions and its effect on roadside soil, vegetation and crops in Nigeria. Environ. Pollution Ser. B 7, 247–254.

Preer, J.R., 1977. Lead and cadmium content of urban garden vegetables, 11th Annual Conference on trace substances in environmental health, Columbia, June 7–9, pp. 179–187.

Rumelhart, D.E., Hinton, G.E, Williams, R.J., 1986. Learning representations by back-propagating error. Nature 323, 533–536.

Shao-Lian, F., Hashimoto, H., Siegel, B.Z., Siegel, S.M., 1989. Period of significant source reduction., Water. Air and Soil Pollution 43, 109–118.

Sukkop, H., 1990. Stadtokologie-Das Beispiel Berlin, chapter 3. Reimer-Verlag, Germany, p. 425.

Tomassone, R., Lesquoy, E., Miller, C., 1983. La regression, nouveaux regards sur une ancienne méthode statistique. INRA, Paris, p. 188.

Moody, J.E., Utans, J., 1991. Principled architecture selection for neural networks: Application to corporate bond rating prediction. In: Moody, J.E., Hanson, S.J., Lippmann, R.P. (Eds.), Advances in Neural Information Processing Systems 4. Morgan Kaufmann Publishers, San Mateo, CA, pp. 683–690.

Weigend, A.S., Huberman, B.A., Rumelhart, D.E., 1992. Predicting Sunspots and Exchange Rates with Connectionist Networks, In: M. Casdagli and S. Eubank (eds.), Nonlinear Modeling and Forecasting, Addison-Wesley, Redwood City, pp. 395–432.

Weisberg, S., 1980. Applied Linear Regression. Wiley, New York, p. 324.

Wong, J.W.C., 1996. Heavy metal contents in vegetables and market garden soils in Hong Hong. Environ. Technol. 17 (4), 407–414.