Cod: 2.1024

ARTIFICIAL NEURAL NETWORK ESTIMATION OF DIOXIN-LIKE PCBS FROM INORGANIC POLLUTANTS IN AGRICULTURAL SOIL

M.G. Bonelli[1], P. Benedetti[2], M. Ferrini[1], E. Guerriero[2], A. Manni[3]

[1]*Dept. ICMA, University of Rome "La Sapienza", Via Eudossiana 18, 00198 Rome, Italy*

[2]*CNR – Institute for Atmospheric Pollution, Via Salaria, Km 29.300, 00015 Monterotondo Scalo, (Rome), Italy*

[3]*Chemical Research 2000 Srl, Via Santa Margherita di Belice 16, 00133 Rome, Italy*

Introduction

Agricultural soils in proximity to industrialized areas are often subject to contamination by heavy metals and organic micropollutants, such as polychlorinated biphenyls (PCBs). Of the 209 PCB congeners, twelve are highly toxic, possessing toxicological properties similar to that of polychlorinated dibenzo-p-dioxins and furans, and are internationally recognized by UNEP (United Nations Environment Programme) as being a danger to the environment and humans[1]: PCB_{dl} or dioxin-like PCBs. In general, it is possible to characterize polluted land using geostatistical techniques[2] by locating the values of a spatial variable (organic and inorganic compounds) in areas where this variable has not been measured. PCBs and other organic micropollutants are poorly soluble in water similar to many inorganic micropollutants such as heavy metals. The anthropogenic release of inorganic and organic species at high levels generates high concentrations of these pollutants at soil surfaces. This phenomenon allows for the determination of these micropollutants by sampling the surface layer of the land (top soils) and also allows for direct metal determination using certain analytical techniques (e.g. FPXRF). The assessment of pollution in agricultural soils is a difficult task due to the vast size of land that needs to be sampled; which can often be a wasteful expenditure of time and money. Alternatively, it could be useful to find a method to estimate the concentration of unknown contaminants based on the statistical relationship between the organic and inorganic pollutants in contaminated soil. The Artificial Neural Network (ANN) algorithm is an efficient technique for the assessment of unknown values of one or more variables of interest using a cause-effect relationship with predictor variables, but without any hypothesis of linearity between dependent and independent variables. The aim of this study was to prove that the ANN technique is a potential method for the estimation of PCB_{dl} from heavy metals in order to characterize polluted agricultural soil.

Materials and methods

In an agricultural site in Northern Italy, different chemical, metallurgical and pigment operating plants were sources of pollution for urban and cultivated soils. Some previous investigations have found high concentrations of metals in this soil, such as Ca, Hg, Mn, Zn, As, Pb and Ni as well as dioxins and PCBs[3]. After systematic sampling, 73 soil samples were analyzed for heavy metals and dl-PCB. Organic compounds were evaluated by HRGC/HRMS (Trace GC Ultra/ DFS, Thermo, USA) according to the EPA1668C method. Native and mass-labelled reference standards were purchased from Wellington Laboratories Inc., Canada. Heavy metals were analyzed by ICP-MS (Agilent, USA) according to the EPA 6020A 2007 method after mineralization (Ethos Touch Control, Milestone, Italy) according to EPA 3051A 2007 method. Standards were purchased from O2Si, Charleston, USA. In 17 of the 73 soil samples, metals were also measured by FPXRF - Field Portable XRF (Genius 1000, Skyray Instruments, China) in order to assess the performance of this analytical technology in comparison to traditional methods using linear regression analysis. An additional 9 soil samples were collected by a random sampling and were analyzed for heavy metals by FPXRF only. The concentrations detected by the portable instrument were "corrected" to the corresponding ICP-MS values using the regression equations calculated on the above mentioned 17 samples.

Figure 1 demonstrates the results of the linear regression analysis for As and Cu showing very good agreement between the results obtained using these two analytical technologies. This "correction" was performed only for metals with a concentration higher than the limit of detection of FPXRF. In fact, this tool is often used as a screening instrument to quickly locate hot spots in large areas, but it has the disadvantage of having high detection limits for some elements[4], such as Hg and Ni. For the 9 samples measured using FPXRF only, dl-PCBs were extracted and purified following the EPA 1668C method. For detection, an HRGC-MS/MS (Trace GC Ultra/TSQ Triple Stage Quantum, Thermo, USA) analysis was carried out. The artificial neural network (ANN) algorithm is suitable for approximation of complex relationships between input and output variables with a process of non-linearity optimization[5]. From the point of view of modeling, ANN works as a black box[6]: the model can provide accurate results from a series of very different input data, but it cannot explain why and how the results were generated. Instead, ANN can process the available data (input) and produce a prediction of target variables (output), artificially reconstructing the effects of unknown complex cause-effect relationship between input and output during the learning process. Neuron (node) is the basic processing unit in neural networks. The general network architecture consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. In this study, a Multi-Layer Perceptron (MLP) model, consisting of three or more layers (an input and an output layer with one or more hidden layers) of nonlinearly activating nodes, was used. MLP is a function of predictors which minimize the prediction error of target variables, measured quantitatively as Root Mean Square Error (RMSE) of prediction.

The network learning process occurs in the following consecutive phases[7]:

a) Training phase, where the original sample set is divided into three subsets: training set, test set and (optional) holdout set.

b) Validation phase, where the model "learns" using the training set and is further validated using the test set.

c) Test phase, where the model accuracy is estimated by comparing the value of the RMSE training set and the RMSE test set. If both are of the same order, the neural network provides reliable predictions.

All statistical calculations and elaborations were performed using the software package SPSS v. 23.

Results and discussion

Neural networks were applied in several phases. The first step was to only consider the 73 samples analyzed by ICP-MS for 14 metals (Ca, Fe, Mg, K, As, Cd, Cr, Mn, Hg, Ni, Pb, Cu, Sn, Zn) and PCB_{dl}. Pearson's correlation coefficient analysis of PCBdl and inorganic micropollutants showed high and positive correlation between PCB_{dl} and Hg only, while the relationship with others metals seemed non-linear (Table 1). There was also multicollinearity between the predictive variables. The detected PCBs may have been produced unintentionally through combustion processes or as synthetic byproducts, while, in the past, they were products of targeted syntheses. The first case occurs, for example, in incinerators or in iron sintering plants. In these examples, PCBs could be the products of incomplete combustion in the presence of ubiquitous chlorine. The latter case relates to the chlorine industry, where chlorine gas was often produced using a mercury cell (Castner-Kellner process) which could result in contamination of the chlorine with mercury. Therefore it was possible that, in the past, the mercury contaminated chlorine could have been used to synthesize PCBs. Presently, employing chlorine for the chemical synthesis of other chlorinated products could lead to the undesired formation of PCBs which would explain the correlation between mercury and PCBs in the soil samples. Figure 2 shows the first ANN model, called model A, with a MLP feedforward network based on 14 inputs, 3 hidden layers and 1 output. Before being input into the ANN, the data were normalized to a value between 0 and 1. The activation function was a tan hyperbolic for the hidden layers and a sigmoid for the output. The training process was a standard error back-propagation rule, and model A was optimized using the bootstrapping technique to evaluate the reliability of the forecasts[8]. At the end of the learning process, the model produced an estimate of the unknown parameters for each activation function to be used for predictions. Three statistical performance measures were used, including the Determination Coefficient (R^2), the Relative Error (RE), and RMSE (Figure 2). Collectively, these statistics were analyzed to characterize the ANN model performance and the impact of input variables on the estimation accuracy of PCB_{dl}[9]. The model accuracy is 95,7%, with a RE in validation set of 0,43 and a RMSE of 0,08. Therefore, in this case, the trained network delivered very reliable forecasts. The network was then trained on the same samples using only As, Cu, Mn, Pb and Zn as predictive variables; examining how the correlation between one or more input variables and the output would affect the accuracy of the model. The result obtained after the optimization process was model B and its features are shown in Figure 3. The trained network was able to correctly predict 86,2% of the PCBs values, with a validation set RE of 0,922 and a RMSE of 0,026. Therefore, the ANN model keeps a good level of reliability even in absence of a linear relationship between input and output variables. Finally, to prove the actual predictive ability of the trained network and to generalize the method, the second model was applied to the 9 samples (prediction set) in which the concentrations of metals were only measured by FPXRF. The ANN forecasts were successively confirmed by traditional laboratory techniques. Table 2 shows the validation results using performance measures of the prediction set, in comparison with model A and B parameters. Despite the R^2 (92,4%) was higher than the model B, the relative error and RMSE values increase indicate less precise, but still reliable, forecasts.

Conclusions

ANN algorithm is an efficient method for forecasting PCB concentrations from inorganic compounds due its ability to estimate variables in absence of linearity. Furthermore, the use of ANN coupled to a metal portable soil analyzer, such as FPXRF, could be a valid technique for screening large polluted agricultural soils and identifying hot spots quickly and more inexpensively than with traditional and geostatistical techniques.

References

1. UNEP Chemicals: http://www.chem.unep.ch/pos
2. H. Wackernagel (2003) *Multivariate Geostatistics: an introduction with applications*, Springer, Berlin, 161-164
3. Regione Lombardia - ERSAF (2007) *Quaderni della ricerca* **61,** 23-51
4. F. S. Mjalli, S. Al-Asheh and H.E. Alfadala *(*2007) *Journal of Environmental Management* **83***,* 329-338
5. Vilas L.G., Spyrakos E., Palenzuela J.M.T. (2011) Remote Sensing of Environment 11, 524-535
6. D. Kriesel *A Brief Introduction to Neural Networks* http://www.dkriesel.com/_media/science/neuronalenetze-en-zeta2-2col-dkrieselcom.pdf
7. R.E. Uhrig (1995) *Industrial Electronics, Control and Instrumentation* **1**, 33-37
8. Janssen P.H.M., Heuberger P.S.C. (1995) *Ecological Modelling*, **83**, 55-66
9. Zhou L., Ma W. Zhang H. Li L. Tang L. (2015)*Water quality, exposure and health*, **7**, 591-602

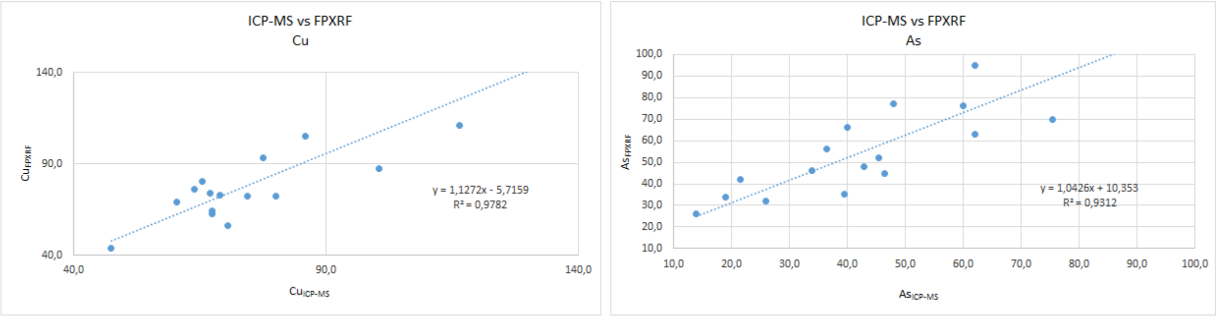

Figure 1. Comparison of ICP-MS and FPXRF performance for Cu and As

Table 1 Correlation matrix between metals and PCB$_{dl}$ (p<0,05, p<0,01)

	Ca	Fe	Mg	K	As	Cd	Cr	Mn	Hg	Ni	Pb	Cu	Sn	Zn	PCB$_{dl}$
Ca	1	-,255*	,268*	-,467**	,508**	,113	,127	,818**	,165	,287**	-,125	,304**	,026	,195*	,175
Fe	-,255*	1	,247*	,722**	-,288**	,150	,327**	-,179	-,172	,109	,154	-,156	,209*	,363**	-,207*
Mg	,268*	,247*	1	,257*	,027	,243*	,371**	,243*	-,189	,232*	-,101	,162	,296**	,360**	-,101
K	-,467**	,722**	,257*	1	-,477**	,191	,403**	-,492**	-,091	,050	,141	-,126	,331**	,342**	-,096
As	,508**	-,288**	,027	-,477**	1	,036	,293**	,542**	,099	,708**	,086	,755**	,132	,473**	,154
Cd	,113	,150	,243*	,191	,036	1	,592**	,286**	,432**	,251*	-,061	,261*	,378**	,333**	,498**
Cr	,127	,327**	,371**	,403**	,293**	,592**	1	,313**	,348**	,667**	,153	,634**	,627**	,744**	,361**
Mn	,818**	-,179	,243*	-,492**	,542**	,286**	,313**	1	,292**	,333**	-,092	,346**	,108	,215*	,233*
Hg	,165	-,172	-,189	-,091	,099	,432**	,348**	,292**	1	,125	,108	,248*	,335**	,152	,840**
Ni	,287**	,109	,232*	,050	,708**	,251*	,667**	,333**	,125	1	,216*	,904**	,474**	,777**	,155
Pb	-,125	,154	-,101	,141	,086	-,061	,153	-,092	,108	,216*	1	,231*	,410**	,179	,006
Cu	,304**	-,156	,162	-,126	,755**	,261*	,634**	,346**	,248*	,904**	,231*	1	,500**	,737**	,340**
Sn	,026	,209*	,296**	,331**	,132	,378**	,627**	,108	,335**	,474**	,410**	,500**	1	,574**	,382**
Zn	,195*	,363**	,360**	,342**	,473**	,333**	,744**	,215*	,152	,777**	,179	,737**	,574**	1	,256*
PCB$_{dl}$	,175	-,207*	-,101	-,096	,154	,498**	,361**	,233*	,840**	,155	,006	,340**	,382**	,256*	1

*. Correlation is significative at level 0,05

**. Correlation is significative at level 0,01

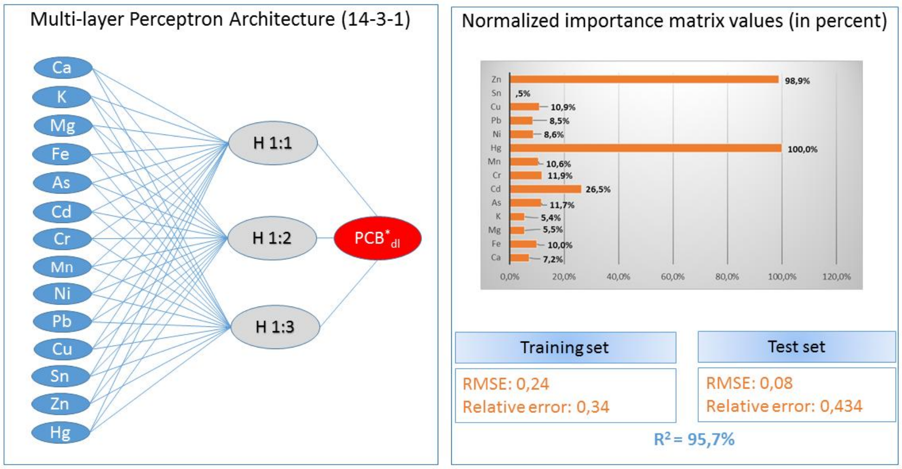

Figure 2. ANNs architecture, Importance matrix and accuracy model A

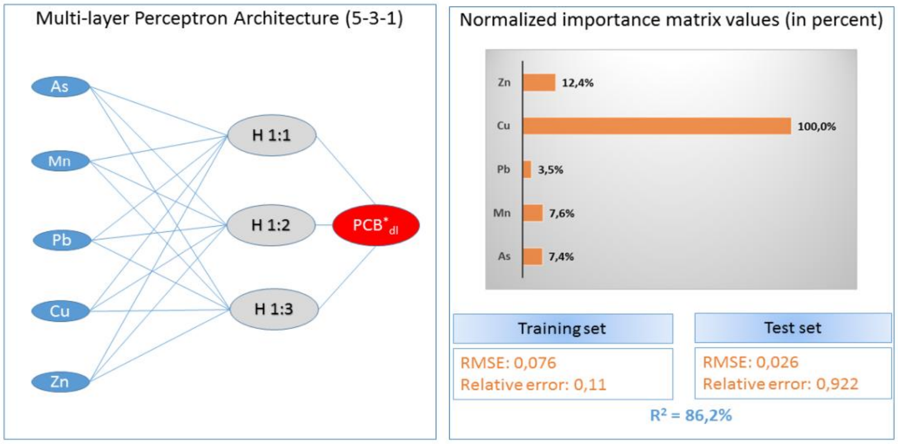

Figure 3. ANNs architecture, Importance matrix and accuracy model B

Table 2 Comparison of the validation parameters in model A, model B and in prediction set

MODEL	DATASET	R^2	RMSE	RE
A	Training	0,957	0,24	0,34
	Validation		0,08	0,43
B	Training	0,862	0,076	0,110
	Validation		0,026	0,922
Prediction set		0.924	7,41	2,24