Artificial neural networks to evaluate organic and inorganic contamination in agricultural soils

Maria Grazia Bonelli, Mauro Ferrini, Andrea Manni

Abstract

The assessment of organic and inorganic contaminants in agricultural soils is a difficult challenge due to the large-scale dimensions of the areas under investigation and the great number of samples needed for analysis. On-site screening techniques, such as Field Portable X-ray Fluorescence (FPXRF) spectrometry, can be used for inorganic compounds, such as heavy metals. This method is not destructive and allows a rapid qualitative characterization, identifying hot spots from where to collect soil samples for analysis by traditional laboratory techniques. Recently, fast methods such as immuno-assays for the determination of organic compounds, such as dioxins, furans and PCBs, have been employed; but several limitations compromise their performance. The aim of the present study was to find a method able to screen contaminants in agricultural soil, using FPXRF spectrometry for metals and a statistical procedure, such as the Artificial Neural Networks technique, to estimate unknown concentrations of organic compounds based on statistical relationships between the organic and inorganic pollutants.

1. Introduction

At the beginning of a contaminated soil characterization procedure it is necessary to know the nature of pollutants, whether they are heavy metals, toxic organics, or both, and the extent of contamination (Perez-Vazquez et al., 2015). Among the organic contaminants, polychlorinated dibenzo-p-dioxins (PCDD) and polychlorinated dibenzofurans (PCDF) are highly toxic and considered persistent organic pollutants (POPs) (UNEP, 2000). In contrast to other chemicals of environmental concern such as polychlorinated biphenyls (PCB), PCDD/PCDFs were never produced intentionally and they are formed as by-products of numerous industrial activities and all combustion processes (Fiedler, 2003).

In contaminated agricultural lands, POPs emissions are often associated with the presence of inorganic compounds, such as heavy metals (Orta-Garcia et al., 2016), due to the use of pesticides, animal manures and fertilizers, and the proximity to emission sources (e.g. solid waste incinerators, urban sites or chemical industries). Metals and metalloids - like Cr, Mn, Fe, Ni, Cu, Zn, Ag, Cd, Pb, Hg and As - are always present in natural soil at trace and non-toxic levels, due to lithogenic inputs, pedogenesis processes and weathering of parent rocks (Alloway, 2013). If their levels increase, however, heavy metals can present problems of toxicity.

In an agricultural site, the assessment of pollution could present several problems due to the vast size of the area subject to investigation and the large number of samples needed for analysis. A conventional approach for environmental soil characterization consists of taking soil samples to analyze in the laboratory through standard techniques such as Atomic absorption spectroscopy (AAS), spectrometry with plasma source (ICP-OES or ICP-MS) and Atomic fluorescence spectroscopy (AFS). High-resolution gas chromatography/high resolution mass spectrometry (HRGC/HRMS) after extensive cleanup are the standard tools to determine PCDD/Fs and PCBs. However, due the long time and high costs related to these methods, only a small number of samples can be analyzed. Geostatistical techniques (i.e. kriging algorithms) are used to quantify uncertainties in the content of contaminants (organic and inorganic) in areas where those variables were not measured (D’Or et al., 2009), generating a continuous map of concentrations to delineate contaminated areas (Huiting et al., 2013) and providing information on the spatial distribution of contamination (Burrough, 2001).

As an alternative, Field Portable X-ray Fluorescence (FP-XRF) spectroscopy is a fast and low cost method to analyze many samples. This is a common analytical technique for on-site screening and quick turnaround analysis of contaminants in environmental samples, but it allows only to measure qualitatively inorganic elements.

In recent times, advances in biotechnology have allowed to rapidly screen dioxin and dioxin-like compounds in many samples, but several limitations compromise their performance (Petronijevic et al., 2015).

Corresponding author.
Email addresses: mariagrazia.bonelli@uniroma1.it (M.G. Bonelli); mauro.ferrini@uniroma1.it (M. Ferrini); info@er2000.it (A. Manni)

© 2017.

Article history:
Received 29 May 2017
Received in revised form 14 July 2017
Accepted 21 July 2017
Available online xxx
Handling Editor: I. Cousins

Keywords:
Artificial Neural Networks
FPXRF
Agricultural soil
Environmental pollution
PCDD/Fs
PCBs

Contents lists available at ScienceDirect
Chemosphere
journal homepage: www.elsevier.com

Chemosphere xxx (2017) xxx-xxx

http://dx.doi.org/10.1016/j.chemosphere.2017.07.116
0045-6535/© 2017.
Instead, the aim of the present study was to find a fast and low cost analytical procedure to screen for hot spots in contaminated soil, using Artificial Neural Network (ANN) data modeling to predict unknown concentrations of POPs using heavy metal values measured by FPXRF.

In this work this procedure has been applied to an Italian agricultural area with soil contaminated by heavy metals, PCDD/Fs and PCBs.

2. Materials and methods

2.1. Sampling site

In 2014, 75 soil samples were collected from a large agricultural field in the region of Lombardy (northern Italy), close to the city of Brescia, where various chemical, metallurgical and pigment factories (some inactive while others still operating) were sources of pollution for urban and cultivated soils.

In the city of Brescia, a few hundred meters from the historic center, the chemical industry Cañaro has been operating since 1906 producing organochlorine compounds, mainly PCBs, in a chloro-alkali plant (ARPA Lombardy, 2001).

Analysis carried out by the Regional Agency for the Protection of the Environment (ARPA Lombardy), Local (ASL Azienda Sanitaria Locale Brescia) and national health authorities (ISS – Istituto Superiore di Sanità -Italian National Institute of Health) proved the high PCB contamination of the area (up to 100 times above the limits of the law) (ASL Brescia, 2002). Consequently, the area was included within the contaminated sites of national interest In the same site, some other investigations, performed by ERSAF Lombardy (Ente Regionale per i Servizi all'Agricoltura e alle Foreste) have found high concentrations of some metals in soil, such as Ca, Hg, Mn, Zn, As, Pb and Ni (ERSAF, 2007).

2.2. Analytical techniques

Heavy metals were analysed by ICP-MS (Agilent, USA) according to the EPA 6020A 1998 method (USEPA, 1998a,b), after their mineralization (Ethos Touch Control, Milestone, Italy) according to EPA 3051A 2007 method (USEPA, 2007). Analytical standards were from O2Si, Charleston, USA.

PCDD/Fs analyses were performed by HRGC/HRMS (Trace GC Ultra/DF, Thermo, USA) according to EPA1613B method (USEPA, 1994). Native and labeled standards were from Wellington Laboratories, Canada. PCBs analyses were performed according to EPA1686C with the above-mentioned instrumentation and analytical standards.

In a subsequent step 45 additional soil samples, collected in 2014 through systematic sampling, were analyzed for heavy metals by FP-XRF - Field Portable XRF (Genius 1000, Skyray Instruments, China), while trained ANN models were used to predict the organic compounds.

EPMA Method 6200 describes an in situ use of FP-XRF (USEPA, 1998a,b), where the tool is placed directly onto the surface soil to analyze or soil samples are scanned in field condition through plastic bags. This method doesn’t require any sample preparation.

Another use of FP-XRF, not described in EPA Method 6200, is the ex situ test, where soil samples are air-dried, ground, sieved and tested through a plastic bag or are powdered, pressed and scanned in a sample cup (Laiho and Perämaa, 2005). Recent studies suggested that thorough sample preparation (e.g. drying, sieving and homogenization, etc.) significantly improved FP-XRF data quality compared to unprepared in situ measurement (Hu et al., 2014). Therefore, to increase the reliability of method, it’s better to consider dried and sieved samples.

2.3. Artificial Neural Networks

An Artificial Neural Network (ANN) algorithm is suitable to approximate complex relationships between the input and output variables with a process of non-linearity optimization (See SI).

In the present work, input variables are metals levels, while PCDD/ Fs and dl-PCB are the output of the ANN model.

All statistical calculations and elaborations were performed using the software package SPSS v. 24. Golden Software Surfer 8.0 was used either for gridding data by ordinary kriging interpolation technique (Dao et al., 2012) and to define contour maps of each element.

2.4. Hot spot screening procedure

The screening procedure used is outlined in the following diagram (Fig. 1):

A preliminary sample set is analysed by traditional instruments, ICP/MS for metals and HRGC/HRMS for POPs and their concentrations are used to train the neural network to create an ANN model.

The same samples, used as secondary certified reference materials, are also analysed by FP-XRF and their metal concentrations are calibrated against AAS, AFS or ICP/MS results considered as standards (Kalnicky and Singhvi, 2001).

In a successive step, a new sample set is collected: metals are measured by FP-XRF, while POPs are estimated by the ANN model.

Calibrated metal values and POPs estimates are used to map soil contamination to detect hot spots, generally characterized by high contaminant concentrations that might be missed by a conventional sampling plain (Horta et al., 2015). Then, a confirmatory test by traditional techniques is necessary. The results of the confirmatory test are used to train again the neural network, improving model accuracy at each repetition of the procedure.

3. Results and discussion

3.1. Preliminary site sampling; site-specific FP-XRF calibration procedure and ANNs model training

In an initial stage of the screening procedure, a total of 30 surface soil samples were collected (10 cm of depth), at distances of 50 m by grid sampling. Each soil sample was analyzed by conventional laboratorial techniques for Cu, As, Zn, Pb, Mn, Hg, Ni, Fe Mg and Cr, PCDD/F and dl-PCB. Metals have also been tested by FP-XRF (ex situ): samples were scanned by being pressed in cup in air-dry and ground (<2,5 mm) condition. For each sample, the analytical design has included ten replicates, considering final concentration value as the median of ten measurements. Then, FP-XRF results have been compared to ICP/MS data to define calibration curve for each metal by linear regression. Good correlation has been found for many of them, (e.g. As, Cu, Zn and Pb with R^2 of 0.789, 0.829, 0.787 and 0.771 respectively). For other metals, such as Hg and Mn, a poor coefficient of determination was attributed to some possible outliers in the distribution and modified Thompson tau (τ) test has been used to identify them (Stauffer, 2017). Outliers have been removed to correct calibration curves but they have been considered in the analysis because they could indicate a hot spot.

As an example, Fig. 2 shows calibration curves for As and Cu.
ICP/MS and HRGC/HRMS analysis results have been used to train the ANN model, using metal levels as predictive or input variables and POPs as output variables. Pearson’s correlation coefficient analysis of organic and inorganic contaminants showed high and positive correlation between PCDD/F and some metals, such as Cu and Ni, and a low correlation with Zn, while there is no linear relationship between dl-PCB and inorganic compounds, except Hg, as well as multicollinearity among predictive variables (Table 1).

In the presence of multicollinearity and non-linear correlation, a traditional forecast method, such as multiple linear regression, does not provide reliable estimates (Farrar and Glauber, 1967; Martins et al., 2007), while the ANNs model does not require any relation form between the input and output variables. Two models, A and B, have been trained for PCDD/Fs and dl-PCBs respectively.

Before being input into the ANN, data were normalized to a value between 0 and 1 to reduce outlier’s influence and to facilitate network learning (Divakar et al., 2007; Azadi and Karimi-Jashni, 2016), using the unity-based normalization (a):

\[
z_i = \frac{x_i - \min (x)}{\max (x) - \min (x)}
\]

Generally, any differentiable function can be used as activation function (Demuth et al., 2008). In this study, hyperbolic tangent (b), sigmoid (c) and linear (d) have been considered, based on the relationships between neurons in layers:

\[
f (x) = \frac{(e^x - e^{-x})}{(e^x + e^{-x})}
\]
<table>
<thead>
<tr>
<th>As</th>
<th>Mn</th>
<th>Pb</th>
<th>Cu</th>
<th>Zn</th>
<th>Hg</th>
<th>Ni</th>
<th>Fe</th>
<th>Mg</th>
<th>Cr</th>
<th>PCDD_F</th>
<th>PCB_DL</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.173</td>
<td>-0.220</td>
<td>0.368</td>
<td>0.329</td>
<td>-0.063</td>
<td>0.162</td>
<td>0.240</td>
<td>-0.133</td>
<td>0.055</td>
<td>0.266</td>
<td>0.154</td>
</tr>
<tr>
<td>Mn</td>
<td>0.173</td>
<td>0</td>
<td>-0.269</td>
<td>0.616</td>
<td>0.678</td>
<td>0.058</td>
<td>0.694</td>
<td>0.832</td>
<td>0.707</td>
<td>0.891</td>
<td>0.259</td>
</tr>
<tr>
<td>Pb</td>
<td>-0.226</td>
<td>-0.269</td>
<td>0</td>
<td>-0.085</td>
<td>-0.079</td>
<td>0.171</td>
<td>-0.153</td>
<td>0.106</td>
<td>-0.197</td>
<td>-0.120</td>
<td>-0.300</td>
</tr>
<tr>
<td>Cu</td>
<td>0.368</td>
<td>0.616</td>
<td>0.086</td>
<td>0</td>
<td>0.388</td>
<td>0.130</td>
<td>0.876</td>
<td>0.591</td>
<td>0.614</td>
<td>0.708</td>
<td>0.82</td>
</tr>
<tr>
<td>Zn</td>
<td>0.329</td>
<td>0.678</td>
<td>-0.079</td>
<td>0.388</td>
<td>0</td>
<td>0.035</td>
<td>0.841</td>
<td>0.817</td>
<td>0.755</td>
<td>0.727</td>
<td>0.636</td>
</tr>
<tr>
<td>Hg</td>
<td>-0.063</td>
<td>0.058</td>
<td>0.174</td>
<td>0.130</td>
<td>0.035</td>
<td>0</td>
<td>0.097</td>
<td>-0.016</td>
<td>-0.050</td>
<td>0.103</td>
<td>0.166</td>
</tr>
<tr>
<td>Ni</td>
<td>0.162</td>
<td>0.684</td>
<td>-0.153</td>
<td>0.878</td>
<td>0.841</td>
<td>0.097</td>
<td>0</td>
<td>0.646</td>
<td>0.757</td>
<td>0.833</td>
<td>0.724</td>
</tr>
<tr>
<td>Fe</td>
<td>0.24</td>
<td>0.632</td>
<td>0.108</td>
<td>0.591</td>
<td>0.817</td>
<td>-0.016</td>
<td>0.646</td>
<td>0</td>
<td>0.681</td>
<td>0.641</td>
<td>0.361</td>
</tr>
<tr>
<td>Mg</td>
<td>-0.133</td>
<td>0.707</td>
<td>-0.157</td>
<td>0.614</td>
<td>0.755</td>
<td>-0.050</td>
<td>0.757</td>
<td>0.681</td>
<td>0</td>
<td>0.802</td>
<td>0.453</td>
</tr>
<tr>
<td>Cr</td>
<td>0.053</td>
<td>0.891</td>
<td>-0.120</td>
<td>0.708</td>
<td>0.727</td>
<td>0.103</td>
<td>0.833</td>
<td>0.641</td>
<td>0.802</td>
<td>0</td>
<td>0.447</td>
</tr>
<tr>
<td>PCDD_F</td>
<td>0.266</td>
<td>0.259</td>
<td>-0.030</td>
<td>0.082</td>
<td>0.636</td>
<td>0.166</td>
<td>0.724</td>
<td>0.361</td>
<td>0.453</td>
<td>0.447</td>
<td>1</td>
</tr>
<tr>
<td>PCB_DL</td>
<td>0.154</td>
<td>-0.103</td>
<td>-0.255</td>
<td>0.088</td>
<td>-0.061</td>
<td>0.556</td>
<td>-0.043</td>
<td>-0.162</td>
<td>-0.245</td>
<td>-0.076</td>
<td>0.070</td>
</tr>
</tbody>
</table>

### Table 1
Correlation matrix.

- **High Correlation** (r ≥ 0.7)
- **Low Correlation** (0.7 < r < 0.5)
- **No correlation** (r ≤ 0.5)

\[
f(x) = \frac{1}{1 + e^{-x}} = \sigma(x)
\]  
\[
f(x) = x = \sigma'(x)
\]

In model A, the activation function was a sigmoid for the hidden layers and a linear for the output layer respectively.

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 (Janssen and Heuberger, 1995). At the end of the learning process, the model produced an estimate of unknown parameters for each activation function to be used for predictions.

Model behavior in training and validation steps was evaluated calculating the following statistical parameters: (e) coefficient of determination \(R^2\), (f) mean absolute error (MAE) and (g) root mean squared error (RMSE) given by:

\[
R^2 = 1 - \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}
\]

\[
MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{Y}_i - Y_i|
\]

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}
\]

\(R^2\) provides the variability measure of the data reproduced in the model. As this test does not give the accuracy of the model, other statistical parameters must be reported. MAE and RMSE measure residual errors, which give a global idea of the difference between the observed and modeled values (Chaloulakou et al., 2003).

Fig. 3 shows model A results, with a Multi-layer Perceptron feed-forward network based on 10 inputs, 1 hidden layer and 1 output (more details about this ANN architecture are reported in Supporting information).

The 30 samples were randomly divided into training and test sets by 72% and 28% respectively. The training set (i.e. 21 data) was used to train the network, the test set (i.e. 9 data) to determine the training stop point and to evaluate the prediction performance: \(R^2\) is 80.5% and MAE \(= 0.866\), RMSE \(= 0.017\) and RMSE \(= 0.098\) are of the same order, therefore model A provides reliable predictions. T-test for means comparison was used to evaluate the forecasting accuracy: the calculated T value \(= 1.81\) is between the endpoints of 95% confidence interval \(= -1.73, 6.48\), supporting the reliability of the trained network.

Normalized importance matrix, computing importance of each predictor in the process of the neural network determination, showed Cu as the predictive variable that more than others is correlated with PCDD/F. Importance matrix has relevance especially if the relationship between variables is not linear.

The same procedure was performed for dl-PCB, training model B. The activation function was a hyperbolic tangent for the hidden layers and a sigmoid for the output layer. According to the form of activation function, output data were normalized to a value between 0 and 1 (Haykin, 1998).

Fig. 4 outlines Model B results, with a MLP feedforward network based on 10 inputs, 1 hidden layer and 1 output.

A training set of 20 samples (66%) and a test set 10 sample (34%) were included. \(R^2\) is 67.5%, RMSE \(= 0.224\) and RMSE \(= 0.153\) are of the same order and T-test (0.979) confirmed the trained network’s reliability. Therefore, the ANN algorithm is an efficient method to forecast POPs concentration also in absence of linearity.

Normalized important matrix proved Hg as the best predictive variable among other inorganics in predicting dl-PCB values.

Model A could approximate a linear regression model. Model B allows to obtain estimate parameters also in presence of no linearity.

### 3.2. Secondary site sampling and soil screening

A successive sampling was carried out in the same site using FPXRF \textit{ex situ}. A total of 45 soil samples were collected and scanned in dried and sieved conditions through plastic bags. Then, metal values were “corrected” using for each element its calibration curve obtained during preliminary site sampling.

PCDD/Fs and dl-PCBs concentrations were predicted by model A and B respectively, considering FP-XRF measures as input variables.

To identify hot spots, spatial distribution of each contaminant (organic and inorganic) was performed by creating a contour map with
cross-validations (Huiting et al., 2013), using ordinary kriging interpolation technique (Dao et al., 2012) (Golden Software Surfer 8.0). As an example, spatial distribution maps have been created for Cu, measured by ICP-MS and ex situ FP-XRF, PCDD/F and dl-PCB measured by HRGC/HRMS and ANN models (Figs. 5–7).

The concentrations analyzed by traditional techniques have been added to FPXRF metal values and ANN POPs forecasts, increasing the number of samples from 30 to 75. In this way, it has been possible to much better define the presence of “hot spots” and avoid loss of time and money analyzing areas with a low concentration of organic and inorganic compounds.

Therefore, an efficient screening design can be produced combining the above-mentioned results together with the preliminary and confirmative results obtained by the standard methods to delineate contaminated areas and to focus the attention where the degree of contamination appears to be higher avoiding the high costs related to the analyses of samples within the law acceptance limit.

Relationships between organic and inorganic micro-pollutants are connected to the geochemical and lithological properties of sampling site; thus, the proposed method is site-specific and it cannot be directly applied to different locations. In other words, the results cannot be generalized without model validation for each site. In fact, it is possible to apply the method to other soil and pollution conditions, but different pedological, geochemical and environmental features could affect either the heavy metals measurements or the relationship between organic and inorganic compounds, influencing both the
metals calibration procedure and the artificial neural network training. Therefore, for each different contaminated site it will be necessary to start the process from the very first step of the method, as shown in Fig. 1.

4. Conclusions

The assessment of pollution in large areas, such as in agricultural soils, can be a difficult challenge, but a sampling design supported by portable field spectrometry and advanced statistical analysis methods, such as Artificial Neural Networks, could feature a new approach for soil contamination screening.

Standard analytical methods, like ICP/MS and HRGC/HRMS, are certainly more precise, but a synergistic use of ANN and a metal portable analyzer could be a valid technique to screen large polluted agricultural lands, for a rapid and cheap detection and spatial mapping of the hot spots which can subsequently be validated via laboratory analysis. In this way, decision models for soil contamination assessment can be optimized by improving the quality and quantity of data, reducing the time needed to evaluate the presence and extent of contamination and the project costs.

Acknowledgment

We are thankful to Dr. Myro Petreas of the California Department of Toxic Substances Control, for her constructive comments on this paper that have improved the manuscript significantly.

We are grateful to ERSAF Lombardy, that has provided some samples and has given us the access to the sampling site.

Finally, we would like to thank Prof. Giuseppe Raspa for several discussions that helped us to focus the geostatistical section.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.chemosphere.2017.07.116

Uncited references

Fig. 6. PCDD/Fs contour maps (ng kg$^{-1}$ TEQ WHO 05). 6.1. PCDD/Fs values measured by HRGC/HRMS according to EPA1613B; 6.2. PCDD/Fs obtained by ANN model after FP-XRF metals determination; 6.3. PCDD/Fs measured by ANN model after FP-XRF metals determination obtained after increasing the number of samples from 30 to 75.
Fig. 7. dl-PCBs contour maps (μg kg$^{-1}$ TEQ WHO 05). 7.1. dl-PCBs values measured by HRGC/HRMS according to EPA1613B. 7.2. dl-PCBs obtained by ANN model after FP-XRF metals determination; 7.3. dl-PCBs measured by ANN model after FP-XRF metals determination obtained after increasing the number of samples from 30 to 75.

References


