

Multiple Linear Regression and Artificial Neural Networks to Predict Time and Efficiency of Soil Vapor Extraction

José Tomás Albergaria & F. G. Martins & M. C. M. Alvim-Ferraz & C. Delerue-Matos

Abstract

The prediction of the time and the efficiency of the remediation of contaminated soils using soil vapor extraction remain a difficult challenge to the scientific community and consultants. This work reports the development of multiple linear regression and artificial neural network models to predict the remediation time and efficiency of soil vapor extractions performed in soils contaminated separately with benzene, toluene, ethylbenzene, xylene, trichloroethylene, and perchloroethylene. The results demonstrated that the artificial neural network approach presents better performances when compared with multiple linear regression models. The artificial neural network model allowed an accurate prediction of remediation time and efficiency based on only soil and pollutants characteristics, and consequently allowing a simple and quick previous evaluation of the process viability.

Keywords

Soil vapor extraction . Artificial neural networks . Multiple linear regression . Remediation time . Process efficiency

1 Introduction

Since the last century, human activities, environmental accidents, or even natural causes have created an enormous number of cases of soil contamination that can represent a risk to public health.

According to the United States Environmental Protection Agency (USEPA), the contaminants that are most commonly found in soils are the halogenated volatile organic compounds and the group of contaminants constituted by benzene, toluene, ethylbenzene, and xylene (USEPA 2007). Soil vapor extraction (SVE) is the remediation technology that is more commonly used to treat soils contaminated with those contaminants (USEPA 2007). This technology uses vacuum pumps to create a movement of air in the soil matrix that remove the contaminant, which is distributed through all the soil phases (Grasso 1993). Even that this technology is considered fast and extremely efficient (Albergaria et al. 2006), there are cases where it can be time consuming and inefficient, namely in soils with high organic matter contents or with low porosity. Following this, the prediction of the remediation time and efficiency becomes essential to avoid unexpected results that could lead to a waste of time, resources, and money.

In the last three decades, the prediction of the remediation time has been the aim of several works that experimented distinct models. Baehr et al. (1989), Falta et al. (1989), and Sleep and Sykes (1989) developed a mathematical models based on differential equations that could be solved numerically. Kaleris and Croise (1997) predicted the remediation time for

continuous and pulsed SVE using the mixed petroleum engineering reservoir numerical model and based on local equilibrium mass transfer. Sawyer and Kamakoti (1998) had a more simple approach by directly using air flow rates for estimating the closure time of SVE. Barnes (2003) incorporated in their model principles of uncertainty analysis, soil gas flow with contaminant vapor transport and decision theory. Alvim-Ferraz et al. (2006) developed the simple mathematical models that predicted remediation times with relative differences not higher than 10 %. However, these models present limitations and were unable to achieve full acceptance from the scientific community. As far as it is known, the utilization of models such as multiple linear regression or artificial neural networks has never been tested as tools to predict time and efficiency of SVE project.

Multiple linear regression (MLR) attempts to fit a linear equation between two or more explanatory variables and a response variable. The equation that generally describes this model is:

$$\hat{Y} = P_0 + P_1X_1 + \dots + P_nX_n \quad (1)$$

where P_i ($i=0, \dots, n$) correspond to the parameters usually estimated by least squares and X_i ($i=1, \dots, n$) are the explanatory variables (Sousa et al. 2006).

Zornoza et al. (2007) used MLR to evaluate the quality of soils through the analysis of several properties and obtained determination coefficients above 0.799 for the relationship between the predicted and the experimental data. Goudarzi et al. (2009) achieved accurate (root-mean-square errors of test set of 0.3705) predictions of soil sorption coefficients of pesticides. In an earlier study, Fass et al. (1999) estimated the environmental half-life of several contaminants (polycyclic aromatic hydrocarbons and herbicides) in soils. A determination coefficient of 0.77 was obtained when the properties of the chemicals and soils were considered in the model construction. However, and as far as it is known, no studies were performed concerning MLR as a predict tool for remediation of contaminated soils using SVE.

Artificial neural networks (ANN) were first introduced as a mathematical tool by McCulloch and Pitts in 1943, whose inspiration was the neural structure of the brain (McCulloch and Pitts 1943). They perform nonlinear transformations of the input data to approximate the output data through learning process from experimental data and exhibiting ability for

generalization beyond training data (Sousa et al. 2006). Being non-parametric and data-driven, neural networks impose a few prior assumptions on the underlying process what turns it less susceptible to model misspecification than most parametric nonlinear methods (Barron 1991). Feedforward Artificial Neural Network (FANN) is one of the most common where the nodes are grouped in three types of layers: input, hidden, and output (Fig. 1). To obtain an accurate model that able to represent the behavior of the system, it is important the choice of the network topology and the neurons' processing function.

In these models, the data is fed to the nodes of the input layer being transmitted to other layers (hidden and output). Activation functions (such as hyperbolic tangent, linear, or sigmoid) are associated with hidden and output nodes (Sousa et al. 2007), transforming the ANN structure in a nonlinear model. The construction of an ANN usually implies three data sets: training, validation, and test. The first set determines the network topology and the associated weights; the validation set is used for avoiding the problem of over fitting and the test set is needed to check the performance of the ANN using different data of the training and validation sets.

Artificial neural networks have been used in several fields of soil studies with excellent results. Poznyak et al. (2007) estimated with an average error of -0.02 ± 0.0795 the anthracene dynamic decomposition and the estimated ozonation rate constant corresponded exactly to the experimental dynamics decomposition of anthracene. Kemper and Sommer (2002) estimated the heavy metal contamination in soils, obtaining coefficients of determination between the predicted and chemically analyzed concentrations, ranging from 0.24 to 0.93. De la

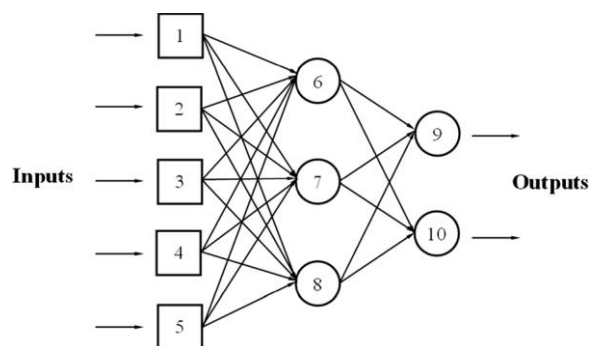


Fig. 1 Feedforward artificial neural network with three layers

Torre-Sanchez et al. (2006) predicted the biodegradation profiles of hydrocarbons in a polluted soil using biopiles with an approximation error below 1.25 %. Therefore, according to the soil heterogeneity and the interaction of a big number of different factors, the application of artificial neural networks to soil vapor extraction seems to be attractive. However, as well as the MLR and as far as it is known, ANN has never been used in experiments involving SVE.

The behavior of both models can be evaluated by calculation of the following performance indexes: mean bias error (MBE), mean absolute error (MAE), root mean squared error, and index of agreement (d_2) (Sousa et al. 2006). The equations to obtain these indexes are:

$$MBE = \frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - Y_i) \quad (2)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{Y}_i - Y_i| \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2} \quad (4)$$

$$d_2 = 1 - \frac{\left[\sum_{i=1}^n |\hat{Y}_i - Y_i|^2 \right]}{\left[\sum_{i=1}^n (|\hat{Y}_i - \bar{Y}_i| + |Y_i - \bar{Y}_i|)^2 \right]} \quad (5)$$

MBE indicates whether the observed values are over or under estimated. MAE and RMSE quantify residual error presenting an overview of the difference between the observed and the estimated values; finally, d_2 compare the difference between the mean, the estimated, and the observed value (Gardner and Dorling 2000; Chaloulakou et al. 2003).

Combining the need to develop a model capable of accurately predict the remediation time of a SVE operation and the good results that ANN have been achieving, it was the objective of this study to predict the remediation time and the process efficiency through a feedforward artificial neural network methodology based on data obtained in SVE pilot experiments and compare it with other mathematical model commonly used as basic prediction model, the multiple linear regression.

Data

2.1 Experimented Soils

SVE experiments were performed in a laboratorial pilot installation using seven simulated soils (sandy and humic) with different moisture (0, 2, 3, and 4 %) and organic matter contents (0, 4, 14, and 24 %), contaminated separately with benzene, toluene, ethylbenzene, xylene, trichloroethylene, and perchloroethylene. The SVE experiments were performed under different airflow rates (2, 5, 10, and 20 L h⁻¹). For each SVE experiment, the vapor pressure and Henry's constant of the contaminant, the moisture, and the organic matter contents of the soils and the airflow used were registered. The preparation of the soil columns as well as

the procedures to perform the SVE experiments is described in Albergaria et al. (2008). All these data were compiled to develop MLR and ANN models which requires a significant amount of data organized in inputs and outputs; the vapor pressure and Henry's constant of

the contaminant, the moisture, and the organic matter contents of the soils and the airflow were used as inputs; and the calculated remediation time and the process efficiency obtained after the SVE (Albergaria et al. 2008) were used as outputs.

SVE experiments were also made in ten real soils with properties within the range of the prepared soils and contaminated separately with the same contaminants. The properties of all experimented soils are presented in Table 1. Altogether, 192 SVE experiments were performed. The properties of all the used soils are indicated in Soares et al. (2010).

2.2 Software Tools

The ANN models were performed using the artificial neural network toolbox of Matlab® (Mathworks, Inc, Natick, MA, USA). The MLR models were performed using subroutines developed in Microsoft Visual Basic applications for Ms-Excel created by the authors.

3 Results and Discussion

The MLR and the ANN models were used to predict the remediation time (t_r) and the efficiency (η) of the SVE. As defined in Albergaria et al. (2008), the remediation time was the operating time required to

Table 1 Properties of prepared (sandy and humic) and real soils

Soil	Moisture content (%)	Organic matter content (%)	pH
Sandy 1	0.0	0.0	8.8
Sandy 2	2.0	0.0	8.8
Sandy 3	3.0	0.0	8.8
Sandy 4	4.0	0.0	8.8
Humic 1	0.7	4.0	6.5
Humic 2	2.3	14	6.1
Humic 3	4.0	24	5.8
Real 1	0.5	0.0	8.1
Real 2	0.6	0.0	8.1
Real 3	0.8	0.0	8.1
Real 4	1.0	0.0	8.1
Real 5	1.2	0.0	8.1
Real 6	2.4	0.0	8.1
Real 7	4.1	0.2	5.5
Real 8	7.2	11	4.0
Real 9	3.0	4.4	4.3
Real 10	6.0	9.4	5.8

achieve a concentration of the contaminant in the gas phase below 1.0 g m^{-3} . The efficiency corresponded to the percentage of the initial amount of contaminant that was removed during the SVE. This time was considered the remediation time. The variables (inputs) considered for these methods were the vapor pressure (P_v) and Henry's constant (H) of the contaminant, moisture (MC), and organic matter contents (OMC) of the soils and the airflow (Q). This choice is based on the influence of these properties on the SVE process. The vapor pressure and the Henry's constant of the contaminant are related with the tendency of the contaminant to be preferentially distributed in the gas phase of the soil hence being easier to be transported by the airflow that percolates through the soil and, consequently, to be remove the contaminant from the soil matrix. The soil water content influences the contaminant partition in the soil matrix and, on the other hand, hinders the air movement through the soil. The organic matter content, responsible for adsorption phenomena of contaminant to the soil, hinders the SVE process because it difficult the mass transfer of the contaminant from the soil phase to the gas phase. Finally, the airflow rate has an important role on the mass transfer of the

contaminant to the gas phase during the remediation process (Albergaria et al. 2006).

3.1 Multiple Linear Regression

Two MLR models were developed: one for the prediction of t_r and other for the prediction of η . The data obtained in the SVE experiments performed with the prepared soils were used to develop these models (Fig. 2).

The models obtained with this method were the following:

$$t_r = 18.75 - 88.86 \times P_v + 5552 \times H + 0.9363 \times \text{OMC} - 1.356 \times Q \quad (6)$$

$$\eta = 90.50 + 94.21 \times P_v + 1.490 \times \text{MC} - 1.202 \times \text{OMC} - 0.4852 \times Q \quad (7)$$

For these models, only the explanatory variables which presented parameters with statistical significance were considered. For these experiments, the moisture content showed to have no significant influence on the t_r while the Henry's constant showed to have no significant influence on the η . The non-significance of the moisture content is not in agreement with other published material (Yoon et al. 2008; Albergaria et al. 2012; Alvim-Ferraz et al. 2006) where it is concluded that the moisture content of the soil hinders the SVE process turning it more time consuming. A possible reason for this result can derive the relatively low number of experiments used in the model development.

Figure 3 illustrates the comparison between the results obtained in the SVE experiments and the results predicted by the MLR models. Table 2 presents the values of the performance indexes of MLR₁ and MLR₂. As can be seen, the data set was divided in two sets: training and test. The training set was composed of 134 experiments and the test set incorporated with 34 experiments. The training set was used to develop the models and the test set was used for testing the model with experiments not used during the models' development. MLR₁ assumed a remediation time of 0 when the predicted value was negative and MLR₂ assumed an efficiency of 100 % when the prediction was higher than 100.

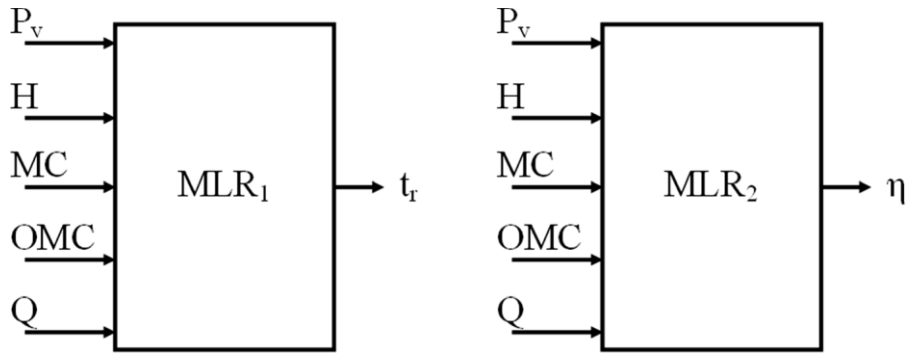


Fig. 2 Structures of the MLR models with the respective inputs and outputs

Attending to Fig. 3 and Table 2, it can be concluded that MLR₁ and MLR₂ were not able to predict accurately neither the remediation time nor the process efficiency, presenting significant errors. This is proved by the values of MAE and RMSE as well as the values of d_2 and R . According to the obtained values of MBE, the training phase of MLR₁ and both test and training phases of MLR₂ gave over-estimated predictions. In order to search for improvements on the results obtained

with the MLR, new models were calculated incorporating data obtained in the tests with real soils. However, worst performance results were obtained, indicating that no benefit is achieved by using this approach.

3.2 Artificial Neural Network

For the ANN model, two approaches were considered. The first approach (ANN₁) considered the

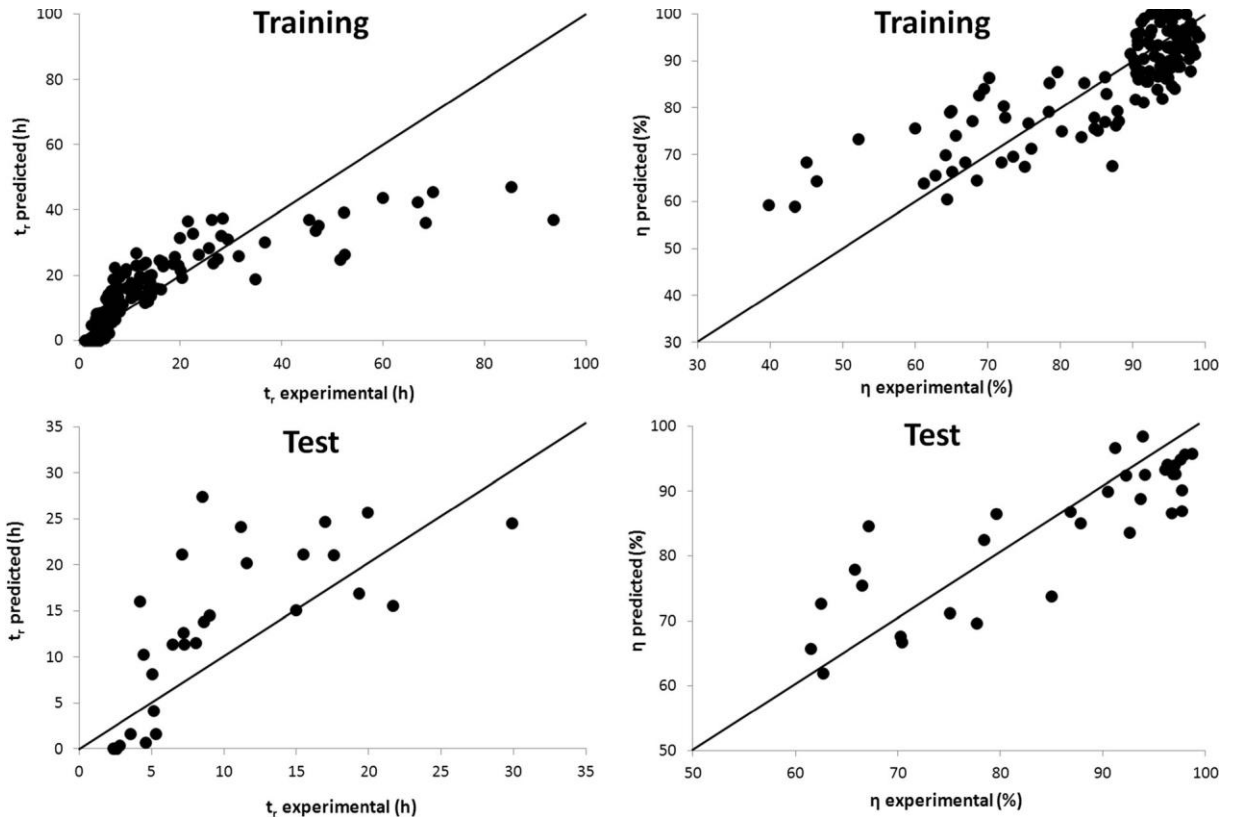


Fig. 3 Comparison between the predicted and experimental results of the efficiency (η) and the remediation time (t_r) of SVE using the MLR models (training and test sets)

Table 2 Performance indexes achieved using MLR₁ and MLR₂

		MBE	MAE	RMSE	d ₂	R
MLR ₁	Training	4.573E-1	6.510	10.07	0.8687	0.8080
	Test	1.691	7.229	9.806	0.8203	0.6863
MLR ₂	Training	-2.131E-1	5.917	7.494	0.8920	0.8208
	Test	-0.1740	5.873	7.678	0.8971	0.8413

utilization of the data exclusively obtained in the experiments with prepared soils (discarding the real soils) to construct the model, while the second approach (ANN₂) considered the utilization of the data of the real soils in the training and in the validation phase. The weights of ANN were obtained by minimization of mean squared error (MSE) of the training data. The cross-validation was applied to avoid the overfitting. For that, the data obtained with simulated soils were divided in training, validation, and test data sets. The training set was composed of 100 experiments, the validation set incorporated with 34 experiments, and the test set also 34. The first data set was used to determine the weights and the second one to decide the time to stop the training procedure (when the minimum value of MSE in validation data set was achieved). This procedure was repeated for different number of hidden neurons that varied from 1 to 8. The best ANN model (Fig. 1) corresponded to the one that had minimum value of MSE in the training and validation data after 1,000 trials and corresponded to architecture of three layers with five neurons in the input layer, three neurons in the hidden layer, and two in the output layer. The variables were the same as in the MLR, and the outputs were also the remediation time and the process efficiency. Hyperbolic tangent and linear were the activation functions used in the neurons of the hidden and output layer, respectively.

The test data set was used to evaluate the performance of the ANN models. The ANN assumed a remediation time of 0 when the predicted value was negative and an efficiency of 100 % when the prediction was higher than 100.

Table 3 presents the values of the performance parameters of ANN₁ both for the training, validation, and test sets.

The values of MBE were negative except in the training and test sets for the prediction of t_r, indicating that in this case the values were overestimated unlike in the others cases, where the predicted values were underestimated. The obtained MAE and RMSE values indicate that the ANN₁ present low residual errors.

Figure 4 illustrates the comparison between the results obtained in the SVE experiments and the results predicted by the ANN₁ model, in all sets.

Comparing Figs. 3 and 4, it is clear that the ANN₁ predicted values accurately and with less errors than those obtained with the MLR models. This observation is confirmed with the results presented in Table 4, which shows the values of the performance indexes of ANN₁ for all sets.

The positive values of MBE show that both predictions were overestimated. Using ANN₁, the values of MAE and RMSE were generally lower indicating that the residual errors were less than with MLR and that the observed values were closer to the predicted

Table 3 Performance indexes obtained for ANN₁ for training, validation, and test sets

	Training		Validation		Test	
	tr	η	tr	η	tr	η
MBE	0.09584	-0.05450	-0.02317	-0.6245	0.6204	-0.9594
MAE	3.116	3.176	4.357	5.117	4.420	4.394
RMSE	4.511	3.954	7.044	6.537	6.467	5.486
d ₂	0.9812	0.9758	0.9443	0.9308	0.9342	0.9633
R	0.9639	0.9541	0.9154	0.8611	0.8775	0.9224

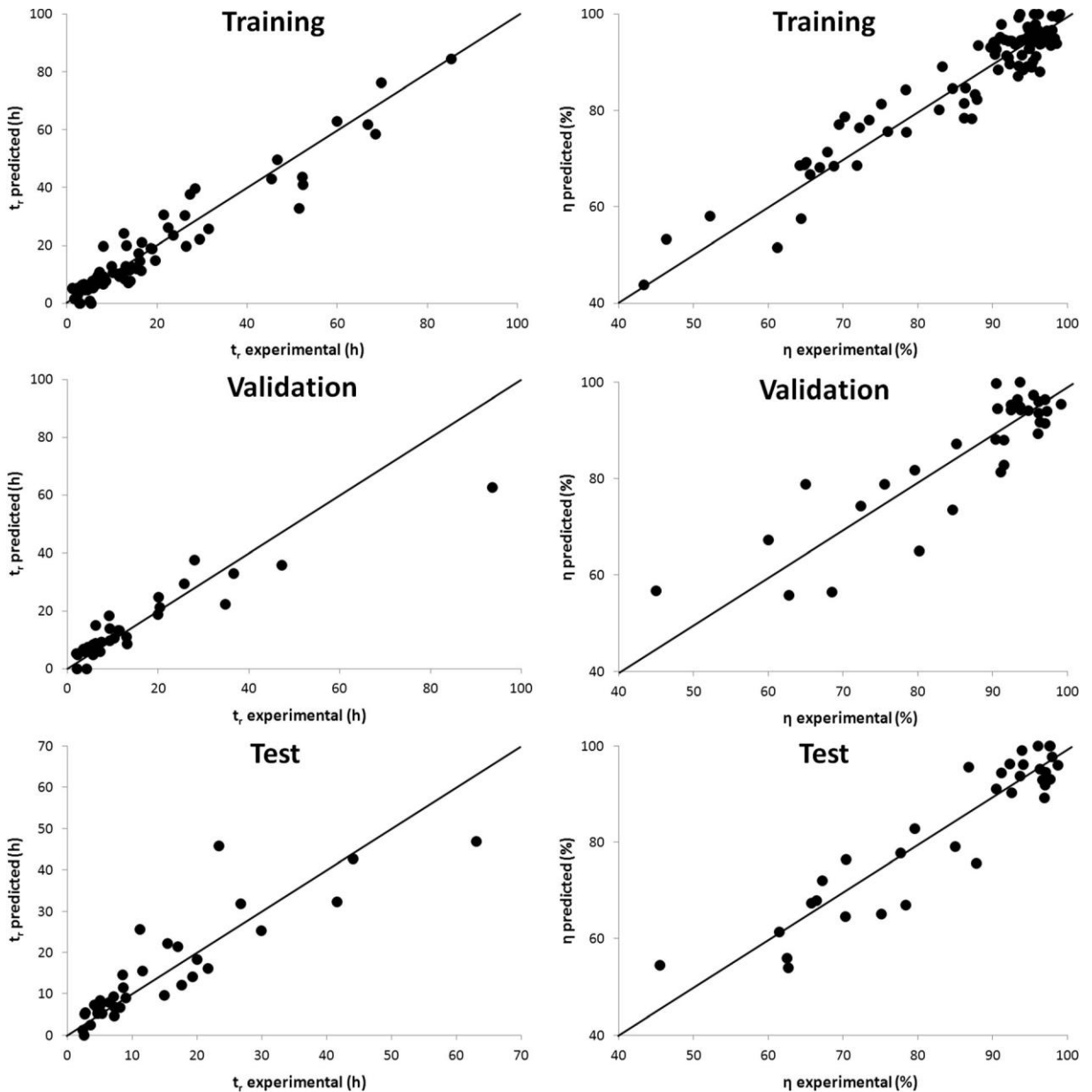


Fig. 4 Comparison between the predicted and experimental results of the remediation time (t_r) and the efficiency (η) of SVE using the ANN₁ model

values. This is also supported by the higher values of R and d_2 .

A second approach (ANN₂) was tested in order to evaluate what was the influence of the utilization of data obtained in SVE performed with real soils in the training and in the validation phases. This aimed the possible inclusion of specificities of real soil in the model construction. The architecture of ANN₂ was similar to ANN₁. The training phase was performed with the results of 116 experiments and the validation and the

test phases with 38. Table 5 shows the performance indexes of ANN₂.

Table 4 Performance indexes achieved using ANN₁ for all data sets

	MBE	MAE	RMSE	d_2	R
t_r	0.05531	3.483	5.337	0.9673	0.9404
η	0.7500	4.601	7.339	0.9165	0.8360

Table 5 Performance indexes achieved using ANN₂ for all data sets

	MBE	MAE	RMSE	d ₂	R
t _r	0.2094	3.799	6.009	0.9591	0.9238
η	0.03344	4.067	6.504	0.9322	0.8738

The values obtained with ANN₂, as well as ANN₁, were also overestimated. Comparing MAE and RMSE and d₂ and R, it can be concluded that both models presented similar performance indexes presenting ANN₂ slightly better results for η and slight worst results for t_r.

Finally, it can be concluded that ANN showed better and very accurate in the prediction of t_r and η in SVE performed with soils similar to those experimented in this work and contaminated with the six contaminants studied, when compared with MLR. These were the first results involving the utilization of these models in this field of research.

4 Conclusions

The work here presented and based on SVE experiments performed in sandy and humic soils contaminated separately with benzene, toluene, ethylbenzene, xylene, trichloroethylene, and perchloroethylene allowed concluding that:

- the MLR models were not able to predict accurately neither the remediation time nor the process efficiency, presenting significant errors,
- the best ANN model corresponded to an architecture of three layers with five neurons in the input layer, three neurons in the hidden layer, and two in the output layer.
- The ANN model showed better and very accurate in the prediction of remediation time and the process efficiency,
- No significant improvements were obtained when data from experiments with real soils were included in the ANN construction.

Acknowledgments This work received financial support from the European Union (FEDER funds through COMPETE) and National Funds (FCT, Fundação para a Ciência e Tecnologia) through project Pest-C/EQB/LA0006/2013.

References

- Albergaria, J. T., Delerue-Matos, C. M., & Alvim-Ferraz, C. M. (2006). Remediation efficiency of vapour extraction of sandy soils contaminated with cyclohexane: Influence of air flow rate and of water and natural organic matter contents. *Environmental Pollution*, 143(1), 146–152.
- Albergaria, J. T., Alvim-Ferraz, M. C. M., & Delerue-Matos, C. (2008). Soil vapor extraction in sandy soils: Influence of airflow rate. *Chemosphere*, 73(9), 1557–1561.
- Albergaria, J. T., Alvim-Ferraz, M. D. M., & Delerue-Matos, C. (2012). Remediation of sandy soils contaminated with hydrocarbons and halogenated hydrocarbons by soil vapour extraction. *Journal of Environmental Management*, 104, 195–201.
- Alvim-Ferraz, M. C. M., Albergaria, J. T., & Delerue-Matos, C. (2006). Soil remediation time to achieve clean-up goals I: Influence of soil water content. *Chemosphere*, 62(5), 853–860.
- Baehr, A. L., Hoag, G. E., & Marley, M. C. (1989). Removing volatile contaminants from the unsaturated zone by inducing advective air-phase transport. *Journal of Contaminant Hydrology*, 4(1), 1–26.
- Barnes, D. L. (2003). Estimation of operation time for soil vapor extraction systems. *Journal of Environmental Engineering-Asce*, 129(9), 873–878.
- Barron, A. R. (1991). Universal approximation bonds for superpositions of a sigmoidal function. Technical report No. 58, Department of Statistics, University of Illinois, Urbana Champaign.
- Chaloulakou, A., Saisana, M., & Spyrellis, N. (2003). Comparative assessment of neural networks and regression models for forecasting summertime ozone in Athens. *Science of the Total Environment*, 313(1–3), 1–13.
- De la Torre-Sanchez, R., Baruch, I., & Barrera-Cortes, J. (2006). Neural prediction of hydrocarbon degradation profiles developed in a biopile. *Expert Systems with Applications*, 31(2), 383–389.
- Falta, R. W., Javandel, I., Pruess, K., & Witherspoon, P. A. (1989). Density driven flow of gas in the unsaturated zone due to the evaporation of volatile organic compounds. *Water Resources Research*, 25(10), 2159–2169.
- Fass, S., Vogel, T. M., Vaudrey, H., Baud-Grasset, F., & Block, J. C. (1999). Prediction of chemicals biodegradation in soils: a tentative of modeling. *Physics and Chemistry of the Earth Part B-Hydrology Oceans and Atmosphere*, 24(6), 495–499.
- Gardner, M. W., & Dorling, S. R. (2000). Statistical surface ozone models: an improved methodology to account for non-linear behaviour. *Atmospheric Environment*, 34(1), 21–34.
- Goudarzi, N., Goodarzi, M., Araujo, M. C. U., & Galvao, R. K. (2009). QSPR modeling of soil sorption coefficients (K-OC) of pesticides using SPA-ANN and SPA-MLR. *Journal of Agricultural and Food Chemistry*, 57(15), 7153–7158.
- Grasso, D. (1993). *Hazardous waste site remediation, source control*. Connecticut: Lewis Publisher Inc.
- Kaleris, V., & Croise, J. (1997). Estimation of cleanup time for continuous and pulsed soil vapor extraction. *Journal of Hydrology*, 194(1–4), 330–356.
- Kemper, T., & Sommer, S. (2002). Estimate of heavy metal contamination in soils after a mining accident using

- reflectance spectroscopy. *Environmental Science & Technology*, 36(12), 2742–2747.
- McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5, 115–133.
- Poznyak, T., Garcia, A., Chairez, I., Gomez, M., & Poznyak, A. (2007). Application of the differential neural network observer to the kinetic parameters identification of the anthracene degradation in contaminated model soil. *Journal of Hazardous Materials*, 146(3), 661–667.
- Sawyer, C. S., & Kamakoti, M. (1998). Optimal flow rates and well locations for soil vapor extraction design. *Journal of Contaminant Hydrology*, 32(1–2), 63–76.
- Sleep, B. E., & Sykes, J. F. (1989). Modeling the transport of volatile organics in variably saturated soils. *Water Resources Research*, 25(1), 81–92.
- Soares, A. A., Albergaria, J. T., Domingues, V. F., Alvim-Ferraz, M. C. M., & Delerue-Matos, C. (2010). Remediation of soils combining soil vapor extraction and bioremediation: Benzene. *Chemosphere*, 80(8), 823–828.
- Sousa, S. I. V., Martins, F. G., Pereira, M. C., & Alvim-Ferraz, M. C. M. (2006). Prediction of ozone concentrations in Oporto city with statistical approaches. *Chemosphere*, 64(7), 1141–1149.
- Sousa, S. I. V., Martins, F. G., Alvim-Ferraz, M. C. M., & Pereira, M. C. (2007). Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. *Environmental Modelling & Software*, 22(1), 97–103.
- USEPA (2007). Treatment Technologies For Site Cleanup: Annual Status Report, 12th Ed., Retrieved January 19, 2014 from http://www.clu-in.org/download/remed/asr/12/asr12_main_body.pdf.
- Yoon, H., Werth, C. J., Valocchi, A. J., & Oostrom, M. (2008). Impact of nonaqueous phase liquid (NAPL) source zone architecture on mass removal mechanisms in strongly layered heterogeneous porous media during soil vapor extraction. *Journal of Contaminant Hydrology*, 100(1–2), 58–71.
- Zornoza, R., Mataix-Solera, J., Guerrero, C., Arcenegui, V., Garcia-Orenes, F., Mataix-Beneyto, J., et al. (2007). Evaluation of soil quality using multiple lineal regression based on physical, chemical and biochemical properties. *Science of the Total Environment*, 378(1–2), 233–237.