Modelling of the adsorption of urea herbicides by tropical soils with an Adaptive-Neural-based Fuzzy Inference System

Babatunde K. Agbaogun1 | Jose M. Alonso2 | Henning Buddenbaum3 | Klaus Fischer1

1Analytical and Ecological Chemistry, University of Trier, Germany 
2Centro Singular de Investigacion en Tecnologias da Informacion (CiTIUS), University of Santiago de Compostella, Santiago, Spain  
3Environmental Remote Sensing and Geoinformatics, University of Trier, Trier, Germany

Correspondence
Babatunde K. Agbaogun, Analytical and Ecological Chemistry, University of Trier, Trier, Germany.
Email: chemie2@uni-trier.de

Funding information
Galician Ministry of Education, University and Professional Training, Grant/Award Numbers: ED431G2019/04, ED431C2018/29, ED431F2018/02; Spanish Ministry of Science, Innovation and Universities, Grant/Award Numbers: RED2018-102641-T, TIN2017-90773-REDT, TIN2017-84796-C2-1-R, RTI2018-099646-B-I00

Abstract
Sorption of pesticides by soils holds a major consequence for their fate in the environment. As such, sorption coefficient ($K_d/K_{oc}$), which is derived from laboratory or field experiments is a fundamental parameter used in almost all screening tools to evaluate the fate or mobility of these compounds. The value of this coefficient is controlled by many soil and solute specific properties, as well as environmental variables. Soft computing techniques such as Adaptive Neuro-Fuzzy Inference System (ANFIS) have been successfully used to predict the equilibrium partitioning of many compounds in various engineered systems. Application of these techniques to natural systems such as soils is however lacking. Here, we present the use of ANFIS in predicting the sorption per unit mass of soil, $Q_e$, used in the calculation of $K_d/K_{oc}$ of compounds in soils. In a previous study, we collected data associated to the adsorption of five phenylurea herbicides in 18 tropical soils. Here, we analysed such data and based on established correlations, nine variables were selected as potential input vectors (i.e., six soil properties, two herbicides molecular descriptors and initial solute concentrations). A total of 255 ANFIS models of one to eight input vectors were elaborated under 10-fold cross-validation. Multiple linear regression (MLR) models were similarly developed, and compared with the ANFIS in terms of mean absolute error (MAE), root-mean-square error (RMSE) and coefficient of determination ($R^2$). The best ANFIS model (M94) has an MAE$_{test}$, RMSE$_{test}$ and $R^2_{test}$ of 3.43 ± 0.43, 4.94 ± 0.80 and 0.95 ± 0.01, respectively, whereas the best MLR model (M13) returned an MAE, RMSE and $R^2$ of 7.71 ± 0.13, 10.11 ± 1.21 and 0.81 ± 0.01, respectively.

Abbreviations: PUHs, phenylurea herbicides; ANN, artificial neural networks; FL, fuzzy logic; TSK, Takagi-Sugeno-Kang; FIS, Fuzzy Inference System; ANFIS, Adaptive Neuro-Fuzzy Inference System; $K_d$, sorption coefficient; $C_{org}$, soil organic carbon; MLR, multiple linear regression; $Q_e$, sorption per unit mass of soil (mg kg$^{-1}$); CEC, effective cation exchange capacity; Fe$_{am}$, amorphous iron content; Mn$_{am}$, amorphous managanese content; log$K_{org}$, PUHs octanol water partition coefficient; Mw, PUH molecular mass; $C_{ini}$, initial solute concentrations (mg L$^{-1}$); stcl, cummulative silt and clay content (%); NFR, number of fuzzy rules; NMF, number of membership function; NM, number of input variables; MF, membership functions; RMSE, root-mean-squared error; MAE, mean absolute error; $R^2$, coefficient of determination.
We observed that generally ANFIS performed better than MLR regarding both accuracy and interpretability. Accordingly, we recommend the use of ANFIS for predicting the sorption coefficients of phenylurea herbicides (PUHs) in soils.

**KEYWORDS**
Adaptive Neural-based Fuzzy Inference System, modelling, multiple linear regression, phenylurea herbicides, sorption, tropical soils

## 1 | INTRODUCTION

Due to several decades of use for agricultural, industrial and urban purposes, phenylurea herbicides (PUHs) and many of their metabolites have been detected in several environmental compartments, and in food products around the world.\(^1\)\(^-\)\(^3\) There are proofs that PUHs exert pathological effects in aquatic organisms, small mammals and humans.\(^1\),\(^4\) Specifically, isoproturon, diuron and linuron are suspected to evoke carcinogenic, mutagenic, teratogenic, endocrine-disrupting and cytogenetic effects in animals as well as humans.\(^1\),\(^5\) Consequently, four PUHs—diuron, chlorotoluron, isoproturon and methabenzthiazuron—have been included in the European Commission’s list of priority substances for European freshwater resources, while diuron, linuron, fluometuron and neburon have also been listed on the U.S. Environmental Protection Agency’s Second Drinking Water Contaminant Candidates.\(^6\),\(^7\)

However, the reactivity, toxicity, bioavailability and mobility of pollutants in the environment are affected by a complex web of many interacting biogeochemical processes. Retention by soils/soil constituents (i.e., sorption) has proven to be the most dominant process.\(^6\),\(^8\) Nevertheless, the degree to which a compound is retained by soil itself is affected by many soil, and solute specific properties.\(^6\),\(^8\),\(^9\) The combination of all these variables creates such complexity that makes it virtually impossible to readily predict how a specific pesticide is likely to be distributed in soils, and/or generally how it will behave in the environment. This therefore necessitates the use of mathematical and computer tools, that is, models, to process the available experimental data (input) and give relevant output. Generally, models are important in environmental research, because they allow complex, nonlinear and multidimensional systems to be investigated, and data from such systems to be interpreted.\(^10\) With models, the interaction of several simultaneous processes and/or factors in a single experiment can be studied. Models are therefore, potentially powerful tools to test and improve our understanding of the environmental fate of pollutants.\(^11\)

Recently, Artificial Intelligence in the broad sense, and Soft Computing (i.e., a family of several preexisting techniques such as Artificial Neural Networks (ANN), Fuzzy Logic (FL), Probabilistic Reasoning and Evolutionary Computation,\(^12\)) in particular, have substantially imparted a new paradigm in the areas of modelling in natural and physical sciences. These techniques are able to work in a cooperative way, taking profit from the main advantages of each individual technique, in order to solve complex real-world problems for which other techniques are not well suited.

FL, which is based on fuzzy set theory, was introduced by Zadeh in 1965 as a logic capable of dealing with approximate reasoning and model uncertainty. Broadly defined, fuzzy set is a class of objects with a continuum of grades of membership ranging from 0–1.\(^13\),\(^14\) Fuzzy sets, usually Mamdani or Takagi-Sugeno-Kang (TSK) systems, provide effective ways of dealing with variable dependencies and their inherent uncertainly through the use of fuzzy ‘IF-THEN’ rules in a process known as a fuzzy inference system (FIS). The ‘IF’ part is called rule premise or antecedent, while the ‘THEN’ part is the conclusion or consequence. FIS are well-known for their interpretability. On the other hand, ANN which represents an advanced mathematical tool, is inspired by the neural structure of the brain/biological nerve system.\(^15\) It is composed of many simple synchronous processing neurons (nodes).\(^16\) The most common structure of an ANN consists of at least three different layers. These layers work similarly as a biological neural system in which a neuron receives inputs from external sources, combines them in some ways, and then predict the output parameters. ANN can identify and learn correlated patterns between input data sets and corresponding target values, and can process nonlinear and complex data.\(^17\) It is therefore widely used for prediction purposes. ANN are well-known because of their accuracy.

The Adaptive Neuro-Fuzzy Inference System (ANFIS) is a hybrid of both ANN and FIS (i.e., a connectionist systems), and as a result it is well-known because of the ability to produce systems with good interpretability-accuracy.
carried out a FIS modelling of Pb II sorption onto mesoporous NiO/ZnCl$_2$-C (i.e., PUH (i.e., diuron) was used soil, using ANN. However, their approach may not be comprehensive enough for the following reasons: (a) only one posed tropical soils. Such dataset was previously built by Agbaogun and Fischer$^6$. The rest of the manuscript is organised building ANFIS models with a sorption dataset for many PUHs of varying molecular descriptors in several differently com-

multivariate techniques like MLR may suffer, it is our postulation that ANFIS offers a better alternative in predicting the sorption coefficients of MLR) models. Notice that, given the likely problems of multicollinearity, which conventional multivariate techniques may suffer, it is our postulation that ANFIS offers a better alternative in predicting the sorption coefficients of ANFIS, ANN and FIS to model Pb (II) adsorption from aqueous solution by ostrich bone ash. Javadian et al.$^27$ also carried out a FIS modelling of Pb II sorption onto mesoporous NiO/ZnCl$_2$-Rosa Canina-L seeds activated carbon nanocomposite. Tanhaei et al.$^28$ reported a neuro-fuzzy modelling of methyl orange removal by magnetic chitosan nanocomposite. Baziar et al.$^29$ applied ANFIS in simulating the sorption of 4-chlorophenol from aqueous solutions by persulphate/nano zero valent iron process. Ghaedi and Vafaie$^{30}$ reviewed the applications of ANN for adsorption of dyes from aqueous solutions. Ghaedi et al.$^{31}$ developed an ANFIS model for adsorption of methylene blue by activated carbon. Ghaedi et al.$^{32}$ used ANFIS to model the adsorption of 1,3,4-thiadiazole-2,5-dithiol by gold nanoparticles-activated carbon. Dolatabadi et al.$^{15}$ compared ANN and ANFIS performances in modelling the simultaneous adsorption of dye and metal ion from aqueous solution by sawdust. ANFIS was also applied by Lashkenari, et al$^{33}$ in predicting the removal of Zn metal ion by $\gamma$-Fe2O3/Polyrhodanine nanocomposite. Several other applications of ANFIS, ANN and FIS can also found in literature, most of which reported successful applications of these techniques.

Despite the reported analytical robustness of these techniques however, as far as we know, there are no such studies in literature which examine their application to modelling sorption of environmentally relevant compounds by soils—especially using the combination of easy to measure/obtain pedotransfer functions (soil properties), and compounds molecular descriptors. Perhaps, the only attempt was that of Ahangar and Shabani$^{34}$ who predicted the $K_d$ of diuron in soil, using ANN. However, their approach may not be comprehensive enough for the following reasons: (a) only one PUH (i.e., diuron) was used—thus, the model may not be applicable to a range of PUHs, and (b) only one input vector (i.e., $C_{org}$) was used, thereby generating only one model. This gave no possibility for comparison and/or ranking of models based on their performance evaluation metrics.

Modelling a fundamental environmental process like adsorption of pollutants by soils, using a few easy to obtain soil and solute specific properties is a very desirable scientific endeavour. This would obviate the need for the error-prone, time consuming, and expensive experimental measurements, and in turn facilitate risk assessment or screening of environmentally relevant compounds. However, given the likely problems of multicollinearity, which conventional multivariate techniques like MLR may suffer, it is our postulation that ANFIS offers a better alternative in predicting the sorption coefficients of PUHs in soils, as well as propagating the inherent uncertainties. In this work, we took a more holistic approach by developing several neuro-fuzzy systems to estimate the sorption capacities ($Q_e$; mg kg$^{-1}$) of soils for PUHs. It is worth noting that the main novelty in this paper is that we searched not only for an accurate prediction but we researched for the smallest combination of inputs to produce such prediction, with high level of accuracy and low uncertainties. We also evaluated the sensitivity of the models to each of the regressed vectors.

Accordingly, we produced robust and trustworthy models with good interpretability-accuracy trade-off. In addition, the predictive capability of the models are successfully compared with similarly developed multiple linear regression (MLR) models. Notice that, given the likely problems of multicollinearity, which conventional multivariate techniques like MLR may suffer, it is our postulation that ANFIS offers a better alternative in predicting the sorption coefficients of PUHs in soils, as well as propagating the inherent uncertainties. Moreover, as far as we know this work is the first to build ANFIS models with a sorption dataset for many PUHs of varying molecular descriptors in several differently composed tropical soils. Such dataset was previously built by Agbaogun and Fischer$^6$.The rest of the manuscript is organised as follows. Section 2 introduces the methodology. Section 3 discusses the main results and provides the reader with final concluding remarks.
2  | MATERIALS AND METHODS

2.1  | Chemicals

Unlabelled standard reference samples (99% purity) of diuron (3-(3,4-dichlorophenyl)-1,1-dimethyl-urea), linuron (3-(3,4-dichlorophenyl)-1-methoxy-1-methylurea), monuron (3-(4-chlorophenyl)-1,1-dimethyl-urea) and chlorotoluron (3-(3-chloro-4-methylphenyl)-1,1-dimethylurea) were obtained from Sigma Aldrich (Germany), while isoproturon (3-(4-Isopropylphenyl)-1,1-dimethylurea) was obtained from Dr Ehrenstorfer, Augsburg, Germany. The chemical structures and properties are shown in Figure 1 and Table 1. Stock solutions (1,000 mg L\(^{-1}\)) were prepared in methanol, from where working solutions of 0.25–27 mg L\(^{-1}\) were prepared in milliQ water (membraPure, 0.055 μs/cm).

2.2  | Soils, sorption isotherm experiments and sample analysis

Eighteen soil samples of different characteristics, originating from the southwestern Nigeria were selected for the sorption experiments. All samples were natural soils collected from different agricultural and non-agricultural fields, between 0- and 20-cm depths. The soils were extensively characterised as previously reported by Agbaogun and Fischer.\(^6\) The sorption experiments were carried out using the batch equilibrium method, while analyses of equilibrium samples were carried out with a Shimadzu 10ADvp Chromatograph (Shimadzu, Tokyo), equipped with a photodiode array detector (SPD-M20A, Prominance). Detailed experimental and analytical procedures are as previously reported by Agbaogun and Fischer\(^6\). From the aqueous phase concentrations \((C_e, \text{mg L}^{-1})\), amounts adsorbed by soils \((Q_e, \text{mg kg}^{-1})\) were calculated based on mass balance as follows:

\[
Q_e = \left(\frac{C_0 - C_e}{m_s}\right) \times V
\]

(1)

where \(C_o\) (mg L\(^{-1}\)) is the concentration of the PUH added to the soil suspension, \(V\) (L) is the volume of the solution and \(m_s\) is the mass of the soil (kg). The solid-liquid (linear) distribution coefficient, \(K_d\) (L kg\(^{-1}\)) is given by Equation 2:

\[
K_d = \frac{Q_e}{C_e}
\]

(2)

From Equations 1 and 2,

\[
K_d = \frac{Q_e V}{C_0 V - Q_e m_s}
\]

(3)

---

**FIGURE 1**  Generic structure of phenylurea herbicides (PUHs)

**TABLE 1**  Common names, structures and properties of the selected phenylureas herbicides

| Common name          | CAS number | \(R^1\) | \(R^2\) | \(R^3\) | log\(k_{ow}\) | Mw (g mol\(^{-1}\)) | \(\alpha\) (A\(^3\)) | \(S\) (mg L\(^{-1}\)) |
|----------------------|------------|---------|---------|---------|--------------|----------------|----------------|-----------------|
| Diuron               | 330-54-1   | Cl      | Cl      | CH\(_3\) | 2.60         | 233.1          | 26.2           | 36              |
| Linuron              | 330-55-2   | Cl      | Cl      | OCH\(_3\) | 3.00         | 249.0          | 27.2           | 64              |
| Monuron              | 150-68-5   | Cl      | H       | CH\(_3\) | 1.81         | 198.0          | 23.9           | 230             |
| Chlorotoluron        | 15,545-48-9| CH\(_3\) | Cl      | CH\(_3\) | 2.51         | 212.0          | 26.1           | 74              |
| Isoproturon          | 34,123-59-6| (CH\(_3\))\(_2\)C | H     | CH\(_3\) | 2.50         | 206.3          | 28.1           | 70              |

Abbreviations. \(\alpha\): molecular polarizability; log\(k_{ow}\), logarithmic value of octanol/water partition coefficient; Mw, molecular weight; \(S\), water solubility at 20°C.
So at any given \( C_0 \), \( K_d \) can be calculated from the predicted values of \( Q_e \) for 1 L of solutes, and 1 kg of soil thus

\[
K_d = \frac{Q_e}{C_0 - Q_e}
\]

### 2.3 Data preparation and input selection

The experimental datasets of \( Q_e \) for five PUHs measured in 18 soil samples earlier reported by Agbaogun and Fischer\(^6\) were adopted for this study. Pearson’s correlation coefficients (\( r \)) were established for the soil physicochemical properties and \( Q_e \), and for the PUHs molecular properties and \( Q_e \), using GNU R (http://cran.r-project.org/). The parameters that showed high and significant correlations with \( Q_e \) were selected as potential regressors for our models. These comprised of six soil properties: organic carbon content (\( C_{org} \)), effective cation exchange capacity (CEC), amorphous Iron and Manganese contents (\( Fe_o \) and \( Mn_o \), respectively), and percentage silt and clay content; and two PUHs molecular properties (octanol water partition coefficient (\( logK_{ow} \)) and molecular mass (\( M_w \))). In addition, the initial solute concentrations, \( C_0 \), which is a standard parameter in phase distributions of chemicals was also included as regressor. However, since percentage silt and clay contents cumulatively represent the non-sand fraction of the soil, we combined the two to make a single regressor (i.e., \( si*cl \)), thus trimming down the number of potential regressors or input variables to eight. This was to avoid complications in the fuzzy rule system. These eight (8) variables were referred to as the input vectors, with \( Q_e \) is the output vector. With these input and output vectors, a dataset of 528 patterns was thence generated, and used for models development as later explained. Table 2 shows the descriptive statistics of the dataset.

### 2.4 ANFIS: Theory, architecture and learning algorithms

The neural fuzzy control systems are based on the TSK fuzzy rules.\(^{35}\) The output of each rule (consequent part) is a linear combinations of input variables (their preconditions) plus a constant term, and the final output is the weighted average of each rule’s output. For instance, assuming that the FIS being considered is of the rule base containing two IF-THEN rules of TSK’s type, with two inputs \( x_1 \) and \( x_2 \), and one output \( y \), then

### Table 2 The statistical description of the experimental data

|                      | \( C_0 \) (mg L\(^{-1} \)) | \( C_{org} \) (%) | CEC (mmolc kg\(^{-1} \)) | \( Fe_o \) (%) | \( Mn_o \) (%) | \( si*cl \) (%) | \( logK_{ow} \) | \( M_w \) (g mol\(^{-1} \)) | \( Q_e \) (mg kg\(^{-1} \)) |
|----------------------|-----------------------------|-------------------|--------------------------|----------------|----------------|----------------|----------------|-----------------------------|-----------------------------|
| Mean                 | 12.3                        | 1.2               | 49.0                     | 0.6            | 0.4            | 39.6          | 2.6            | 220.2                       | 28.2                        |
| Standard error       | 0.4                         | 0.0               | 1.0                      | 0.0            | 0.0            | 0.5           | 0.0            | 0.8                         | 1.0                         |
| Median               | 10.4                        | 1.0               | 44.2                     | 0.6            | 0.3            | 38.2          | 2.5            | 212.0                       | 22.9                        |
| Standard deviation   | 8.5                         | 0.8               | 23.2                     | 0.4            | 0.5            | 10.8          | 0.4            | 18.6                        | 23.2                        |
| Sample variance      | 72.8                        | 0.6               | 539.3                    | 0.2            | 0.2            | 117.3         | 0.2            | 347.2                       | 539.5                       |
| Kurtosis             | −1.3                        | 7.1               | 0.7                      | 2.6            | 8.9            | −0.4          | −0.5           | −1.3                        | 1.4                         |
| Skewness             | 0.1                         | 2.6               | 1.0                      | 1.6            | 3.1            | 0.5           | −0.8           | 0.4                         | 1.1                         |
| Range                | 27.7                        | 3.5               | 90.3                     | 1.7            | 2.2            | 40.7          | 1.2            | 51.0                        | 131.8                       |
| Minimum              | 0.3                         | 0.5               | 21.5                     | 0.2            | 0.1            | 24.4          | 1.8            | 198.0                       | 0.5                         |
| Maximum              | 27.9                        | 4.0               | 111.7                    | 1.9            | 2.3            | 65.0          | 3.0            | 249.0                       | 132.3                       |
| Count                | 528                         | 528               | 528                      | 528            | 528            | 528           | 528            | 528                         | 528                         |
| Correlation with \( Q_e \) | 1.00**                      | 0.90**           | 0.82**                   | 0.64**         | 0.85**         | 0.73**        | 0.90**         | 0.99**                      | 1.00                        |

\*Significance at \( p < 0.01 \).

\**Significance at \( p < 0.001 \).
\[ R_1 : if \ x_1 \ is \ A_{11} \ and \ x_2 \ is \ A_{21}, \ then \ y_1 = f_1 = p_1 x_1 + q_1 x_2 + r_1 \]  
\[ R_2 : if \ x_1 \ is \ A_{12} \ and \ x_2 \ is \ A_{22}, \ then \ y_2 = f_2 = p_2 x_1 + q_2 x_2 + r_2 \]

where \( f_i \) is output and \( p_i, q_i, r_i \) and \( f \) are the consequent parameters of the \( i \)th rule. \( A_i \) and \( A_j \) are linguistic labels whose membership function parameters are premise parameters and are represented by fuzzy sets. The use of linguistic propositions in the antecedent of rules favour interpretability.

Then, the inferred output \( y^* \) is calculated:

\[
y^* = f = \frac{(w_1 f_1 + w_2 f_2)}{w_1 + w_2} = w_1 f_1 + w_2 f_2
\]

where \( w_i \) is the firing strength of \( R_i, j = 1,2, \ldots \), and it is given by

\[
w_j = w_{1A_i}(x_1) + w_{2A_i}(x_2), j = 1,2
\]

As shown in Figure 2, the architecture of ANFIS for this particular example consists of five layers. In the first layer (also known as the fuzzification layer), every node \( i \) is a square node with a node function:

\[
O_1^i = \mu_{A_i}(x)_i
\]

where \( x \) is the input to node \( i \), and \( A_i \) is the linguistic label (e.g., small and large) associated with this node function. In other words, \( O_1^i \) is the membership function of \( A_i \) and it denotes the degree to which the given \( x \) satisfies the label \( A_i \). In FL, there are several membership functions (e.g., linear, triangular, trapezoidal and Guassian). The most popular in ANFIS being the generalised bell (gbellmf) and the gaussian function (gaussmf), but we used the gbellmf in this study. Gbellmf is given as

*Figure 2*  General architecture of an Adaptive Neuro-Fuzzy Inference System (ANFIS) with two input vectors (adapted from Riahi-Madvar et al. with slight modifications)
\[ \mu_{A_i}(x) = \frac{1}{1 + \left( \frac{x - a_i}{a_i} \right)^2 b_i} \]  

(10)

where \( a_i, b_i \) and \( c_i \) are premise parameters whose changes in values cause corresponding changes in the bell-shaped functions, thus exhibiting various forms of membership functions on \( A_i \). In the second layer (product layer), every node multiplies the incoming signals and sends the product out. The firing strength of each node output is giving as

\[ \omega_i = \mu_{A_i}(x_1) \times \mu_{A_j}(x_2), \ i, j = 1, 2, 3 \]  

(11)

the sum of all the rules' firing strengths:

\[ \bar{\omega}_i = \frac{\omega_i}{\omega_1 + \omega_2}, i = 1, 2, \ldots \]  

(12)

The output of this layer is called the normalised firing strength. In the fourth (defuzzification) layer, the node function is given as:

\[ O_i^4 = \bar{\omega}_i f_i = \bar{\omega}_i (p_i x_1 + q_i x_2 + r_i) \]  

(13)

where \( \bar{\omega}_i \) is the product of the third layer, and \( p_i, q_i \) and \( r_i \) in this instance are known as consequent parameters. Finally, the last layer labelled \( \Sigma \) computes the overall output as the summation of all the incoming signals, that is,

\[ O_i^5 = \text{overall output} = \sum_i \bar{\omega}_i f_i = \frac{\sum_i \omega_i f_i}{\sum_i \omega_i} \]  

(14)

There are two types of ANFIS: ANFIS_GRID partitioning and ANFIS subtractive clustering (i.e., ANFIS_SUB). Notice that ANFIS_GRID partitioning was used in this study for the sake of interpretability. Grid partitioning is a method of dividing a data space into rectangular sub-spaces using axis paralleled partition based on predefined number of membership functions and their types, so that overlapping parts do not occur in the input space.\(^{19,38} \) For this method, the number of fuzzy rules is giving as \( NFR = NMF^{NM} \); where \( NM \) is the number of the input variables, \( NMF \) is the number of membership functions per each input, and \( NFR \) is the number of fuzzy rules that will be created by all inputs.\(^{19} \) It has been argued that grid-type fuzzy partitions cannot handle high-dimensional problems with many input variables due to the curse of dimensionality.\(^{19} \) However, since ANFIS has been claimed to make good models even when the MFs is small,\(^{36} \) we only used two (2) MFs in this case in order to overcome this problem. Therefore, even in M1 where all the eight input vectors were used, we had a maximum of 256 fuzzy rules (i.e., \( 2^8 \)).

2.5 | MLR

MLR is a statistical procedure used in predicting the values of a response (dependent) variable from a collection of predictor (independent) variables. In other words, it is the explanation of the outcome value \( y \) as the weighted sum of influences from multiple independent variables \( x_1, x_2, x_3, \ldots \). It is expressed as follows:

\[ y = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_n x_n + \varepsilon \]  

(15)

where \( a_1, \ldots, a_n \) are the regression coefficients which represent the amount of corresponding changes in \( y \) when the independent variables change by 1 unit; \( a_0 \) is a constant, where the regression line intercepts the \( y \) axis, and represents the value of \( y \) when all of the independent variables are zero; \( \varepsilon \) is the random error term.\(^{37} \)
2.6 Models development and performance evaluation/error propagation

With a given inputs/outputs dataset, the FIS uses the backpropagation algorithm in combination with a least square type error method to learn a model from the data. Subsequently, another dataset is used to check the generalisation capability of the resulting ANFIS. In this study, the 528 inputs/outputs dataset was randomly divided into 90% (for training) and 10% (for testing) under 10-fold cross validation. Then, several ANFIS and MLR models were developed with MATLAB version 2019b. We used exactly the same training and testing datasets for both ANFIS and MLR for the sake of a fair comparison.

In a considered opinion on models evaluation, any single metric provides only one projection of the model errors and, therefore, only emphasises a certain aspect of the error characteristics. Thus, a combination of metrics is often required. In this study therefore, we used three error metrics: root-mean-square error (RMSE), mean absolute error (MAE) and the coefficient of determination ($R^2$) to evaluate and compare the performance of the generated models.

These three metrics are given by Equations 16–18.

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{n} (y_i^* - y)^2} \quad (16)$$

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{n} |y_i^* - y| \quad (17)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i^* - y)^2}{\sum_{i=1}^{n} (y_i^* - y_m)^2} \quad (18)$$

where $y^*$ and $y$ are the observed and predicted values, respectively. $y_m$ is the mean value of $y^*$, and $N$ is the number of observations.

RMSE indicates the absolute fit of the model to the data (i.e., how close the observed data points are to the model’s predicted values). The lower the values of RMSE, the better the fit. $R^2$ gives the degree of association between predicted and measured values. One of its useful properties is that its scale is intuitive (i.e., it ranges from zero to one; with zero indicating that the proposed model does not have any prediction power, while one indicates perfect prediction). MAE on the other hand, measures the average magnitude of the errors in a set of forecasts, without considering their directions. Just like RMSE, the lower the values of MAE, the lower the prediction errors. Therefore, the best or optimal model is that which has the least values of MAE and RMSE, and the highest values of $R^2$.

3 RESULTS AND DISCUSSION

Often, theory and experience give only general direction as to which of a pool of candidate variables should be included in the regression model. However, finding the exact best subset of regressors involves two opposing objectives. First, the model must be as complete and as realistic as possible. This implies inclusion of every regressor that even remotely affects the dependent variable in the model. Second, inclusion of as few variables as possible, because each relatively irrelevant regressor decreases the precision of the estimated coefficients and predicted values. Additionally, the presence of extra variables increases the complexity of data collection and model maintenance. Moreover, the smaller the number of variables, the higher the interpretability of the model. Thus, the goal of variable selection is one of parsimony: to realise a balance between simplicity (as few regressors as possible), and fit (as many regressors as needed). There are many strategies to achieve this. However, since we did not have more than eight input vectors, the ‘All Possible Regressions’ approach was adopted. The methodology followed was to start with ‘the all-eight-vector model’, and drop one vector at each subsequent stage, until the last stage of one-vector models. Arising from this approach, a total of 255 models (i.e., 1 eight-vector model, 8 seven-vector models, 28 six-vector models, 56 five-vector models, 70 four-vector models, 56 three-vector models, 28 two-vector models and 8 one-vector models) were developed.
| Model | Input Vectors | ANFIS | | | MLR | | | |
|-------|---------------|-------|-------|-------|-------|-------|-------|-------|
|       |               | Train | Test | Train | Test | Train | Test |
|       |               | MAE   | RMSE  | R²    | MAE   | RMSE  | R²    | MAE   | RMSE  | R²    | MAE   | RMSE  | R²    | MAE   | RMSE  | R²    |
|       |               | Avrg  | S.D.  | Avrg  | Avrg  | Avrg  | S.D.  | Avrg  | Avrg  | Avrg  | Avrg  | S.D.  | Avrg  | Avrg  | Avrg  |
| M1    | Co, Corg, CEC, Feo, Mno, si*cl, logKow, Mw | 1.11  | 0.04 | 1.47  | 0.06 | 1.00  | 0.00 | 5.43  | 2.41 | 10.42 | 7.70 | 0.81  | 7.45 | 0.13 | 9.80 | 0.12 |
|       |               |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |
| M2    | Corg, CEC, Feo, Mno, si*cl, logKow, Mw | 14.44 | 0.22 | 18.63 | 0.26 | 0.36  | 0.36 | 17.81 | 1.98 | 22.74 | 2.44 | 0.12  | 15.06 | 0.22 | 19.36 | 0.22 |
|       |               |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |
| M3    | Co, CEC, Feo, Mno, si*cl, logKow, Mw | 1.20  | 0.04 | 1.60  | 0.08 | 1.00  | 1.00 | 4.66  | 1.83 | 8.84  | 6.34 | 0.84  | 7.60  | 0.11 | 9.95  | 0.12 |
|       |               |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |
| M4    | Co, Corg, Feo, Mno, si*cl, logKow, Mw | 1.36  | 0.05 | 1.89  | 0.10 | 0.99  | 0.99 | 4.43  | 1.65 | 8.23  | 5.28 | 0.86  | 7.57  | 0.10 | 9.88  | 0.13 |
|       |               |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |
| M5    | Co, Corg, CEC, Mno, si*cl, logKow, Mw | 1.27  | 0.04 | 1.68  | 0.07 | 0.99  | 0.99 | 4.60  | 1.61 | 8.60  | 5.48 | 0.85  | 7.46  | 0.10 | 9.82  | 0.12 |
|       |               |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |
| M6    | Co, Corg, CEC, Feo, si*cl, logKow, Mw | 1.22  | 0.05 | 1.62  | 0.09 | 1.00  | 1.00 | 4.95  | 1.74 | 9.16  | 5.63 | 0.84  | 7.45  | 0.10 | 9.80  | 0.12 |
|       |               |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |
| M7    | Co, Corg, CEC, Feo, Mno, logKow, Mw | 1.28  | 0.04 | 1.71  | 0.08 | 0.99  | 0.99 | 4.21  | 1.41 | 7.83  | 5.27 | 0.87  | 7.48  | 0.10 | 9.82  | 0.12 |
|       |               |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |
| M8    | Co, Corg, CEC, Feo, Mno, si*cl, Mw | 2.94  | 0.08 | 4.48  | 0.13 | 0.96  | 0.96 | 5.70  | 0.85 | 8.56  | 1.49 | 0.86  | 7.61  | 0.10 | 9.96  | 0.13 |
|       |               |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |
| M9    | Co, Corg, CEC, Feo, Mno, si*cl, logKow | 3.54  | 0.05 | 5.41  | 0.13 | 0.95  | 0.95 | 6.36  | 1.08 | 9.29  | 2.11 | 0.84  | 8.89  | 0.14 | 11.56 | 0.16 |
|       |               |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |
| M10   | Co, Corg, CEC, Mno, logKow, Mw | 1.59  | 0.06 | 2.22  | 0.09 | 0.99  | 0.99 | 4.09  | 1.56 | 7.99  | 5.70 | 0.86  | 7.48  | 0.13 | 9.84  | 0.12 |
|       |               |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |
| M12   | Co, CEC, Feo, si*cl, logKow, Mw | 1.50  | 0.05 | 2.06  | 0.10 | 0.99  | 0.99 | 3.84  | 1.38 | 6.39  | 3.62 | 0.92  | 7.66  | 0.13 | 10.10 | 0.12 |
|       |               |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |
| M13   | Co, Corg, CEC, si*cl, logKow, Mw | 1.53  | 0.04 | 2.11  | 0.08 | 0.99  | 0.99 | 3.88  | 1.48 | 7.29  | 5.60 | 0.88  | 7.48  | 0.13 | 9.82  | 0.12 |
|       |               |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |
| M14   | Co, Corg, CEC, Feo, Mno, si*cl | 9.51  | 0.20 | 12.60 | 0.17 | 0.70  | 0.70 | 11.55 | 1.84 | 15.42 | 1.94 | 0.57  | 11.02 | 0.13 | 14.06 | 0.13 |
|       |               |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |
| M38   | Co, Corg, CEC, logKow, Mw | 2.39  | 0.04 | 3.42  | 0.11 | 0.98  | 0.98 | 3.54  | 0.59 | 5.12  | 1.03 | 0.95  | 7.52  | 0.13 | 9.85  | 0.12 |
|       |               |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |

(Continues)
### TABLE 3 (Continued)

| Model | Input Vectors | ANFIS | Train | Test | MLR | Train | Test |
|-------|---------------|-------|-------|------|-----|-------|------|
|       |               | MAE   | RMSE  | R²   | MAE | RMSE  | R²   |
|       |               | Avg   | S.D.  | Avg  |     | Avg   | S.D.  |     |
| M39   | Co, Corg, si*cl, logKow, Mw | 2.37  | 0.06  | 3.31 | 0.09 | 5.02  | 1.12  | 0.95 |
| M40   | Co, CEC, si*cl, logKow, Mw | 2.29  | 0.04  | 3.28 | 0.08 | 5.11  | 1.10  | 0.95 |
| M41   | Corg, CEC, Feo, Mno, si*cl | 17.12 | 0.26  | 21.71| 0.27 | 22.44| 2.44  | 0.10 |
|       |               | MAE   | RMSE  | R²   | MAE | RMSE  | R²   |
|       |               | Avg   | S.D.  | Avg  |     | Avg   | S.D.  |     |
| M94   | Co, Corg, logKow, Mw | 3.04  | 0.05  | 4.51 | 0.09 | 4.94  | 0.80  | 0.95 |
| M95   | Co, Corg, CEC, Mw | 3.77  | 0.04  | 5.42 | 0.10 | 7.25  | 1.64  | 0.89 |
| M96   | CEC, Feo, Mno, si*cl | 17.12 | 0.26  | 21.71| 0.27 | 22.44| 2.44  | 0.10 |
|       |               | MAE   | RMSE  | R²   | MAE | RMSE  | R²   |
|       |               | Avg   | S.D.  | Avg  |     | Avg   | S.D.  |     |
| M164  | Co, Corg, Mw | 4.18  | 0.06  | 6.07 | 0.10 | 7.17  | 1.73  | 0.90 |
| M165  | Co, si*cl, Mw | 4.46  | 0.08  | 6.44 | 0.10 | 8.00  | 2.62  | 0.85 |
| M166  | Feo, Mno, si*cl | 17.14 | 0.27  | 21.72| 0.27 | 22.44| 2.43  | 0.10 |
| M167  | Co, logKow, Mw | 6.35  | 0.16  | 10.49| 0.26 | 10.75| 2.43  | 0.79 |
| M220  | Co, Mw | 6.85  | 0.15  | 11.14| 0.26 | 11.30| 2.35  | 0.77 |
| M221  | Co, logKow | 7.17  | 0.15  | 11.40| 0.26 | 11.56| 2.30  | 0.76 |
| M222  | logKow, Mw | 15.73 | 0.22  | 20.74| 0.27 | 20.81| 2.47  | 0.19 |
|       |               | MAE   | RMSE  | R²   | MAE | RMSE  | R²   |
|       |               | Avg   | S.D.  | Avg  |     | Avg   | S.D.  |     |
| M248  | Co | 11.39 | 0.12  | 16.00| 0.24 | 16.00| 2.05  | 0.52 |
| M249  | Mw | 15.98 | 0.23  | 20.95| 0.27 | 21.00| 2.48  | 0.18 |
| M250  | logKow | 16.03 | 0.23  | 21.02| 0.27 | 21.08| 2.50  | 0.18 |
| M255  | Feo | 17.6  | 0.24  | 22.2 | 0.23 | 22.39| 2.09  | 0.1  |
for ANFIS and MLR. Each network was trained and tested, and the corresponding RMSE, MAE and $R^2$ of all the models were calculated and compared. This approach enabled us to select the best models that combine optimal orthogonality with maximum explanatory power. Due to limitation of space and ease of discussion, only a few of the models and their performance evaluation indexes were highlighted for discussion (see Table 3). These models were selected to espouse the major intricacies of both the training and the testing phases of the 255 models.

Regarding ANFIS, the all-eight-vector model (M1) had the least $\text{MAE}_{\text{training}}$ (1.11) and the highest $R^2_{\text{training}}$ (1.00). Nevertheless, it did not compete favourably with the best of seven-, six-, five-, four-, and three-vector models at the testing phase. The best six-vector model (M12) showed better performance than M7 (the best of the seven-vector models). Out of the five-vector models, M39 showed the best performance, followed by M40 and M38. These three also performed better than the best six-vector model (M12). However, there was no significant difference between M39 and the best of four-vector systems (M94). Further reductions in the number of input vectors (beyond four) did not improve the quality of the models, but rather worsened it. Therefore, relying on their performance at both training and testing phases, the following ANFIS models: M94 $>$ M39 $>$ M38 $>$ M40 are taken to be the best models, and ranked in that order. All the above mentioned ANFIS models performed better (i.e., they have lower MAE and RMSE, and higher $R^2$) than the best of MLRs (M6). As a matter of fact, all the ANFIS models recorded better performance metrics than their corresponding MLRs. In order to analyse deeper the effects of random errors in the models, the student's distribution.

**Figure 3** Scatterplots of adsorbed quantity, $Q_e$ (experimental) versus $Q_e$ (predicted) mg kg$^{-1}$ for (a) M94, (b) M39, (c) M38, (d) M40 and (e) M248, for both Adaptive Neuro-Fuzzy Inference System (ANFIS) and multiple linear regression (MLR) test systems.
was applied to calculate the scattering range of the predicted outputs versus actual outputs, at 95% (significance level),
that is, the confidence range in which 95% of all values are expected. Based on the scattering ranges and the distribu-
tions of points around the fitted lines ($y = x$) in Figure 3, one could again infer that ANFIS performed better than MLR.

Arising from all these observations, we can confirm our hypothesis that ANFIS performs better than MLR in
predicting the sorption capacities of soils for PUHs. Ghaedi et al.\textsuperscript{30,31} and several other authors have also concluded that
ANFIS performs better than MLR in predicting adsorption efficiency and/or coefficients. For illustrative purpose, the
interpretability of fuzzy rules in M94 can be observed in Table 4. For instance, the first rule can be read as follows: If $C_o$, $C_{org}$, Mw and $\log k_{ow}$ take small values then the predicted adsorption is computed as the following linear regression:

$$Q_e = -4.127903 + 5.687517 C_o - 0.055349264 C_{org} - 1.5698494 Mw + 0.077525616 \log k_{ow}.$$ 

Therefore, all four selected ANFIS models (M94, M39, M40 and M38) are good estimators of adsorption capacities of soils for PUHs.

In addition, the corresponding MLR equations for the best four ANFIS are given in Equations 19–22, as follows:

$$Q_e = -139 + 1.98 C_o + 9.86 C_{org} - 9.08 \log k_{ow} + 0.70 Mw \quad (19)$$

$$Q_e = -142 + 1.98 C_o + 6.57 C_{org} + 0.14 CEC - 8.81 \log k_{ow} + 0.70 Mw \quad (20)$$

$$Q_e = -145 + 1.98 C_o + 7.38 C_{org} + 0.22 \text{sicl} - 9.26 \log k_{ow} + 0.70 Mw \quad (21)$$

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure3.png}
\caption{Continued}
\end{figure}
3.1 Sensitivity analysis

It is pertinent to point out some intricacies in the developed models, especially with respect to the sensitivity of the models to each of the input vectors. For instance, M1 which was constructed with all the eight variables had ANFIS MAE\text{training} and MAE\text{test} of 1.11 and 5.43, respectively. Whereas, when $\text{Co}$ was excluded (i.e., M2), the MAEs increased to 14.44 and 17.81, respectively. RMSE values also increased from 1.47 and 10.42 to 18.63 and 22.74, respectively, for training and testing. In addition, $R^2$ dropped from 1.00 and 0.81 for training and testing, respectively, to 0.36 and 0.12. Similar trends were observed for MLR. Generally from Table 3, it can be noticed that $\text{Co}$ has the greatest effects on the performance metrics, followed by $\text{Mw}$. However, apart from these two, our results did not support crisp ranking of the remaining six input vectors into third, fourth, fifth, sixth, seventh and eighth orders of importance, respectively. They can be said to have almost the same predictive powers (i.e., degeneracy in chemistry).

Furthermore, as stated earlier, three sets of input variables were used in this study: (a) six soil physico-chemical properties, (b) two PUHs molecular descriptors and (c) one operational variable. From the reported results (see Table 3), it was observed that soil properties alone returned a very poor model (M41), with an $R^2$ of 0.10. This

\[
Q_e = -151 + 1.97\text{Co} + 0.19\text{CEC} + 0.34\text{si/C} - 9.12\log K_{ow} + 0.71\text{Mw}
\]
| Rules | Premise parameters | Consequent parameters |
|-------|-------------------|-----------------------|
| 1     | IF v1 IS in1mf1 AND v2 IS in2mf1 AND v3 IS in3mf1 AND v4 IS in4mf1 THEN | out1mf1 – z = -4.127903 + 5.687517v1 + -0.055349264v2 + -1.5698494v3 + 0.077525616v4 |
| 2     | IF v1 IS in1mf1 AND v2 IS in2mf1 AND v3 IS in3mf1 AND v4 IS in4mf2 THEN | out1mf2 – z = 180.19455 + 0.8085636v1 + 0.38135612v2 + -0.15521085v3 + 29.112919v4 |
| 3     | IF v1 IS in1mf1 AND v2 IS in2mf1 AND v3 IS in3mf1 AND v4 IS in4mf1 THEN | out1mf3 – z = -3.8253684 + -7.142671v1 + 0.04568574v2 + 0.59210825v3 + -0.120955646v4 |
| 4     | IF v1 IS in1mf1 AND v2 IS in2mf1 AND v3 IS in3mf1 AND v4 IS in4mf2 THEN | out1mf4 – z = -2.1034973 + -9.113726v1 + 0.10809665v2 + -6.879371v3 + 2.2672744v4 |
| 5     | IF v1 IS in1mf1 AND v2 IS in2mf2 AND v3 IS in3mf1 AND v4 IS in4mf1 THEN | out1mf5 – z = 1.0726746 + 18.549412v1 + -0.20757745v2 + -5.1022186v3 + 0.1714676v4 |
| 6     | IF v1 IS in1mf1 AND v2 IS in2mf2 AND v3 IS in3mf2 AND v4 IS in4mf2 THEN | out1mf6 – z = 285.2058 + -2.70496v1 + -5.1133356v2 + -0.53474575v3 + 55.43162v4 |
| 7     | IF v1 IS in1mf1 AND v2 IS in2mf2 AND v3 IS in3mf2 AND v4 IS in4mf2 THEN | out1mf7 – z = -11.831079 + -23.211313v1 + 0.48203982v2 + 1.9483917v3 + 3.1973724v4 |
| 8     | IF v1 IS in1mf1 AND v2 IS in2mf2 AND v3 IS in3mf2 AND v4 IS in4mf1 THEN | out1mf8 – z = -7.477879 + -29.670479v1 + 0.59202933v2 + -22.410894v3 + 2.7649508v4 |
| 9     | IF v1 IS in1mf2 AND v2 IS in2mf1 AND v3 IS in3mf1 AND v4 IS in4mf1 THEN | out1mf9 – z = 47.408077 + 79.503296v1 + -0.19606948v2 + -21.89567v3 + -3.1777375v4 |
| 10    | IF v1 IS in1mf2 AND v2 IS in2mf1 AND v3 IS in3mf1 AND v4 IS in4mf2 THEN | out1mf10 – z = 1353.7849 + -11.468742v1 + -10.301582v2 + -2.2396328v3 + 87.1057v4 |
| 11    | IF v1 IS in1mf2 AND v2 IS in2mf1 AND v3 IS in3mf1 AND v4 IS in4mf2 THEN | out1mf11 – z = -7.830193 + -99.63051v1 + 1.4203159v2 + 8.3186643v3 + -2.9155204v4 |
| 12    | IF v1 IS in1mf2 AND v2 IS in2mf1 AND v3 IS in3mf2 AND v4 IS in4mf2 THEN | out1mf12 – z = -3.4263732 + -127.25193v1 + 2.463414v2 + -96.0921v3 + -2.5602899v4 |
| 13    | IF v1 IS in1mf2 AND v2 IS in2mf2 AND v3 IS in3mf1 AND v4 IS in4mf1 THEN | out1mf13 – z = -39.965275 + 250.51112v1 + -0.83661354v2 + -69.014435v3 + -4.752681v4 |
| 14    | IF v1 IS in1mf2 AND v2 IS in2mf2 AND v3 IS in3mf1 AND v4 IS in4mf2 THEN | out1mf14 – z = 1178.5208 + -36.064487v1 + -25.527784v2 + -7.0259757v3 + 108.59624v4 |
| 15    | IF v1 IS in1mf2 AND v2 IS in2mf2 AND v3 IS in3mf2 AND v4 IS in4mf2 THEN | out1mf15 – z = 24.690424 + -314.05127v1 + 2.4320133v2 + 26.184992v3 + 2.0234947v4 |
| 16    | IF v1 IS in1mf2 AND v2 IS in2mf2 AND v3 IS in3mf2 AND v4 IS in4mf2 THEN | out1mf16 – z = 53.110725 + -401.05057v1 + 5.979349v2 + -302.82263v3 + -4.276886v4 |

Note: v1, v2 and v4 are Co, Conge, Mw and logkow, respectively.
contradicts Ahangar and Shabani\textsuperscript{34} who reported an $R^2$ of 0.99 for $K_d$ prediction using only $C_{org}$. As pointed out earlier, the scantiness of their data (i.e., only 36 observations which were further divided into 24/6/6—for training, validation and testing, respectively) might have further exacerbated the problem of overfitting, which is even the natural drawback of ANN which they used. Also, PUHs molecular descriptors alone returned a very poor model (M222), with an $R^2$ of 0.19. A subset of soil properties and operational variable alone returned an average model, with an $R^2$ of 0.57, while the subset of operational variable plus PUHs molecular descriptors resulted in a fairly good model (M167), with an $R^2$ of 0.79. Some models combining the three sets of variables however gave very good performances, with $R^2$ as high as 0.95. Other evaluation metrics (MAE and RMSE) also followed similar trends.

4 | CONCLUSION

The mobility, bioavailability, bioaccumulation and transformation reactions of chemicals in soils depend largely on their sorption coefficients. Thus, adequate knowledge of the sorption coefficients of xenobiotics in soils is important to environmental/analytical chemistry and allied fields. Determining this experimentally is often not a practical option, as it can be time-consuming, cost-intensive and error prone.\textsuperscript{41} Direct measurements may even be practically impossible in cases where partitioning is extremely high or low, the target chemicals are not commercially available, or a large number of compounds must be evaluated\textsuperscript{41}. Considering the high spatial variations of soil compositions, even where direct measurements are possible, data stemming from sorption investigation in soils, just like any other geochemical/ecological investigations, are usually imprecise.\textsuperscript{42} Using such data to predict the partition coefficients of chemicals in the environment may be fraught with uncertainties. Therefore, there is a need for reliable methods that can predict the sorption coefficients of compounds in soils, from few easy to collect or readily available soil attributes and compound specific molecular properties. One of the most successful ways of dealing with such problems is the FL approach. From the practical point of view, and within the scope and limit of this work, we have used ANFIS and MLR to model the $Q_e$ of some PUHs in tropical soils. We have shown how ANFIS generate models which take profit from the interpretability of FIS and accuracy of ANN; thus yielding models with a good interpretability-accuracy trade-off. This is the first time such task was carried out using ANFIS. Comparatively, ANFIS performed better than MLR in terms of all considered evaluation metrics. Thus, our hypothesis was confirmed, and we conclude that ANFIS punishes a linear combination of potential regressors with respect to multicollinearity and accumulated errors better than MLR.

Overall, the following models or subsets of regressors are considered to have displayed the best performance: ANFIS M94 with four input vectors ($C_o$, $C_{org}$, log$k_{ow}$ and Mw), ANFIS M39 with five input vectors ($C_o$, $C_{org}$, si*cl, log$k_{ow}$ and Mw), ANFIS M40 with five input vectors ($C_o$, CEC, si*cl, log$k_{ow}$) and ANFIS M38 also with five input vectors ($C_o$, $C_{org}$, CEC, log$k_{ow}$ and Mw). Each of them was found to account for at least 95% (i.e., $R^2$ 0.95) of the predicted $Q_e$ values, with relatively low uncertainties. They are hereby presented for the task of predicting the retention capacities of soils for any compounds in the PUH molecular group, without a need for laboratory or field measurements. Nevertheless, as future work, with the aim of predicting compounds and soils scenarios outside the range of the present study, we intend to further train the models with additional data so as to improve their generalisation and prediction capabilities.

ACKNOWLEDGEMENT

The first author is grateful to a.i.m rlp—agency for international HEI mobility Rhineland-Palatinate, Germany, for the award of an Erasmus+ sponsorship for research stay at Centro Singular de Investigacion en Tecnoloxias da Information (CiTIUS), University of Santiago de Compostella, Spain. This research is partially supported by the Spanish Ministry of Science, Innovation and Universities grants: RTI2018-099646-B-I00, TIN2017-84796-C2-1-R, TIN2017-90773-REDT, RED2018-102641-T, and the Galician Ministry of Education, University and Professional Training grants: ED431F2018/02, ED431C2018/29, ED431G2019/04 awarded to the second author. The authors would be glad to oblige the request for the MATLAB source codes for further work based on these models.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/cem.3335.

ORCID

Babatunde K. Agbaogun  https://orcid.org/0000-0002-5352-6108
29. Baziar M, Nabizadeh R, Mahvi AH, Alimohammadi M, Naddafi K, Mesdaghinia A. Application of adaptive Neural Fuzzy Inference System and Fuzzy C—means algorithm in simulating the 4-chlorophenol elimination from aqueous solutions by persulfate/nano zero valent iron process. Eurasian J Anal Chem. 2018;13(1):em03. https://doi.org/10.12973/ejac/80612
30. Ghaedi AM, Vafaei A. Applications of artificial neural networks for adsorption removal of dyes from aqueous solution: a review. Adv Colloid Interface Sci. 2017;245:20-39. https://doi.org/10.1016/j.cis.2017.04.015
31. Ghaedi M, Ghaedi AM, Abdi F, Roosta M, Vafaei M, Asghari A. Principal component analysis-adaptive neuro-fuzzy inference system modeling and genetic algorithm optimization of adsorption of methylene blue by activated carbon derived from Pistacia khinjuk. Ecotoxicol Environ Saf. 2013;96:110-117.
32. Ghaedi M, Hosaininia R, Ghaedi AM, Vafaei A, Taghizadeh F. Adaptive neuro-fuzzy inference system model for adsorption of 1,3,4-thiadiazole-2,5-dithiol onto gold nanoparticles-activated carbon. Spectrochim Acta, Part A. 2014;131:606-614.
33. Lashkenari MS, KhazaiePoul A, Ghasemi S, Ghorbani M. Adaptive neuro-fuzzy inference system prediction of Zn metal ions adsorption by γ-Fe2O3/polyrhodanine nanocomposite in a fixed bed column. Int J Eng. 2018;31(10):1617-1623.
34. Ahangar AG, Shabani A. Predicting soil sorption coefficients of environmental pollutant herbicide (Diuron) using a neural network model. Health Scope. 2014;3(2):e14974. https://doi.org/10.17795/jhealthscope-14974
35. Takagi T, Sugeno M. Fuzzy identification of systems and its application to modeling and control. IEEE Trans Syst Man Cybern Part A Syst Humans. 1985;15:116-132. https://doi.org/10.1109/TSMC.1985.6313399
36. Jang JSR. ANFIS: Adaptive-Network-Based Fuzzy Inference System. IEEE Trans Syst Man Cybern Part A Syst Humans. 1993;23(3):665-685. https://doi.org/10.1109/21.256541
37. Riahi-Madvar H, Dehghani M, Seifi A, et al. Comparative analysis of soft computing techniques RBF, MLP, and ANFIS with MLR and MNLR for predicting grade-control scour hole geometry. Eng Appl Comput Fluid Mech. 2019;13(1):529-550. https://doi.org/10.1080/19942060.2019.1618396
38. Sadouki L, Haddad B. Adaptive neuro-fuzzy inference system for echoes classification in radar images. In Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP). 2016; 4:VISAPP,149–166. https://doi.org/10.5220/0005717401590166
39. Chai T, Draxler RR. Root mean square error (RMSE) or mean absolute error (MAE)?—arguments against avoiding RMSE in the literature. Geosci Model Dev. 2014;7(3):1247-1250. https://doi.org/10.5194/gmd-7-1247-2014
40. Martin G. Assessing the fit of regression models. 2020. https://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/ Accessed March 4, 2020.
41. Santoshi E, Kai-Uwe G. Applications of polyparameter linear free energy relationships in environmental chemistry. Environ Sci Technol. 2014;48:12477-12491.
42. Salski A. Fuzzy knowledge-based models in ecological research. Ecol Model. 1992;63(1-4):103-112.

How to cite this article: Agbaogun BK, Alonso JM, Buddenbaum H, Fischer K. Modelling of the adsorption of urea herbicides by tropical soils with an Adaptive-Neural-based Fuzzy Inference System. Journal of Chemometrics. 2021;35:e3335. https://doi.org/10.1002/cem.3335