Influence of soil properties on the sorption and retention of cadmium, copper and lead, separately and together, by 20 soil horizons: Comparison of linear regression and tree regression analyses

F.A. Vega, M.L. Andrade, E.F. Covelo*

Departamento de Bioloxía Vexetal e Ciencia do Solo, Facultade de Bioloxía, Universidade de Vigo, Lagoas, Marcosende, 36310 Vigo, Pontevedra, Spain

ARTICLE INFO

ABSTRACT

In this paper we compare linear regression with tree regression for analysis of the influence of soil properties on the sorption and retention of added Cd, Cu and Pb by 20 soil horizons typical of cropped soils in Galicia (N.W. Spain); our measure of sorption/retention capacity was K_r, a recently introduced adimensional parameter. Sorption and retention of Cd was depressed by the presence of Cu and Pb. The soil parameters that were most associated, overall, with differences in Cd, Cu and Pb sorption and retention were cation exchange capacity (CEC), pH and Mn oxides' content. Tree regression, which can take into account variation on both global and local scales, afforded better-fitting models than linear regression, which only reflects global tendencies; but for coherent interpretability of tree regression results it is just as important to avoid overfitting as in the case of linear regression.

© 2009 Published by Elsevier B.V.

1. Introduction

The availability of heavy metals and other soil pollutants for uptake by plants, and the risk of their finding their way into surface or underground waters, depend on their sorption and desorption by soil components, where "sorption" encompasses adsorption, precipitation on soil particle surfaces, and fixation, and "desorption" indicates the release of sorbed species into the medium surrounding the particles by which they had been sorbed [1–3]. The sorption and desorption of metals – especially metal cations, the directly bioavailable form – predominantly involves negatively charged surfaces on organic matter, clay minerals, and metallic oxides or hydroxides, especially Fe and Mn oxides [4–8]. Depending on the relative proportions and compositions of these soil fractions, and on soil texture and total cation exchange capacity (CEC) (which itself depends mainly on soil organic matter and clay content), a soil will have more or less capacity to bind a particular heavy metal species, and will bind it more or less strongly. Accordingly, the distribution of metal cations that are added to the soil among these fractions and others depends on the properties of these fractions as well as on the nature and quantity of the metal added [7,9,10]. In fact, soil properties appear to be more determining than the characteristics of the metal in this respect [11,12]. Given the complexity of soil properties, this makes empirical characterization of their influence on sorption and retention processes essential.

Cd, Cu and Pb are among the potentially most toxic heavy metals, and are present – often together – in numerous polluting spills and in agrochemicals. In a previous work, in which we evaluated the sorption and retention of Cd, Cu and Pb, separately or in mutual competition, by 20 soil horizons, we found that the most useful measure of the capacity of a soil for sorption and retention of these metals was K_r, a novel adimensional parameter that integrates data obtained using a series of different dosage levels [13]. The influence of soil properties on a compound K_r value (a factor score afforded by a principal components analysis of K_r values obtained under various conditions) was modelled by multivariate linear regression (LR). In the present paper we compare the performance of LR of K_r values with that of tree regression [14], which for a series of native soils and different measures and different measures of sorption and retention capacity has been reported to be superior to LR [15].

Data used in previous work [15,16] were from competitive sorption and desorption experiments of Cd, Cr, Cu, Ni, Pb, and Zn. Nevertheless, in this work the models obtained are not only for competitive sorption of Cd, Cu and Pb, but also for individual sorption and retention. Moreover, results from both types of data are compared and discussed.

* Corresponding author. Tel.: +34 986812630; fax: +34 986812556.
E-mail address: emmaf@uvigo.es (E.F. Covelo).

1.1. Sampling

We selected 20 soil horizons as representative of the most widely cropped soil orders in our region that commonly receive inputs containing Cd, Cu and/or Pb. Six samples of each horizon were collected using an Eijkelkamp Model A sampler and were transported in polyethylene bags to the laboratory, where they were air dried, passed through a 2 mm mesh sieve, pooled, and homogenized in a Fritsch Laborette 27 vibratory solid sample homogenizer. The homogenized sample was split into three subsamples, and each of these replicates was in turn subsampled for soil analyses and for sorption/desorption experiments.

2. Variables and methods

2.1. Soil characterization

The following soil properties were determined with a view to evaluating their influence on the capacity of the soil horizons to sorb and retain Cd, Cu and Pb.

Soil pH was determined with a pH meter in 2:1 water/soil suspensions [17]. Particle size distributions were determined following oxidation of organic matter with hydrogen peroxide; the fraction >50 mm was separated by sieving, and the sub-50 mm fraction was characterized as per Day [18]. Organic carbon was quantified by the method of Walkey and Black [19]. CEC and exchangeable cations (Ca^{2+}, Mg^{2+}, K^+ and Na^+; LODs: 5.8, 1, 26, and 4.5 $\mu g\,L^{-1}$, respectively) were determined by extraction with 0.2 M ammonium chloride buffered at the soil pH [20–21], followed by quantification by inductively coupled plasma, atomic emission spectrometry (ICP-OES). Exchangeable acidity was determined using a 1 M KCl replacing solution and titration against 0.1 M NaOH, with phenolphthalein as indicator [22]. Oxides were determined using the method of Mehra and Jackson [23]: samples were shaken in a solution of 0.11 M sodium hydrogen carbonate and 0.27 M sodium citrate, and the Fe, Al and Mn contents of the extracts were determined by inductively coupled plasma, atomic emission spectrometry (ICP-OES) (LODs: 14, 11, and 0,2 $\mu g\,L^{-1}$, respectively). The measured resolution of the system is 0.006 nm at 200 nm. The 80 by 160 mm echelle grating (manufactured by Perkin-Elmer) has 79 lines per mm and a blaze angle of 63.4°.

2.2. Sorption and desorption experiments

Sorption experiments were conducted following the methods of Alberti et al. [24] and Gomes et al. [11], as modified by Harter and Naidu [25]. Non-competitive sorption was evaluated using single-metal sorption solutions of Cu, Cd or Pb nitrates at concentrations of 0.01, 0.03, 0.05, 0.08, 0.1, 0.2, 0.3, 0.4, 0.5, 1, 2 and 3 mmol L^{-1}; and competitive sorption using multi-metal solutions (Cu + Cd + Pb) in which each metal had the same concentration (again 0.01, 0.03, 0.05, 0.08, 0.1, 0.2, 0.3, 0.4, 0.5, 1, 2 or 3 mmol L^{-1}). Both single- and multi-metal solutions also contained 0.01 M $NaNO_3$ as background electrolyte. The heavy metals were used in the form of nitrates because of the high solubility of these salts, and the concentrations were chosen to range from normal values to values representative of severe pollution. Triplicate suspensions of 6 g soil samples in 100 mL of solution in polyethylene tubes were shaken in a rotary shaker for 24 h at 25 °C and then centrifuged at 5000 rpm; the pellet was set aside for use in the desorption stage of the experiment; the supernatant was filtered through Whatman 42 paper; the resulting filtrate was analysed by ICP-OES in a Perkin-Elmer Optima 4300 DV apparatus (USA); and the quantity of each metal that had been sorbed was calculated from the difference between its concentrations in solution before the addition of soil and after equilibration (shaking) with the soil.

Following Madrid and Díaz-Barrientos [26], desorption experiments were conducted using the pellets obtained in the sorption phase of the experiments. The pellets were dried at 45 °C and weighed; each pellet was shaken for 24 h in a polyethylene tube with 100 mL of 0.01 M $NaNO_3$ solution at 25 °C, which was then centrifuged at 5000 rpm; the supernatant was filtered through Whatman 42 paper; the resulting filtrate was analysed by ICP-OES; and the quantity of each metal that had been retained on the soil sample was calculated from the quantity sorbed (determined in the sorption stage of the experiment) and the concentration of the metal in solution following desorption.

For both sorption and retention data, and in both cases for competitive and non-competitive situations, the parameter K_r [13] was calculated as follows. Sorption data were fitted with equations of the forms

$$C_{s,i} = K_{r1}C_{p,i} \tag{I}$$

and

$$C_{p,i} - C_{s,i} = K_{r2}C_{p,i} \tag{II}$$

where $C_{s,i}$ is the amount of metal i that was sorbed per gram of soil and $C_{p,i}$ is the amount of metal i that was potentially sorbable, i.e. the amount in the initial sorption solution (per gram of soil). K_r was then defined as K_{r1} if the coefficient of determination of I was larger than that of Eq. (II), and $1 - K_{r2}$ otherwise. When thus calculated from sorption data, K_r varies from 0 for totally non-sorbent soils to 1 for a soil that completely eliminates metal i from solution. To measure capacity for retention of sorbed metal in desorption experiments, K_r was calculated in the same way as for sorption, except that sorbed metal was replaced by metal retained at equilibrium; in this case K_r is 0 for a soil that completely releases all sorbed metal, and for a soil that releases no metal during the desorption phase of the experiment it adopts the value obtained using the corresponding sorption data.

2.3. Regression analyses

For each metal, sorption and retention K_r data obtained as above were regressed on soil properties using both forward stepwise multiple linear regression as implemented in SPSS version 16.0 for Windows, and tree regression as implemented in STATISTICA 7. In the latter case, because of the inclusion of the three replicate samples of each horizon, the tree selected was not chosen on the basis of its cross-validated R^2 value; instead, from the tree sequence generated we chose the smallest tree with more than two terminal nodes in which the intranodal variances of all terminal nodes were no greater than 0.011. As an indication of the degree to which the relationships between the dependent and each independent variable conformed globally to the structure of each regression tree, we also calculated importance values defined by

$$I(j) = \sum_t \Delta_S(j, t)$$

where $I(j)$ is the importance of variable x_j and $\Delta_S(j,t)$ is the reduction in mean squared error S that would be achieved if node t of the tree were split using x_j [14].

3. Results and discussion

The properties of the soil horizons used have been published elsewhere, together with the corresponding K_r values obtained from the experiments described above [13]; these data are summarized in Table 1. The K_r values were regressed on soil properties using both tree regression and LR, and Table 2 compares the corresponding goodness-of-fit values (coefficients of determination R^2, calculated as usual as $\sum_s(\hat{y}_s - \bar{y})^2 / \sum_s(y_s - \bar{y})^2$, where $\bar{y}$ is the

Table 1
Descriptive statistics of the properties of the 20 horizons studied, and of their K_r values for competitive and non-competitive sorption and retention of Pb, Cu, and Cd.

	Minimum	Maximum	Mean	S.D.
pH	3.660	8.390	5.466	1.183
Organic matter (mg kg^{-1})	0.020	96.920	36.831	28.448
Al oxides (mg kg^{-1})	1.580	19.790	8.245	4.939
Fe oxides (mg kg^{-1})	2.230	65.490	19.528	16.814
Mn oxides (mg kg^{-1})	0.000	2.220	0.323	0.509
CEC$_e$ (cmol$_{(+)}$ kg^{-1})	0.790	156.810	17.638	35.850
Sand (%)	19.500	82.020	56.971	16.470
Clay (%)	6.400	66.600	19.785	1.837
Non-competitive cadmium sorption	0.074	0.986	0.370	0.316
Non-competitive copper sorption	0.046	0.999	0.474	0.331
Non-competitive lead sorption	0.079	1.000	0.546	0.316
Competitive cadmium sorption	0.033	0.621	0.231	0.174
Competitive copper sorption	0.028	0.981	0.392	0.283
Competitive lead sorption	0.104	0.996	0.530	0.259
Total sorption when added together	0.053	0.884	0.400	0.236
Non-competitive cadmium retention	0.000	0.984	0.284	0.352
Non-competitive copper retention	0.012	0.998	0.445	0.345
Non-competitive lead retention	0.034	1.000	0.518	0.334
Competitive cadmium retention	0.000	0.600	0.160	0.184
Competitive copper retention	0.000	0.981	0.365	0.294
Competitive lead retention	0.075	0.996	0.506	0.271
Total retention when added together	0.023	0.877	0.363	0.247

Table 2
Goodness-of-fit (R^2) of the models fitted by tree regression and by forward stepwise multiple linear regression (LR) to the sorption and retention K_r and soil property data.

	Sorption		Retention	
	R^2 LR	R^2 tree	R^2 LR	R^2 tree
Non-competitive cadmium	0.94	0.98	0.95	0.97
Non-competitive copper	0.99	0.99	0.90	0.99
Non-competitive lead	0.85	0.95	0.85	0.95
Competitive cadmium	0.96	0.97	0.97	0.98
Competitive copper	0.95	0.98	0.94	0.99
Competitive lead	0.91	0.98	0.90	0.98
Total when added together	0.95	0.97	0.95	0.98

mean of the observed values y_s, and $\hat{y}_s$ the value predicted by the regression equation). Tree regression always achieved greater R^2 than LR, the difference being most marked for K_rs for lead, for which LR R^2 was lowest.

Tables 3 and 4 list the LR regression coefficients for standardized soil variables. The three largest in absolute value always included pH and CEC (except that the coefficient of CEC was not significant in the case of competitive sorption or retention of cadmium), with pH in top position in all cases except for the non-competitive sorption or retention of lead. However, for any given heavy metal, Mn oxides' content only had a significant influence for either single-metal experiments (Cd) or multi-metal experiments (Cu and Pb),

Table 3
Coefficients of multiple linear regression of heavy metal sorption on standardized soil variables (only significant coefficients are shown).

	Non-competitive			Competitive			Total
	Cd	Cu	Pb	Cd	Cu	Pb	
pH	0.82	0.41	0.29	0.83	0.87	0.73	0.82
Organic matter	0.07	0.20	0.29	−0.18			
Al oxides	−0.11			0.08	0.14		0.08
Fe oxides	0.18	0.20	0.23		0.09		
Mn oxides	−0.30				−0.48	−0.41	−0.36
CEC$_e$	0.34	0.29	0.33		0.50	0.55	0.42
Sand		−0.26	−0.27	0.25		−0.50	−0.20
Clay				0.58	0.10	−0.32	

Table 4
Coefficients of multiple linear regression of heavy metal retention on standardized soil variables (only significant coefficients are shown).

	Non-competitive			Competitive			Total
	Cd	Cu	Pb	Cd	Cu	Pb	
pH	0.89	0.43	0.29	0.86	0.94	0.75	0.87
Organic matter		0.20	0.29	−0.05		0.10	
Al oxides	−0.09				0.14		0.11
Fe oxides	0.15	0.20	0.24		0.12		0.07
Mn oxides	−0.33				−0.62	−0.19	−0.38
CEC$_e$	0.33	0.28	0.34		0.59	0.28	0.38
Sand		−0.25	−0.26	0.37		−0.19	
Clay				0.59			0.13

Table 5
Importance values in regression trees for heavy metal sorption.

	Non-competitive			Competitive			Total
	Cd	Cu	Pb	Cd	Cu	Pb	
pH	0.960	0.963	0.907	0.979	0.988	0.949	0.995
Organic matter	0.190	0.455	0.555	0.140	0.318	0.391	0.317
Al oxides	0.217	0.282	0.337	0.202	0.217	0.270	0.269
Fe oxides	0.393	0.408	0.415	0.359	0.342	0.352	0.361
Mn oxides	0.871	0.820	0.763	0.782	0.834	0.870	0.829
CEC$_e$	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Sand	0.376	0.565	0.582	0.658	0.488	0.493	0.519
Clay	0.381	0.455	0.466	0.648	0.452	0.413	0.478

never for both. Furthermore, either the LR coefficient of pH was appreciably greater than that of any other variable (the most usual situation), or all the significant coefficients were quite similar (the case of non-competitive sorption and retention of Cu or Pb).

Tables 5 and 6 list tree regression importance values. The three variables with greatest importance values were in all cases, in increasing order, Mn oxides' content, pH, and CEC, and these importance values were in all cases appreciably larger than those of any other variable. CEC is a quantitative measure of the ability of a soil to exchange cations with the soil solution and is expressed in terms of cmols$_{(+)}$ kg^{-1} of soil.

In what follows we discuss these results in greater detail for each metal in turn.

3.1. Cadmium

By far the largest LR coefficient for the non-competitive sorption and retention of cadmium was that of pH, followed a long way behind by CEC$_e$ and Mn oxides' content. Sorption and retention, as measured by K_r, increased with pH and CEC$_e$ (and also with Fe oxides' content and, in the case of sorption, organic matter content); and decreased with increasing Mn and Al oxides' contents. For competitive sorption and retention, pH again had the largest LR coefficient, but the other major predictor of K_r was now clay content, followed by sand content; sorption and retention increased with all these variables (and also to a small extent with Al oxides'

Table 6
Importance values in regression trees for heavy metal retention.

	Non-competitive			Competitive			Total
	Cd	Cu	Pb	Cd	Cu	Pb	
pH	0.962	0.969	0.912	0.975	0.990	0.965	0.986
Organic Matter	0.170	0.435	0.553	0.119	0.317	0.390	0.304
Al oxides	0.237	0.264	0.323	0.157	0.226	0.272	0.251
Fe oxides	0.398	0.403	0.405	0.335	0.346	0.357	0.361
Mn oxides	0.866	0.823	0.779	0.807	0.844	0.877	0.824
CEC$_e$	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Sand	0.350	0.550	0.579	0.547	0.477	0.492	0.483
Clay	0.366	0.441	0.456	0.529	0.453	0.409	0.456

Tree 13 graph for Individual Cadmium Sorption
Num. of non-terminal nodes: 3, Num. of terminal nodes: 4

Tree 10 graph for Competitive Cadmium Sorption
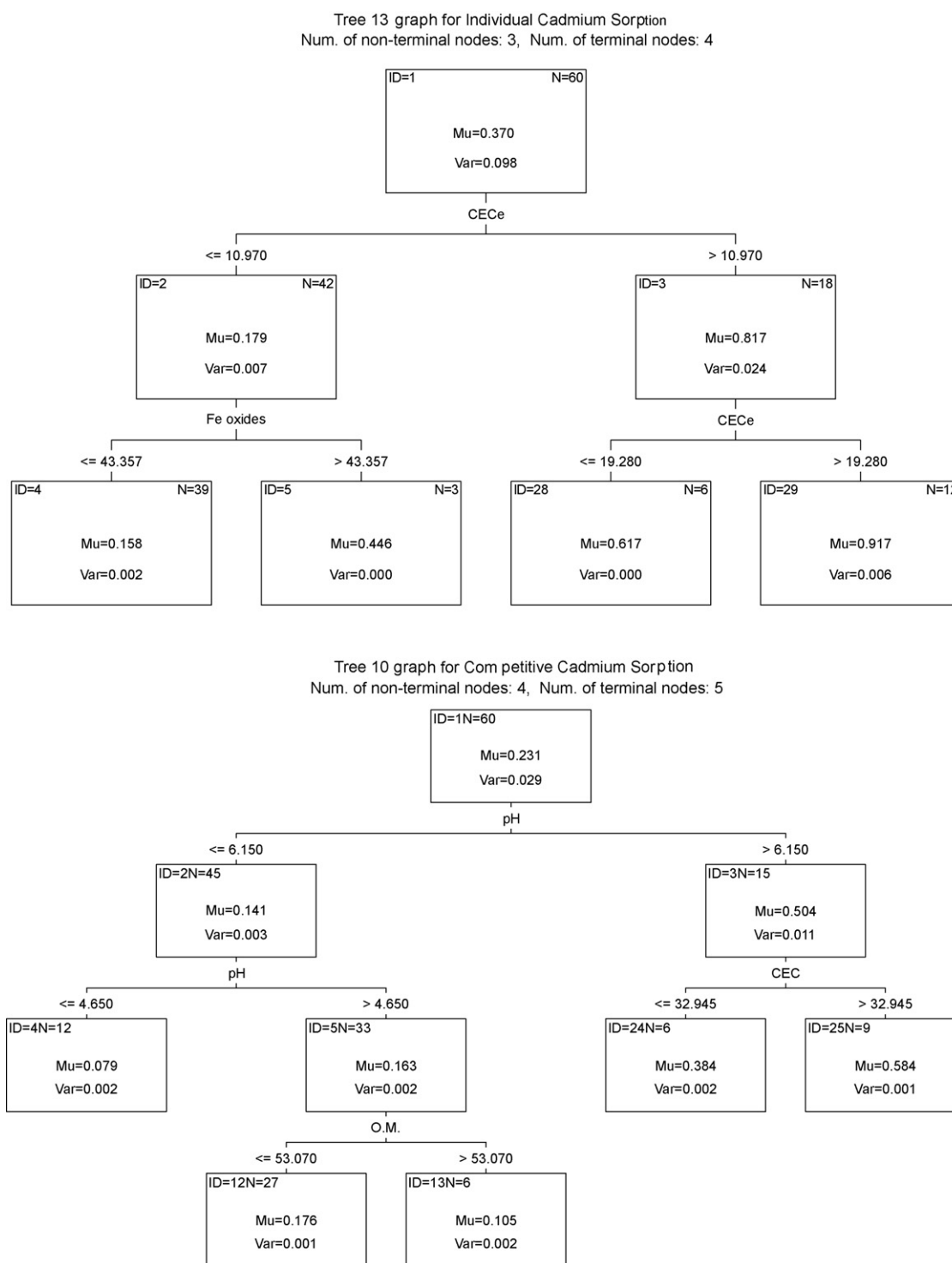Num. of non-terminal nodes: 4, Num. of terminal nodes: 5

Fig. 1. CART binary regression tree for sorption of cadmium by 20 cropped soil horizons. Top, non-competitive; bottom, competitive.

content in the case of sorption), and decreased slightly with increasing organic matter content.

Whereas the LR equations for non-competitive sorption and retention of cadmium are more complex than those for competitive sorption and retention, in the sense that they have more significant variables, the reverse is the case for the regression trees (Figs. 1 and 2), the trees for the non-competitive process being less complex than those obtained for the competitive process. In fact, the non-competitive trees involve only CEC and either Fe oxides' content (in the case of sorption) or Mn oxides' content (reten-

tion). The tree for non-competitive sorption predicts a K_r value of around 0.9, indicative of high sorption capacity, for horizons with $CEC_e > 19.28$ cmol kg^{-1}, and a value of only about 0.15, indicative of low sorption capacity, for horizons with $CEC_e \leq 10.97$ cmol kg^{-1} and Fe oxides' contents lower than 43.4 g kg^{-1}. The major bifurcation of the tree for non-competitive retention is also determined by CEC, and both the splits on this variable have the same splitting values as in the sorption tree, but least retention capacity ($K_r \cong 0.05$) is predicted for horizons with both $CEC \leq 10.97$ cmol kg^{-1} and Mn oxides' content ≤ 0.35 g kg^{-1}. Strikingly, it is low Mn oxides' content

Tree 12 graph for Individual Cadmium retention
Num. of non-terminal nodes: 3, Num. of terminal nodes: 4

Tree 11 graph for Competitive Cadmium retention
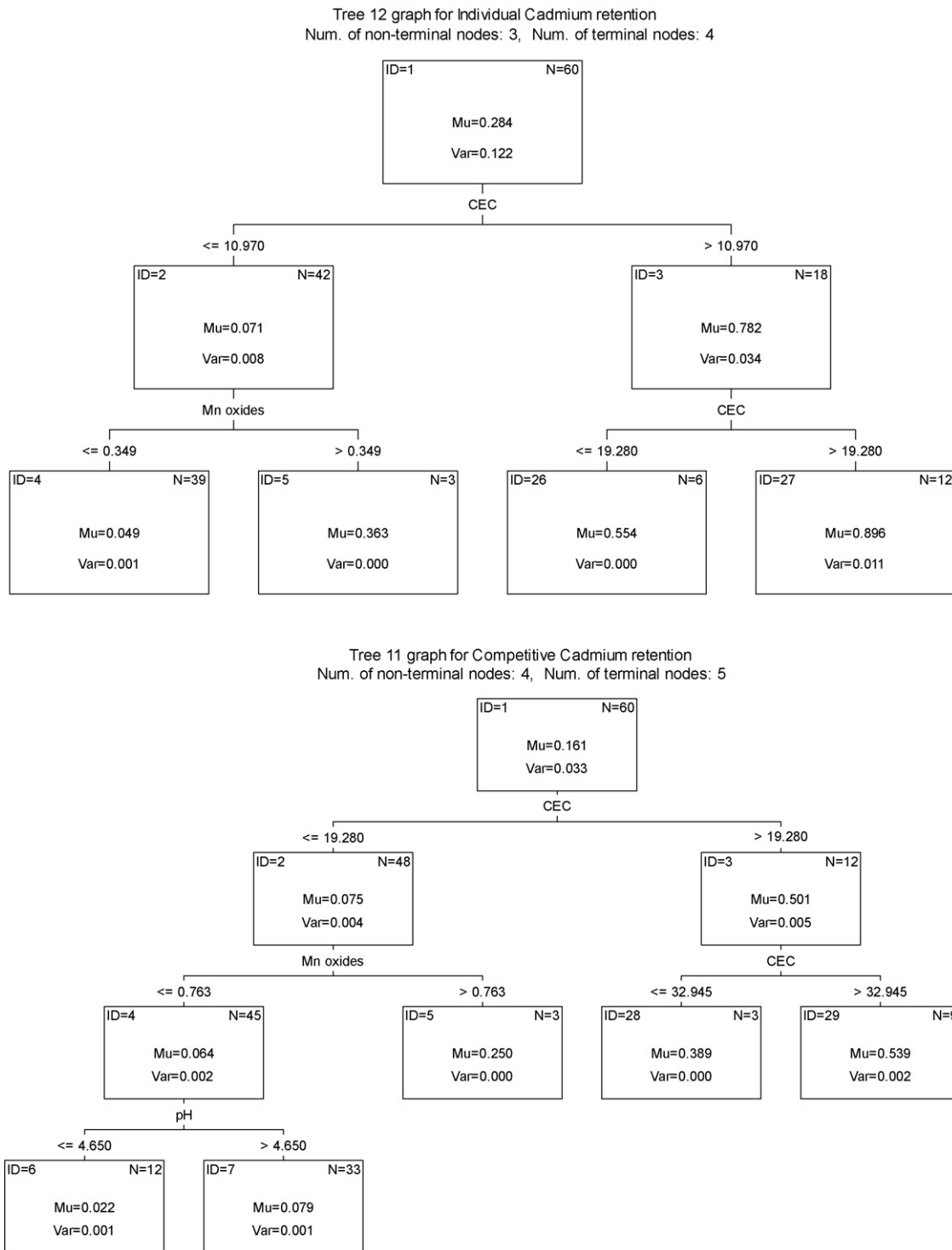Num. of non-terminal nodes: 4, Num. of terminal nodes: 5

Fig. 2. CART binary regression tree for retention of cadmium by 20 cropped soil horizons. Top, non-competitive; bottom, competitive.

that indicates lower Cd retention capacity, whereas in the LR model the coefficient of this variable has a negative sign (this is discussed further below, in the general remarks and Section 4).

The major bifurcation of the regression tree for competitive sorption of Cd is determined by pH, and other splits by pH, CEC, and organic matter content. In this case, even the greatest predicted sorption capacity, predicted for horizons with pH >6.15 and $CEC_e > 32.9\,\mathrm{cmol\,kg^{-1}}$, is only moderate ($K_r \cong 0.6$). Least capacity ($K_r \cong 0.08$) is predicted for horizons with pH ≤4.65; and in keeping with the negative LR coefficient of organic matter, horizons with

pH in the range 4.65–6.15 are predicted to sorb more or less cadmium depending on whether their organic matter content is less than or greater than $53.1\,\mathrm{g\,kg^{-1}}$. Since organic matter is generally one of the most important immobilizers of heavy metals in soils, and since the LR coefficient of organic matter for non-competitive sorption of Cd is positive, this latter finding suggests that under these acidic conditions competition with Cu and Pb is particularly unfavourable for the binding of Cd to organic matter.

The regression tree for competitive retention of Cd, like that for competitive sorption, has five internally homoge-
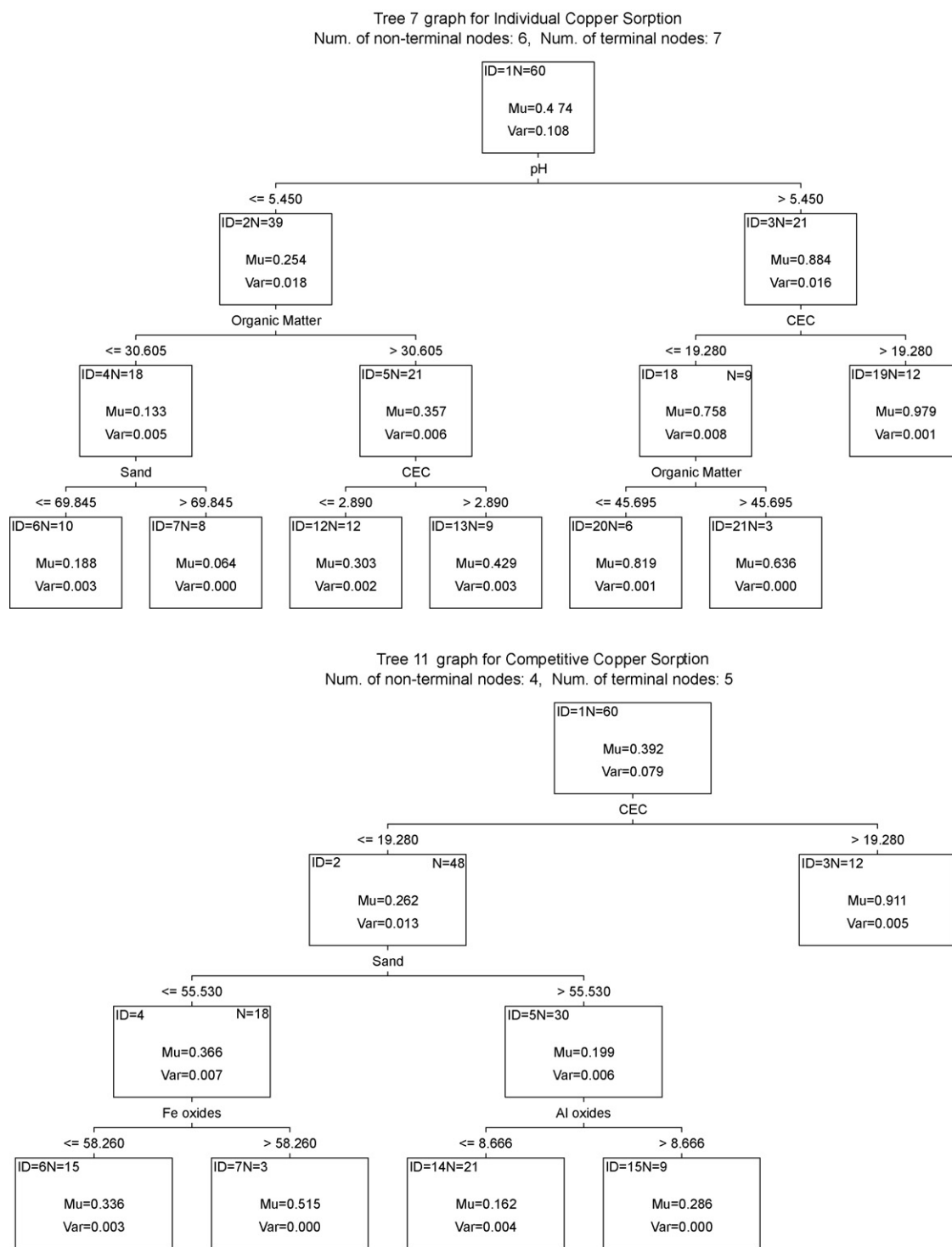
Tree 7 graph for Individual Copper Sorption
Num. of non-terminal nodes: 6, Num. of terminal nodes: 7

Tree 11 graph for Competitive Copper Sorption
Num. of non-terminal nodes: 4, Num. of terminal nodes: 5

Fig. 3. CART binary regression tree for sorption of copper by 20 cropped soil horizons. Top, non-competitive; bottom, competitive.

neous terminal nodes. Greatest retention capacity (again no more than moderate; $K_r \cong 0.54$) is predicted for horizons with $CEC_e > 32.95$ cmol kg^{-1}, while a $K_r < 0.1$ is predicted for all horizons with $CEC_e \leq 19.28$ cmol kg^{-1} and Mn oxides' content ≤ 0.76 g kg^{-1}, and a K_r of 0.02 if additionally pH ≤ 4.65. Once more, it is low Mn oxides' content that is predicted as indicating low Cd retention capacity. Note also that of the four trees for cadmium, only the tree for competitive retention has as its splitting variables all three variables of greatest importance: CEC_e, pH and Mn oxides' content.

3.2. Copper

According to the LR equations for the non-competitive sorption and retention of copper, the corresponding K_r values increase with pH, CEC_e, Fe oxides' content and organic matter's content, and decrease with increasing sand content; the magnitudes of all these effects are quite similar, except for that of pH, which is rather large. For competitive sorption and retention, K_r increases with pH and CEC_e (and also to some extent with Al and Fe oxides' contents and, in the case of sorption,
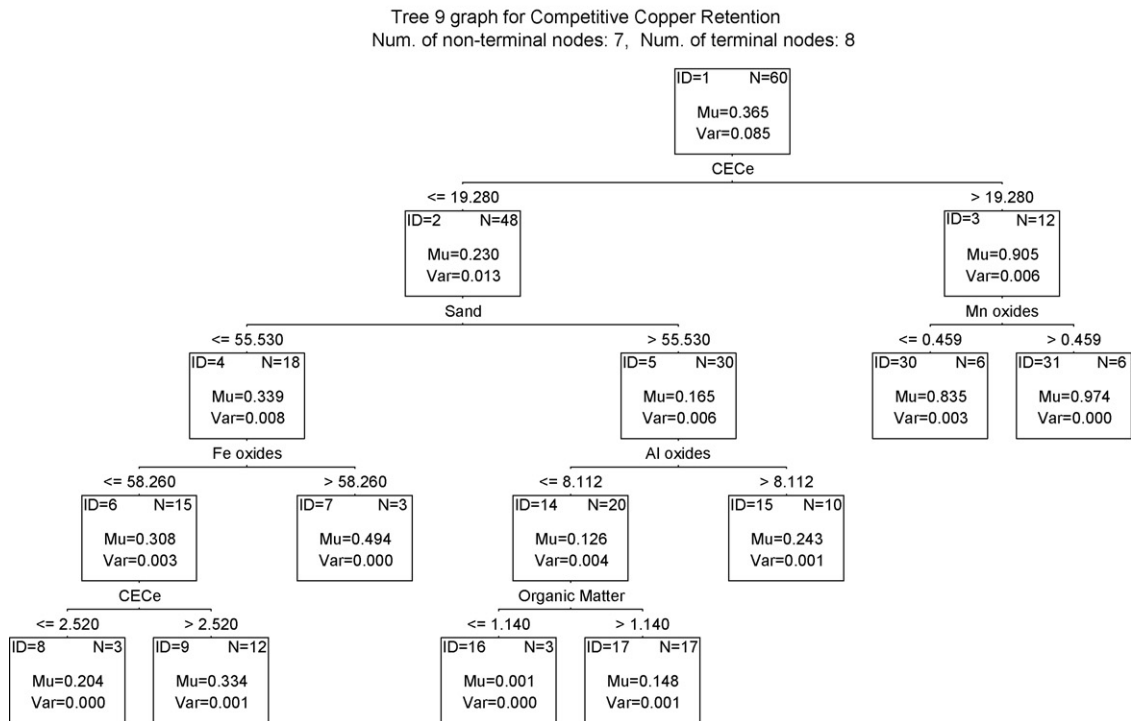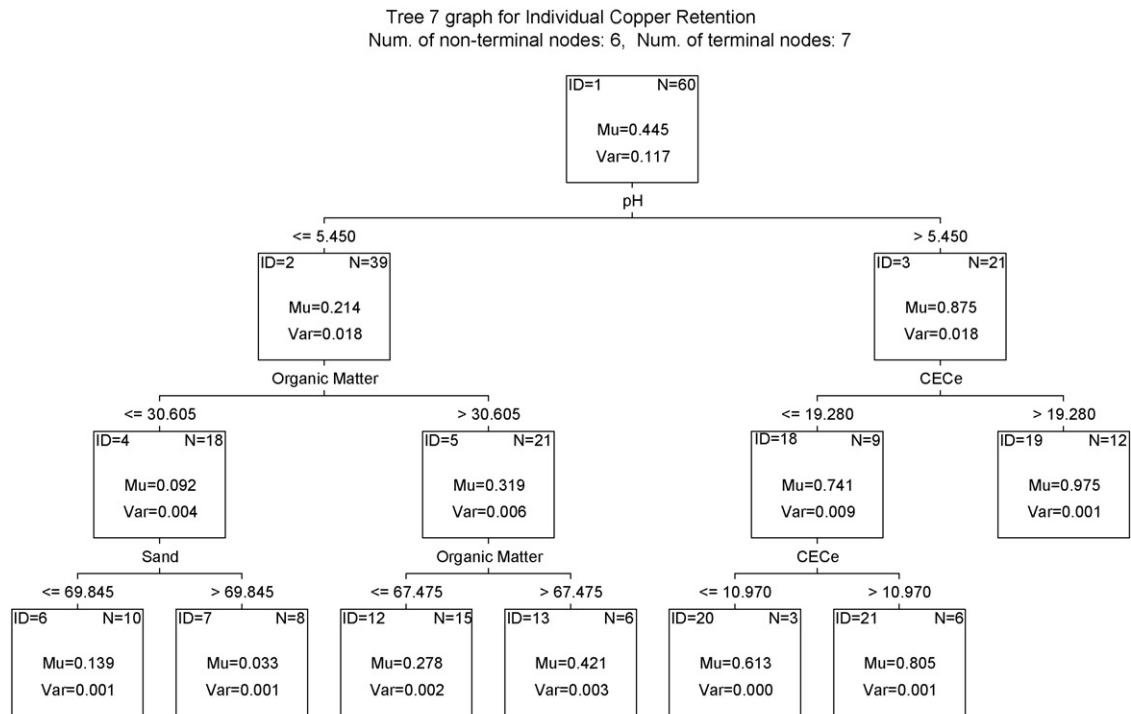
Tree 7 graph for Individual Copper Retention
Num. of non-terminal nodes: 6, Num. of terminal nodes: 7

ID=1 N=60
Mu=0.445
Var=0.117
pH

<= 5.450

ID=2 N=39
Mu=0.214
Var=0.018
Organic Matter

> 5.450

ID=3 N=21
Mu=0.875
Var=0.018
CECe

<= 30.605

ID=4 N=18
Mu=0.092
Var=0.004
Sand

> 30.605

ID=5 N=21
Mu=0.319
Var=0.006
Organic Matter

<= 19.280

ID=18 N=9
Mu=0.741
Var=0.009
CECe

> 19.280

ID=19 N=12
Mu=0.975
Var=0.001

<= 69.845

ID=6 N=10
Mu=0.139
Var=0.001

> 69.845

ID=7 N=8
Mu=0.033
Var=0.001

<= 67.475

ID=12 N=15
Mu=0.278
Var=0.002

> 67.475

ID=13 N=6
Mu=0.421
Var=0.003

<= 10.970

ID=20 N=3
Mu=0.613
Var=0.000

> 10.970

ID=21 N=6
Mu=0.805
Var=0.001

Tree 9 graph for Competitive Copper Retention
Num. of non-terminal nodes: 7, Num. of terminal nodes: 8

ID=1 N=60
Mu=0.365
Var=0.085
CECe

<= 19.280

ID=2 N=48
Mu=0.230
Var=0.013
Sand

> 19.280

ID=3 N=12
Mu=0.905
Var=0.006
Mn oxides

<= 55.530

ID=4 N=18
Mu=0.339
Var=0.008
Fe oxides

> 55.530

ID=5 N=30
Mu=0.165
Var=0.006
Al oxides

<= 0.459

ID=30 N=6
Mu=0.835
Var=0.003

> 0.459

ID=31 N=6
Mu=0.974
Var=0.000

<= 58.260

ID=6 N=15
Mu=0.308
Var=0.003
CECe

> 58.260

ID=7 N=3
Mu=0.494
Var=0.000

<= 8.112

ID=14 N=20
Mu=0.126
Var=0.004
Organic Matter

> 8.112

ID=15 N=10
Mu=0.243
Var=0.001

<= 2.520

ID=8 N=3
Mu=0.204
Var=0.000

> 2.520

ID=9 N=12
Mu=0.334
Var=0.001

<= 1.140

ID=16 N=3
Mu=0.001
Var=0.000

> 1.140

ID=17 N=17
Mu=0.148
Var=0.001

Fig. 4. CART binary regression tree for retention of copper by 20 cropped soil horizons. Top, non-competitive; bottom, competitive.

clay content), and decreases with increasing Mn oxides' content.

The regression trees for non-competitive sorption and retention of copper (Figs. 3 and 4) are morphologically identical and have an almost identical distribution of splitting variables and splitting values, the only differences being that although the third-generation nodes 5 and 18 both split with respect to CEC_e in one tree and organic matter in the other, it is node 5 that splits with respect to CEC_e in the sorption tree, and node 18 in the retention tree. The splitting variables of both trees are the two of greatest importance value (CEC_e and pH), together with sand content and organic matter content; the affinity of organic matter for Cu is well documented [27–29]. In both trees, greatest Cu-binding capacity ($K_r \cong 0.98$) is predicted for horizons with pH >5.45 and $CEC_e > 19.28$ cmol kg^{-1}; in fact, the similarity between the sorption and retention K_r values predicted for these high-Cu-capacity horizons suggests that almost all their sorbed Cu is retained, i.e. binding is very strong and irreversible. Least sorption and retention capacity ($K_r < 0.1$) is predicted

for horizons with pH $\leq$5.45, organic matter content $\leq$30.6 g kg^{-1}, and sand content greater than 70%, i.e. less than 30% of more finely divided material with greater metal-binding capacity.

The regression trees for competitive sorption and retention of Cu are respectively the simplest and the most complex of the four Cu trees, with respectively five and eight terminal nodes. The former is nevertheless identical, as regards splitting variables and values, with a subtree of the latter. In both these trees, the major split separates horizons with CEC$_e$ $\leq$ 19.28 cmol kg^{-1} from those with greater values, for which average K_r values of around 0.9 are predicted; low-CEC$_e$ horizons are split with respect to sand content (higher sand content indicating lower Cu-binding capacity); low-CEC$_e$, low-sand horizons are split with respect to Fe oxides' content; and low-CEC$_e$, high-sand horizons are split with respect to Al oxides' content, the lowest Cu-binding capacity (K_r < 0.2) being predicted for horizons with CEC$_e$ $\leq$ 19.28 cmol kg^{-1}, sand content >55% and Al oxides' content $\leq$8.7 g kg^{-1}. Highest retention capacity is predicted for horizons that, besides having CEC$_e$ > 19.28 cmol kg^{-1}, have Mn oxides' contents >0.46 g kg^{-1}; and lowest retention capacity (K_r $\cong$ 0.001) for those that, in addition to the above-noted requisites for both low sorption and low retention, have organic matter contents $\leq$1.14 g kg^{-1}. Note that, as in the case of cadmium, there is an apparent contradiction between the influence of Mn oxides' content predicted by LR and that predicted by the regression tree.

3.3. Lead

The LR models for non-competitive Pb sorption and Pb retention capacities are virtually identical, with regression coefficients of 0.33 or 0.34 for CEC$_e$, 0.29 for pH and organic matter, 0.23 or 0.24 for Fe oxides, and −0.26 or −0.27 for sand content (Tables 3 and 4). The presence of other metals disrupts this identity somewhat. For both sorption and retention, competitive Pb binding increases with both CEC$_e$ and pH (though sorption increases about twice as fast with CEC$_e$ as does retention), and decreases with increasing sand and Mn oxides' contents (again about twice as fast in the case of sorption as in the case of retention); however, clay content reduces sorption but has no significant effect on retention, while organic matter content aids retention but has no significant effect on sorption.

The splitting variables of the regression trees for non-competitive sorption and retention are those with the largest coefficients in the corresponding LR models, CEC$_e$, pH and organic matter content, although the main split is in both cases with respect to pH (Figs. 5 and 6). Horizons with pH >5.45 are predicted to have the highest Pb-binding capacities (average K_r > 0.9), and, in particular, horizons with both pH >5.45 and CEC$_e$ greater than about 11 cmol kg^{-1} are predicted to have Pb sorption capacities of around 0.96. Among horizons with pH $\leq$5.45, both trees predict that those with relatively low organic matter contents (less than about 37 g kg^{-1} for sorption, or 30.6 g kg^{-1} for retention) have least Pb-binding capacity (K_r < 0.25).

CEC$_e$ and organic matter content, but not pH, are likewise among the splitting variables of the regression trees for competitive sorption and retention of Pb, alongside sand content and Fe oxides' contents in the former case and Al oxides' content in the latter. Greatest sorption and retention capacity (K_r $\cong$ 0.96) is predicted for horizons with CEC$_e$ > 19.28 cmol kg^{-1}, while least sorption capacity (K_r $\cong$ 0.16) is predicted for horizons with CEC$_e$ < 11 cmol kg^{-1} and sand content greater than about 73%, and least retention capacity (K_r $\cong$ 0.1) for those with CEC$_e$ < 11 cmol kg^{-1}, organic matter content less than about 44 g kg^{-1}, and Al oxides' content less than about 2 g kg^{-1}.

3.4. Total heavy metal-binding capacity

In the LR models of total heavy metal-binding capacity, as in most of the other cases, the largest coefficient is that of pH, followed

by CEC$_e$ and, with negative sign, Mn oxides' content. Al oxides' content has a positive influence (as do Fe oxides' and clay contents for retention), while sand content has a negative influence on sorption. In keeping with this, in both the regression trees the major split is on pH (Fig. 7), and both also use CEC$_e$, sand and Al oxides' contents, while Mn oxides' content is used only to split a third-generation node in the retention tree. Note, however, that the splits on both Mn oxides and Al oxides are of opposite sign to the corresponding LR coefficients.

3.5. The relationship between the LR and tree regression results

The variable involved in the major bifurcation of each regression tree was always either CEC$_e$ or pH: seven trees started with a split on CEC$_e$ and the other seven with a split on pH, with a tendency towards pH for non-competitive sorption or retention trees and CEC$_e$ for their competitive counterparts. These two variables were thus identified by the tree regressions as of global significance, which both supports the relevance of their inclusion in the corresponding LR models, and is in keeping with their generally having the largest LR regression coefficients (pH especially). By contrast, as noted above, there is an apparent contradiction between the LR and tree regression results in regard to Mn oxides' content: whereas it is low Mn oxides' content that is predicted by certain tree regressions to indicate lower metal-binding capacities, the coefficients of this variable in the corresponding LR models have negative signs. In Fig. 7 there is a similar apparent contradiction regarding Al oxides' content. This highlights the fact that whereas LR coefficients represent global trends, regression tree bifurcations indicate behaviour at the scale corresponding to their level in the tree: whereas LR coefficients attempt to characterize with a single number the relationship that holds between the dependent variable and the independent variable in the whole of the space defined by the independent variables, the scope of a regression tree splitting variable is delimited by the splits that have preceded it on the tree. However, when local trends contradict global trends, the possibility that the model is overfitted should also be borne in mind.

CEC$_e$ and pH were also, in all trees, the variables with greatest importance values. However, whereas CEC$_e$, with an importance value of 1 in all trees, was used as a splitting variable a total of 24 times, and at all tree levels, pH (mean importance 0.964) was used as a splitting variable just three times in addition to the seven top-level splits mentioned above. Furthermore, while Mn oxides' content, which always had the third largest importance value (mean 0.828), was only used as a splitting variable four times in all 14 trees (splitting three second-generation nodes and one third-generation node in retention trees), sand (mean importance 0.511) was used twice as often (splitting 5 second-generation and 3 third-generation nodes), and organic matter (mean importance 0.332) was used nine times (splitting 5 second-, 5 third- and 1 fourth-generation nodes). That the importance value of a variable does not reflect the number of times it is used for splitting nodes is as expected, because it was situations such as that of Mn oxides that originally prompted the definition of importance values [14]; but nor does the global importance represented by importance values necessarily correspond to the global importance reflected by the size of LR coefficients. In these trees, the LR coefficient of pH is generally much larger than that of CEC$_e$, while its importance value is always smaller; and the difference in sign between the LR coefficients of Mn oxides and the direction of the splits it performs indicates that although splitting on this variable would generally eliminate a large amount of variance in K_r, the split would take one direction at some nodes and the opposite direction at others.
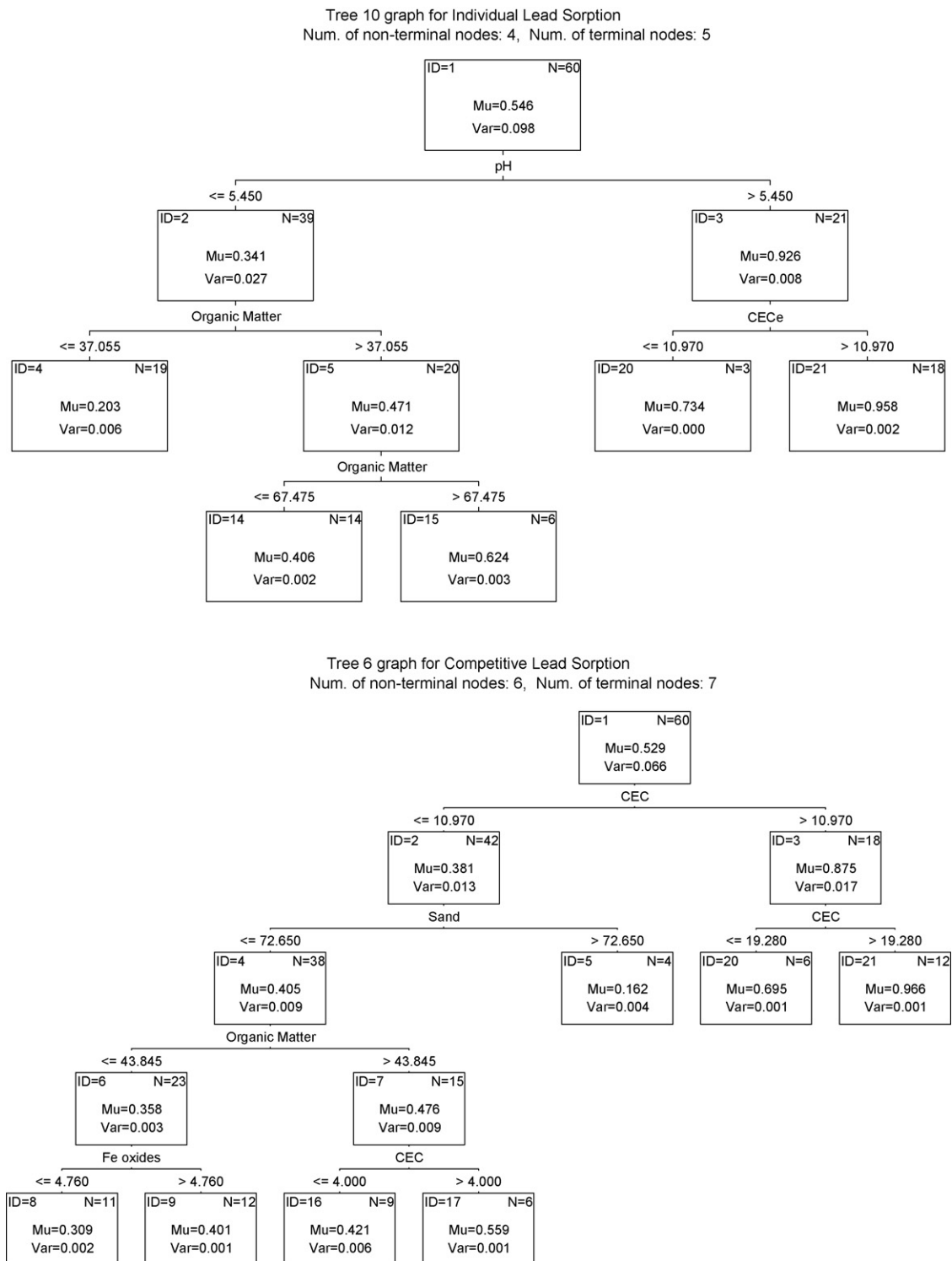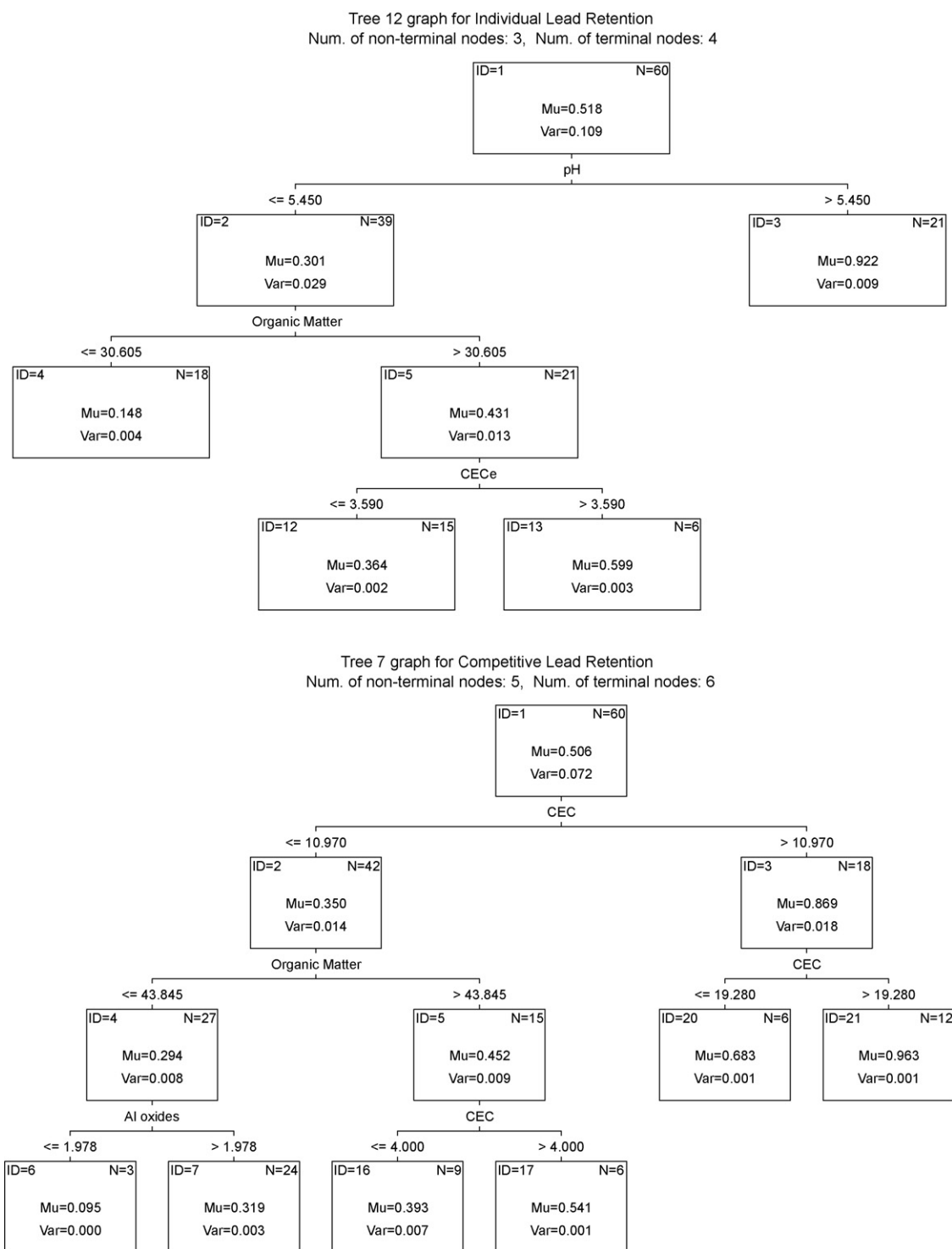
Fig. 5. CART binary regression tree for sorption of lead by 20 cropped soil horizons. Top, non-competitive; bottom, competitive.

3.6. Tree regression predictions of moderate-to-good metal-binding capacity

The mean K_r values listed in Table 1 show that, on average, the sorption or retention of these metals by the soil horizons studied increased in the order Cd < Cu < Pb. The least of the maximum K_r values listed in Table 1 is 0.6, corresponding to competitive retention of cadmium. Here we identify, for each experimental situation, the soil characteristics that are predicted by tree regression to be asso-

ciated with mean K_r values greater than 0.6 in all the corresponding terminal nodes.

For non-competitive sorption of Cd: $CEC_e > 10.97\,\mathrm{cmol\,kg^{-1}}$.
For non-competitive retention of Cd: $CEC_e > 19.28\,\mathrm{cmol\,kg^{-1}}$.
For competitive sorption and retention of Cd, there are no terminal nodes with mean $K_r > 0.6$.
For non-competitive sorption and retention of Cu: pH > 5.45.

Fig. 6. CART binary regression tree for retention of lead by 20 cropped soil horizons. Top, non-competitive; bottom, competitive.

For competitive sorption and retention of Cu: $CEC_e > 19.28$ cmol kg^{-1}.

For non-competitive sorption of Pb: pH > 5.45 or organic matter content > 67.5 g kg^{-1}.

For non-competitive retention of Pb: pH > 5.45 (although the combination of organic matter content > 30.6 g kg^{-1} and CEC_e > 3.6 cmol kg^{-1} may be considered as borderline conditions).

For competitive sorption and retention of Pb: CEC_e > 10.97 cmol kg^{-1}.

For total sorption and retention of Cd, Cu and Pb when added together: pH > 6.15, Al oxides' content ≤ 9.8 g kg^{-1}.

Whether K_r > 0.6 or not (i.e. whether or not more than 60% of added metal was sorbed or retained) thus depended only on CEC_e, pH and organic matter content, and in all three cases, as the values of these parameters increased, Pb was among the metals for which that value of K_r was first reached. Non-competitive sorption or retention of >60% of added Pb or Cu was predicted by pH >5.45 (or by
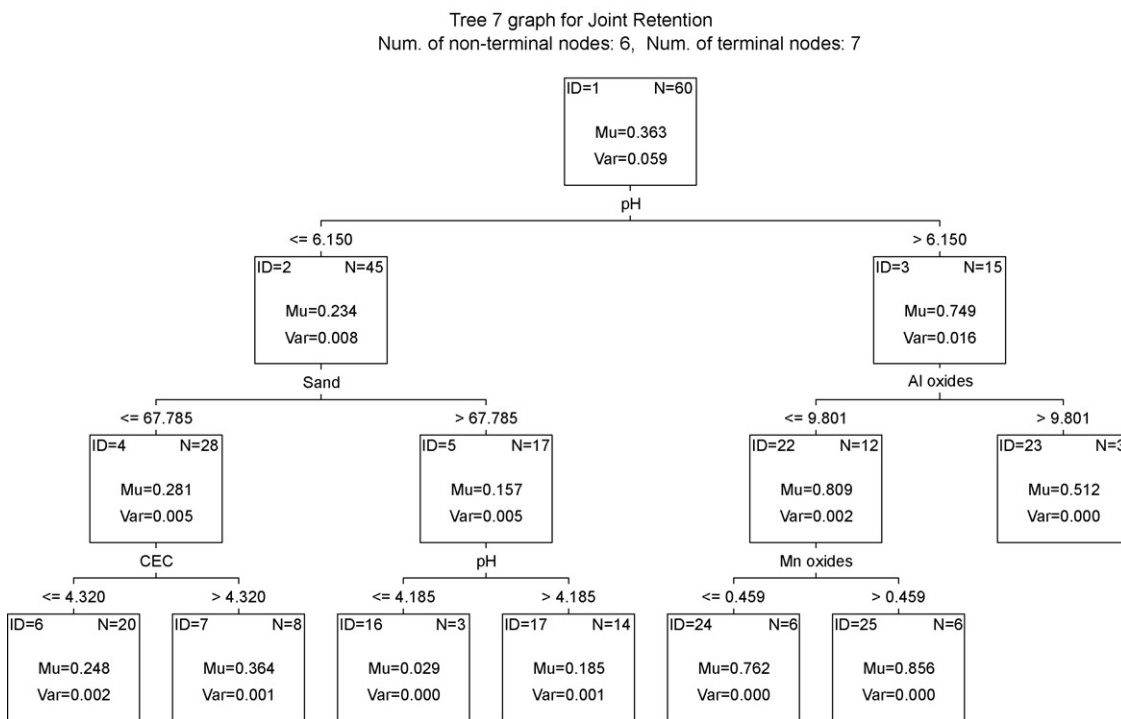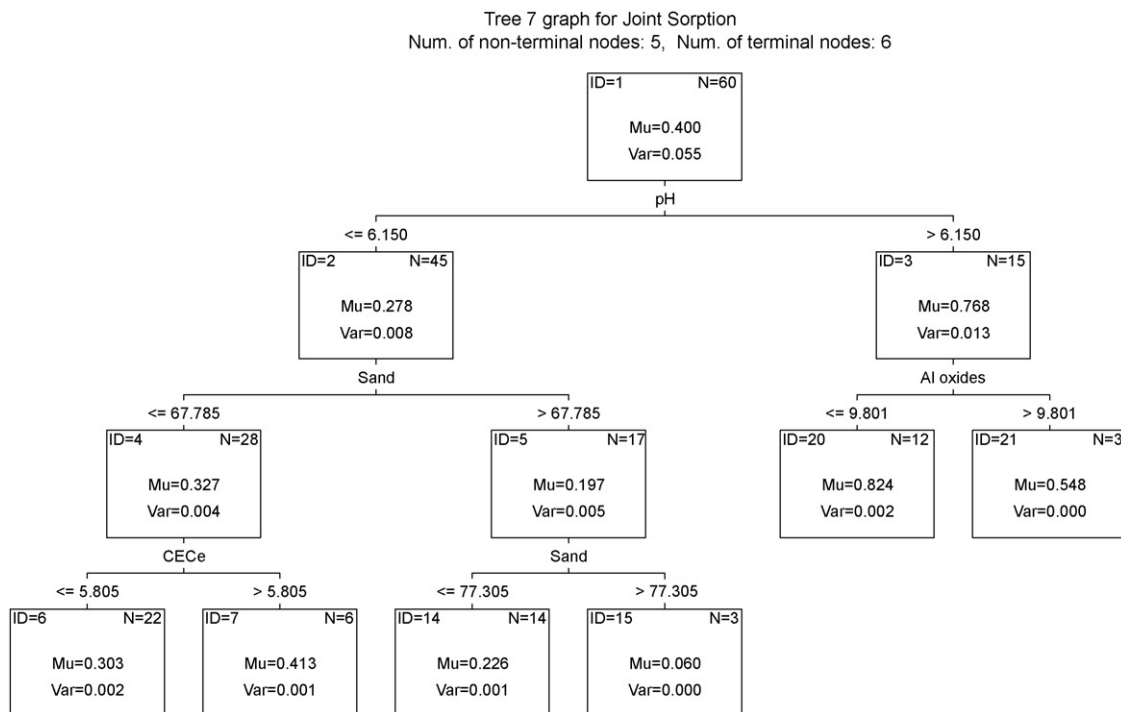
Fig. 7. CART binary regression tree for total sorption (top) and retention (bottom) of Cd, Cu, and Pb by 20 cropped soil horizons.

organic matter content >67.5 g kg^{-1} in the case of Pb), and competitive sorption or retention of >60% of added Pb or Cu was predicted by a high enough CEC$_e$ (>11 cmol kg^{-1} for Pb, >19.3 cmol kg^{-1} for Cu). Non-competitive sorption or retention of >60% of added Cd was also predicted by a high enough CEC$_e$, but Cd sorption and retention were predicted never to attain this level in the presence of Cu and Pb.

4. Conclusions

These horizons sorb and retain cadmium, copper and lead in the order Cd < Cu < Pb. Sorption and retention of Cd is depressed by the presence of Cu and Pb. The soil parameters that are most associated, overall, with differences in Cd, Cu and Pb sorption and retention are CEC$_e$, pH and Mn oxides' content, although the sorption or reten-

tion of >60% of added heavy metal is dependent almost exclusively on CEC_e and pH. Tree regression, which can take into account variation on both global and local scales, can afford better-fitting models than linear regression, which only reflects global tendencies; but for coherent interpretability of tree regression results it is just as important to avoid overfitting as in the case of linear regression.

Acknowledgements

References

[1] D.L. Sparks, A.M. Schidegger, D.G. Strawn, K.G. Scheckel, in: D.L. Sparks, T.L. Grundl (Eds.), Mineral–Water Interfacial Reactions: Kinetics, Am. Chem. Soc., Washington, DC, USA, 1999, p. 108.
[2] R. Apak, in: A. Hubbard (Ed.), Encyclopaedia of Surface and Colloid Science, Dekker, New York, 2002, p. 385.
[3] F.A. Vega, E.F. Covelo, M.L. Andrade, J. Colloid Interface Sci. 298 (2006) 582–592.
[4] G. Sposito, C.S. LeVesque, J.P. LeClaire, A.C. Chang, Soil Sci. Soc. Am. J 47 (1983) 898–902.
[5] S.P. McGrath, J. Cegarra, J. Soil Sci. 43 (1992) 313–321.
[6] D.G. Kinniburgh, W.H. Van Riemsdijk, L.K. Koopal, M. Borkovec, M.F. Benedetti, M.J. Avena, Colloids Surf. A: Physicochem. Eng. Aspects 151 (1–2) (1999) 147–166.
[7] F.X. Han, W.L. Kingery, H.M. Selim, in: I.K. Iskander, M.B. Kirkham (Eds.), Trace Elements in Soil: Bioavailability, Fluxes and Transfer, CRC Press, Boca Raton, FL, 2001, pp. 141–168.
[8] H.B. Bradl, J. Colloid Interface Sci. 277 (2004) 1–18.
[9] G.W. Bruemmer, J. Gerth, K.G. Tiller, J. Soil Sci. 39 (1998) 37–52.
[10] F.X. Han, A. Banin, Water Air Soil Pollut. 114 (1999) 221–250.
[11] P.C. Gomes, M.P.F. Fontes, D.G. da Silva, E. de, S. Mendonça, A.R. Netto, Soil Sci. Soc. Am. J. 65 (2001) 1115–1121.
[12] E.F. Covelo, M.L. Andrade, F.A. Vega, Commun. Soil Sci. Plant Anal. 35 (19–20) (2004) 2709–2729.
[13] F.A. Vega, E.F. Covelo, M.L. Andrade, J. Colloid Interface Sci. 327 (2) (2008) 275–286.
[14] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees, Wadsworth, Belmont, CA, 1984.
[15] E.F. Covelo, J.M. Matías, F.A. Vega, M.J. Reigosa, M.L. Andrade, Geoderma 147 (1–2) (2008) 75–85.
[16] F.A. Vega, J.M. Matías, M.L. Andrade, M.J. Reigosa, E.F. Covelo, J. Hazard. Mater. 167 (1–3) (2009) 615–624.
[17] F. Guitián, T. Carballas, Técnicas de Análisis de Suelos, in: Editorial Pico Sacro, Santiago de Compostela, Spain, 1976.
[18] P.R. Day, Particle size analysis, in: C.A. Black (Ed.), Methods of Soil Analysis. Part I, American Society of Agronomy, Madison, WI, 1965.
[19] A. Walkey, I.A. Black, Soil Sci. 34 (1934) 29–38.
[20] N. Reeve, M.E. Sumner, Soil Sci. Soc. Am. Proc. 35 (1) (1971) 38–42.
[21] O. Rodríguez, A. Rodríguez, Rev. Fac. Agric. LUZ 19 (3) (2002) 253–263.
[22] G.W. Thomas, Exchangeable cations, in: A.L. Page, R.H. Miller, D.R. Keeney (Eds.), Methods of Soil Analysis. Part 2. Chemical and Microbiological Properties, 2nd ed., American Society of Agronomy, Madison, WI, 1982, pp. 159–166.
[23] O.P. Mehra, M.L. Jackson, Clay and Clay Minerals, in: 7th Conference, 1960, pp. 317–327.
[24] C. Alberti, A. Cristini, A. Loi, P. Melis and G. Pilo, Copper and lead sorption by different fractions of two Sardinian soils. in: R. Prost (Ed.), Contaminated Soils: Third International Conference on the Biogeochemistry of Trace-Elements. D: A data/communic/111.PDF, Colloque no 85, INRA Editions Paris, France, 1997.
[25] R.D. Harter, R. Naidu, Soil Sci. Soc. Am. J. 65 (2001) 597–612.
[26] L. Madrid, E. Díaz-Barrientos, J. Soil Sci. 43 (1992) 709–721.
[27] G.J. Lair, M.H. Gerzabek, G. Haberhauer, Environ. Chem. Lett. 5 (2007) 23–27.
[28] R.G. McLaren, J.G. Williams, R.S. Swift, Geoderma 31 (1983) 97.
[29] C.T. Chen, D.Y. Lee, M. Houng, J. Chin. Agric. Chem. Soc. 30 (1) (1992) 83–99.