Geoderma 170 (2012) 359–368

Contents lists available at SciVerse ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma

Mapping soil Pb stocks and availability in mainland France combining regression trees with robust geostatistics

Eva Lacarce [a,*], Nicolas P.A. Saby [a], Manuel P. Martin [a], Ben P. Marchant [b], Line Boulonne [a], Jeroen Meersmans [a], Claudy Jolivet [a], Antonio Bispo [c], Dominique Arrouays [a]

[a] INRA, US 1106 Infosol, INRA Orléans, CS 40001 Ardon, F45075 Orleans Cedex 2, France
[b] Rothamstead Research, Harpenden, Hertfordshire AL5 2JQ, United Kingdom
[c] ADEME, Centre d'Angers, Direction Déchets et Sols, 20 avenue du Grésillé, BP 90406, F49004 Angers Cedex 01, France

ARTICLE INFO

Article history:
Received 26 April 2011
Received in revised form 11 November 2011
Accepted 21 November 2011
Available online 29 December 2011

Keywords:
Soil Pb
Availability
Diffuse contamination mapping
Regression tree
Robust geostatistics

ABSTRACT

Maps of lead (Pb) stocks in soils and estimates of its availability are needed to assess risks of contamination. Stocks in soils of total and ethylenediamine tetraacetic acid (EDTA) extractable Pb, as well as Pb availability, assessed by EDTA/total Pb ratio, were measured and calculated to a depth of 30 cm with the French soil monitoring network at sites defined by a regular 16×16 km grid. Setting aside punctual anomalies by winsorizing, these properties were mapped using linear mixed models (LMM). LMMs combined conditional partitioning trees upon 5 predictors (pH, texture, parent material, land use, population density) with robust geostatistics to avoid distortion due to outlying values. Rather than selecting the fixed effects according to expert-knowledge, regression trees were used to account for explanatory variables in a single classification. This original method stressed both the necessity for a geostatistical component to complement regression tree models when spatial correlation is evident, and the usefulness of these trees to interpret maps. Pb stocks varied widely with peak concentrations and availability in densely populated areas. Lithology, texture and forestation also affected total Pb stocks. With regards to availability, forestation and pH appeared as key factors.

1. Introduction

Many soils are critically contaminated with Pb as a consequence of its use in industry or of its presence in ores (e.g. metal smelters, mining, glass, paints). Because Pb is generally not very mobile in soils (Semlali et al., 2004), surroundings of former factories and waste deposits remain contaminated (Arrouays et al., 1996; Douay et al., 2009). Additionally, more diffuse pollution arises from processes such as long range air-borne particles deposition. In mainland France, anthropogenic atmospheric emissions have declined 98% in the 1990–2008 period due to leaded gasoline abandonment (CITEPA, 2010). Soil deposition of Pb due to past leaded gasoline consumption is proportional to traffic intensity and to the distance from the roads (Laidlaw and Taylor, 2011; Markus and McBratney, 2001). Anthropogenic air emissions are now estimated at 95 t Pb per year and are essentially due to manufacturing (CITEPA, 2010). Additional to atmospheric deposition, agricultural soils in mainland France receive Pb by composts and sewage sludge (20% of the inputs) and animal manure used as fertilizers (44% of the inputs) (SOGREAH, 2007). Lead arsenate was intensively used in orchards and vineyards before it was slowly replaced by DDT (dichlorodiphenyltrichloroethane) in the sixties (Peryea, 1998) and then forbidden. Local redistribution of topsoil by eolian and hydrological erosion may also lead to land contamination (Quinton and Catt, 2007).

Contamination of soils by heavy metals has implications in terms of human health, agricultural production and environmental quality. Human poisoning with Pb in soils happens through 1) direct exposure to soil dust, 2) contamination of crops and 3) drinking water pollution by erosion or leaching processes that rely greatly on acidic pH and on the presence of bio-colloids (Citeau et al., 2003; Denaix et al., 2001). For example, at present low-level chronic exposure in cities is mainly due to soil dust re-suspension and deposition which affects child health worldwide (Demetriades et al., 2010; Glorennec et al., 2007; Laidlaw and Taylor, 2011; Mielke et al., 2011). However far from cities and industries, farmers have the highest blood lead levels among French adults (Falq et al., 2011) possibly because of diffuse pollution of soils. When the risk of exposure is considered, both point sources of pollution and diffuse contamination need to be mapped to implement proper land use policies (Markus and McBratney, 2001). To appraise potential risk due to contamination, stock estimation of Pb in soils is an essential basic input to flux models. Moreover its calculation is the only reliable way to use data from different monitoring networks and

Abbreviations: EDTA, ethylenediamine tetraacetic acid; RMQS, Réseau de Mesures de la Qualité des Sols i.e. the French soil monitoring network; LMM, Linear Mixed Model; REML, Restricted maximum likelihood; CI, confidence interval.

* Corresponding author. Tel.: +33 (0)1 43 38 65 42.
E-mail address: eva.lacarce@laposte.net (E. Lacarce).

0016-7061/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.geoderma.2011.11.014

to include organic top horizons. Beyond Pb stock, contamination risk depends on its availability and mobility (Markus and McBratney, 2001; Rieuwerts et al., 2006; Rodrigues et al., 2010; Sauvé et al., 2000). Ethylenediamine tetraacetic acid (EDTA) extraction of Pb was approved by the Measurements and Testing Programme of the European Commission and generally correlates well with plant available Pb (Quevauviller et al., 1998).

Soil surveys and monitoring networks (Arrouays et al., 2002; Lacarce et al., 2009; Morvan et al., 2008) have been designed for mapping soil properties and diffuse pollution at country to continental scale. However to derive maps from collected data, appropriate statistical methods should be used to avoid distortion by extreme observations (Marchant et al., 2010). These methods can also yield an understanding of the factors which cause contamination. Such an understanding is required by policy makers. Maps of some trace elements in soils across France have been derived from punctual data using robust geostatistics (Marchant et al., 2010; Saby et al., 2011). Similar methods were used by Marchant et al. (2011) to map soil metal contamination in an industrial region of Wales. These studies decomposed the variation of the properties into 3 effects within the framework of linear mixed models (LMM): 1) expert defined geological classes that constituted the fixed effect of the LMM, 2) spatially continuous variation within these classes that formed the random effect of the LMM, and 3) anomalous processes separated from the underlying model by winsorizing. Unfortunately, the number of external explanatory factors used was limited and hence the factors leading to the observed variation were not fully identified. However if too many factors are included the maps which result will be difficult to interpret and there will be a danger of the model being over-fitted. Therefore we propose the use of non-parametric intuitive models known as regression trees to parsimoniously select relevant factors to be included.

In this paper, we predict maps for total Pb and EDTA extractable Pb stocks as well as an index of Pb availability. These maps are based on stocks originally measured by the French soil monitoring network. External factors indicative of Pb stocks (pH, texture, parent material, land use, population density) were analysed and amalgamated with regression trees. We suspected these factors would not encompass the entire variation of the properties as spatial correlations exist and as additional predictors, for example accounting for punctual anomalies, were not available. Therefore the adopted methodology coupled 1) regression trees to understand factors that affect the properties with 2) robust geostatistics presented by Saby et al. (2011) to obtain maps by adding the fixed effects derived from regression trees to simple kriging predictions on winsorized data.

2. Material and methods

2.1. Data

2.1.1. Pb stocks calculation

Soil Pb stocks and properties and most explanatory variables were derived from the French National Soil Monitoring Network (Réseau de Mesures de la Qualité des Sols or RMQS). The RMQS surveys soils and their properties on a regular 16 km grid across the French mainland territory (550 000 km²) (Arrouays et al., 2002). It consists of 2091 sites discarding Corsica because it introduces spatial discontinuity in the data set. Where the sites defined by the grid could not be sampled (e.g. water bodies, no soil), an alternative location was selected as nearby as possible within a 1 km radius. When that was impossible the site was omitted. At each site, 25 individual core samples were taken for the organic layer when it existed and for mineral layers at depths 0–30 cm (topsoil) and 30–50 cm (subsoil) using an unaligned sampling design within a 20×20 m area. For each layer, core samples were bulked to make a composite sample. Then composite samples were air dried and sieved (AFNOR, 2006) before analysis by INRA Soil Analysis Laboratory at Arras. In parallel to composite sampling, adjacent to the composite sampling area, a soil pit was described. Bulk density was also measured in that pit. The cylinder method (AFNOR, 1992) was mainly used except for stony soils where an excavation method was preferred. When no density measurement was available, it was estimated by a pedotransfer function obtained on the full RMQS data based on the methodology of Martin et al. (2009).

Concerning Pb, two types of analyses were performed. 1) Total Pb content was determined by inductively coupled plasma mass spectrometry (ICP-MS) after dissolution with hydrofluoric and perchloric acids (AFNOR, 2001; AFNOR, 2003a). 2) Ethylenediamine tetraacetic acid (EDTA) extractable Pb content was determined by ICP atomic emission spectrometry after extraction with a 0.05 M EDTA solution at pH 7 in a ratio 1/10 (Quevauviller et al., 1998). Stocks in the fine earth were calculated, where possible, from the surface of the soil to 30 cm below the mineral soil surface by multiplying Pb contents by the mass of the fine earth in the total volume of soil. If organic horizons existed, they were added to the top 30 cm of mineral horizon or to the whole mineral horizon if it did not reach 30 cm. Organic horizons were present at 102 sites. Their mean and median thicknesses were 6.6 cm and 5.5 cm respectively. Stocks are all expressed in Pb $g.m^{-2}$. An index of Pb availability was calculated as the ratio EDTA/total Pb contents.

2.1.2. Predictor variables

pH was measured in water (AFNOR, 1994) and texture was derived from sand, silt and clay fractions measurements (AFNOR, 2003b) projected in the five class texture triangle from the soil map of Europe at 1:1 000 000 scale (King et al., 1995). The codes were as follows: 1 = coarse, 2 = medium, 3 = medium fine, 4 = fine and 5 = very fine. Surveyors reported parent material following the soil map of Europe nomenclature and land use according to Agreste classification (Agreste, 2009). Because these classifications include numerous classes many of which have few members, they were combined into broader classes listed in Tables 1 and 2. Land use nomenclature is a simplified Corine Land Cover classification while parent material classes follow a general lithological classification. Population density (people per km²) was also used as a predictor and was derived from the French 1990 census (INSEE, 2000).

2.2. Statistical methods

2.2.1. Regression trees

Statistical models, such as linear models, relate a response variable to various explanatory variables or predictors, and can be used to both predict and explain its variation. In soil science, explanatory variables often can be qualitative or quantitative, and frequently contain extreme values. These can have high leverage and adversely impact the fitted parameters. In that framework, regression trees are very handy and appropriate tools. They sequentially partition (or segregate in branches) the data according to predictors so that the response variable is as homogeneous as possible within each terminal node (also named class) and differs as much as possible between them. Regression trees have been shown to perform better for prediction than linear regression in certain cases (Vega et al., 2010). Unlike multiple additive regression trees, the interpretation of a single regression tree is straightforward: each division of the population corresponds to one predictor threshold or to a split in two groups of modalities. The estimator of each terminal node is usually the mean, but we preferred a more robust estimator, the median of the response at each terminal node. Because regression trees define classes, they are suitable to re-blend explanatory variables into a sound classification to be used instead of expert-defined classes.

Regression trees algorithms such as "CART" (Breiman et al., 1984) present some drawbacks. To avoid 1) over-fitting and the need for pruning and 2) bias in variable selection for the splits (preference is