Implementation of a COM-based decision-tree model with VBA in ArcGIS

Wei Cheng a,b, Ke Wang a,*, Xiuying Zhang c

a Institution of Remote Sensing and Information System Application, Zhejiang University, Hangzhou 310029, China
b Tourism, Resource and Environment Department, Zaozhuang University, Zaozhuang 277160, China
c International Institute for Earth System Science, Nanjing University, Nanjing 210093, China

Abstract

The problem of soil pollution by heavy metals has been receiving an increasing attention in the last few decades. Geostatistics module in ArcGIS, could not however efficiently simulate the spatial distribution of heavy metals with satisfied accuracy when the spatial autocorrelation of the study area severely destroyed by human activities. In this study, the classification and regression tree (CART) has been integrated into ArcGIS using ArcObjects and Visual Basic for Application (VBA) to predict the spatial distribution of soil heavy metals contents in the area severely polluted. The overall CART accuracy of assigning samples to the right Pb classes is 89.62% and 85.71%, the Kappa coefficient is 0.8444 and 0.7575, respectively for learning data and test data. This is a great improvement comparing with ordinary Kriging method in ArcGIS. The integrated approach allows for relatively easy, fast, and cost-effective estimation of spatially distributed soil heavy metals pollution. The methods and results described in this study are also valuable for understanding the relationship between heavy metals pollution risk and environmental factors.

1. Introduction

Soils are critical environments where rock, air and water interface. Alongside air, water and the biota, the soil is of central significance in ecosystem research as it is the place where many kinds of interactions take place between minerals, air, water and the living environment (Bloemena, 1995). The problem of soil pollution by heavy metals has been receiving an increasing attention in the last few decades. Heavy metals occur naturally in rocks and soils, but increasingly higher quantities of them are being released into the environment by anthropogenic activities. The soil is the primary recipient by design or accident of a myriad of waste products and chemicals used in modern industrial society (Brady & Weil, 2002). One challenge in predictive modeling of heavy metals contents varying in the anthropogenic activities is to achieve the prediction over space with acceptable accuracy. In practice, soil pollution by heavy metals is commonly assessed by interpolating concentrations of heavy metals sampled at point locations. Geostatistics module in ArcGIS, or more specifically, kriging estimation or stochastic simulation, provides a tool for the evaluation of spatial distribution of soil properties. However, relationships between heavy metals pollution and environmental factors are often non-parametric and involve complex interactions when humans play an important role in its dynamics. Because of this complexity, common linear and parametric models (Barrera-Bassols, Zinck, & Van Ranst, 2006; ChoT & David, 2005; Hillyer, McDonagh, & Verlinden, 2006; McGrath, Zhang, & Carton, 2004; Salgueiro, Ávila, Pereira, & Santos Oliveira, in press) that try to explain heavy metals pollution with associated environmental variables often do not provide good model fits. The geostatistical methods could not however efficiently simulate the spatial distribution of heavy metals with satisfied accuracy when the spatial autocorrelation of the study area severely destroyed by human activities, for a prior requirement of these methods is to quantify the spatial autocorrelation of the study area severely destroyed by human activities, for a prior requirement of these methods is to quantify the spatial autocorrelation between properties at different locations so that the information from samples can be weighted into an estimator of the values at unsampled locations (Yao, 1999).

Contemporary GIS applications often include tools to develop customizations that extend the capabilities of the system, thereby presenting the opportunity to link a GIS with even more powerful analytical modules that may not have been previously used for spatial analysis (Crossman, Perry, Bryan, & Ostendorf, 2007). However, there is still no implementation of the decision-tree models in the standard functionality of one of the most widespread GIS solutions, ArcGIS. In this study, the integration has been realized in ArcGIS by integrating the classification and regression tree (CART) using ArcObjects and Visual Basic for Application (VBA) into ArcGIS.
9.0 to predict the spatial distribution of soil heavy metals contents in the area severely polluted.

2. Methodology

2.1. Classification and regression tree (CART)

Classification and regression trees (Breiman, Friedman, Olshen, & Stone, 1984) are a data mining technique for empirical model building and hypothesis formulation. A major advantage of CART is that assumptions which are required for the appropriate use of parametric statistics, such as Gaussian distribution of predictor variables, do not need to be satisfied (Rothwell, Futter, & Dise, 2008). CART enables processing large sets of mixed data, i.e. nominal, ordinal and metric scale data. CART also allows uncovering hierarchical and non-linear relationships among one dependent variable and several predictors (Schröder et al., 2008). This is achieved by nested binary “if–then–else” splits, each maximizing the homogeneity of the target variable. CART handles both categorical and parametric data without data transformation and produces classification results that immediately indicate the variable that significantly discriminates between classes (Schröder, 2006). So the machine-learning, probabilistic, non-parametric decision-tree method has been widely applied in agriculture management (Waheed, Bonnell, Prasher, & Paulet, 2006; Zhang, Valentine, tree method has been widely applied in agriculture management (Waheed, Bonnell, Prasher, & Paulet, 2006; Zhang, Valentine, Kemp, & Lambert, 2006), traffic accidents (Chang & Wang, 2006), medical problem (Brady and Weil, 2002; Deconinck, Hancock, Coomans, Massart, & Vander Heyden, 2005), materials properties prediction (Li, 2006) and even archaeological research (Espa, Benedetti, De Meo, Ricci, & Espa, 2006).

Generally, CART analysis consists of three basic steps. The first step consists of tree building, during which a tree is built using recursive splitting of nodes. After a large tree is identified, the second stage of the CART methodology uses a pruning procedure that incorporates a minimal cost-complexity measure. The result of the pruning procedure is a nested subset of trees starting from the largest tree grown and continuing the process until only one node of the tree remains (Lee, Chiu, Chou, & Lu, 2006). A testing sample will be used to provide estimates of future classification errors for each sub-tree. The last stage of the methodology is to select the optimal tree, which corresponds to a tree yielding the lowest testing set error rate.

2.1.1. Building the maximal tree

In this study, we used measure of Gini impurity that used for categorical target variables. The Gini index at node \( t_i \), \( I(t) \), is defined as (Breiman et al., 1984):

\[
I(t) = 1 - \sum_{i=1}^{s} (P_i(t))^2
\]

The tree building process starts by dividing the root node into two nodes. CART searches for the best split, over all possible splits for each explanatory variable, which divides the objects of the root node into two more homogeneous nodes. The best split is found when the impurity function between the parent and the two new child nodes is minimized. The Gini criterion function for split \( s \) at node \( t_i \) is defined as (Kurt, Ture, & Kurum, 2008):

\[
\Delta(s, t) = I(t) - (p_{L}(t_s) + p_{R}(t_s))
\]

where \( s \) is the candidate split of a variable \( i \), \( t \) the parent node, \( i(t) \) the impurity of the node \( t \), \( p_{L} \) and \( p_{R} \) the proportions of objects going to the left \( (t_L) \) or right \( (t_R) \) child nodes, respectively, \( i(t_L) \) and \( i(t_R) \) their impurities. Several impurity measures have been proposed as splitting criteria. When a classification tree is being built, three criteria are usually used to choose the best split.

2.1.2. Tree-pruning

The over-large tree, built on the initial step of CART, describes the training set almost perfectly, but usually provides poor prediction results for new samples. Thus, finding a smaller tree with better predictive ability without losing much accuracy is essential. The optimal tree size is found by pruning, i.e., by successive cutting back branches of the over-large tree. This procedure determines a sequence of smaller trees and establishes which is the most accurate by calculating its cost-complexity. The cost-complexity measure \( R_s \) is defined as a linear combination of the cost of the tree and its complexity (Caetano, Aires-de-Sousa, Daszykowski, & Vander Heyden, 2005):

\[
R_s = R(T) + \alpha |T| \iff \alpha = \frac{R_s - R(T)}{|T|}
\]

where \( R(T) \) is the resubstitution estimated error, which for a classification tree is given by the misclassification error, \( |T| \) the size of the sub-tree (number of terminal nodes), and \( \alpha \) the complexity parameter. During the pruning procedure \( \alpha \) takes values between 0 and 1, and a sequence of nested trees of decreasing size is found.

2.1.3. Selection of the optimal tree

The principle behind selecting the optimal tree is to find a tree with respect to a measure of misclassification cost on the testing dataset (or an independent dataset), so that the information in the learning dataset will not be overfit.

2.2. Implementation of the CART in ArcGIS

By exploiting the modeling power of GIS through integration of GIS with decision-tree models, a GIS can be transformed from a simple spatial query and visualization tool to a powerful analytical and spatially distributed modeling tool (Satti & Jennifer, 2004). Recent advances in GIS technology facilitate the seamless integration of GIS and computer-based modeling. Some methods for integrating models with GIS have been categorized as ‘loose’, ‘close’, or ‘tight’ coupling. Loose coupling methods usually involve data exchange. An interface program is normally used to convert and organize GIS data into a format required by the model (Liao & Tim, 1997). Close coupling passes information between the GIS and the model via memory-resident data models rather than external files, leading to improved model interactions and performance (Di Luzio, Srinivasan, & Arnold, 2004). Tightly coupled model integration focuses on incorporating the functional components of one system within the other (i.e. the model within the GIS program) (Liao & Tim, 1997). In this study, the CART has been integrated into ArcGIS using ArcObjects and Visual Basic for Application (VBA) to predict the spatial distribution of soil heavy metals contents. ESRI’s ArcGIS 9.0 was chosen because it is widely used, powerful in functionality and allows easy project expansion (Winterton & Livermore 2004). ArcObjects is a COM compliant, object-oriented programming structure developed by Environmental Systems Research Institute Inc (ESRI). Because ArcObjects are the same software libraries on which the ArcGIS suites of applications are built, any function available in ArcGIS can be implemented programmatically through ArcObjects. Also ArcObjects provides application programming interfaces (APIs) that allow a model developer to programmatically access ArcGIS to automate repetitive tasks and extend its functionality using third-party Component Object Model-compliant (COM-compliant) programming languages such as Visual Basic, C++, Java, or Python (Stevens, Dragicevic, & Rothley, 2007).

There are three main reasons for us to develop the program into a VBA macro in the ArcGIS environment. (1) VBA macros in ArcGIS applications can use the ArcGIS functionality to its full extent; (2)
VBA macros can take advantage of global ArcGIS variables; (3) create, test and debug macros in the ArcGIS Visual Basic editor are the same as in the VB development environment (Marinoni, 2004).

The integration of the CART model in a geographic information system (GIS) combines decision support methodology with powerful visualization and mapping capabilities which in turn should considerably facilitate the creation of maps that indicating areas with pollution risks.

3. Study area and materials

Fuyang County, situated at the north of Zhejiang Province, near to the southwestern periphery of Hangzhou city, was selected as the study area. The county is located at 119°25′00″−120°19′30″E, 29°44′45″−30°11′58.5″N, and covers a region of 1831 km². The location of the study area is presented in Fig. 1. During the past three decades since economic reform in 1978, industrialization has increased at unprecedented rate, and paper making, mining and smelting are well developed.

Soil samples (302) were collected from different locations in March 2005 to take account of uniformity of soil sample distribution and soil types in the study area. The distribution of the 302 soil sampling points is presented in Fig. 1. All samples were taken at a depth of 0–20 cm and air-dried to remove stones and coarse plant roots or residues. The samples were thoroughly mixed and ground to pass through a 0.15 mm sieve, then stored in polythene bags for chemical analysis. Pb was determined by digesting the soil sample with a mixture of nitric acid (HNO₃) and perchloric acid (HClO₄) followed by Pb measurement in the digest by atomic absorption spectrometry. Soil pH was determined in a 1:2.5 soil: water ratio and organic matter was determined by wet oxidation at 180 °C with a mixture of potassium dichromate and sulfuric acid (Agricultural Chemistry Committee of China, 1983).

The CART model is implemented in VBA (Visual Basic for Applications) and integrated within the ArcObjects. The CART extension is automatically created in ArcMap during installation. When turned on, the extension operates as a dockable toolbar, similar to other ArcMap extensions. After selecting ‘Data Input’ from the ‘CART model’ pull down menu, a form appears as shown in Fig. 2. In our study area, the natural background soil Pb content was set to be 25 mg/kg (Zhejiang soil survey office. 1994). The content index was classified into six categories to indicate the level of Pb contamination.

Soil pH was included in the model for it is strongly correlated with soil Pb content determined by the method described here (Zhejiang Soil Survey Office, 1994). The other reason is that soil pH data are often more readily available from soil investigations than heavy metal data, and their value are relatively stable. Agricultural practices such as the use of manure or inorganic fertilizers could add heavy metals to soils, thus the agriculture land use practice was also selected to estimate heavy metals content. There are seven main agricultural land use in Fuyang county: paddy field (PF), dry land (DL), vegetable land (VT), tea garden (TG), orchard (OR), woodland (WO) and wasteland (WL). The independent variable was named LandUse.

Different industrial plants have different impacts on soil Pb accumulation.

(1) Smelting factories and disused smelting factories (SM), they produce and stock significant quantities of smelting wastes into piles and soil heaps (Jiang, Yang, Ye, Shi, & Jiang, 2002);
(2) Hardware machining (HM), especially factories related to Pb materials;
(3) Paper mills (PM), influence Pb concentrations due to paper mill sludge waste (Battaglia, Calace, Nardi, Paterno, & Pietroletti, 2007);
(4) Chemical plants (CM) influence Pb concentrations due to additives used;
(5) Others (OT), refers to the industry plants without relationship to heavy metals.
(6) NO, soil samples located outside the 500 m buffer zone of any industrial plants.

The independent variable was named INTType.

It has been noted that location close to roads are severally polluted by heavy metals such as Pb, Zn, Cu, Cd, etc. from traffic (Al-Khashman Omar, 2004). The Cd, Cu, Pb and Zn metal contents in road soils and total contents in grasses confirmed the effect of the traffic as source of pollution. To represent the influence of roads, soil samples within 100 m, 200 m, 300 m, 500 m, 1000 m

![Fig. 1. Location of study area and of sampling points.](image-url)
and outside 1000 m main roads buffer zone were respectively selected in this study. The independent variable was named RoadDist.

The developed VBA macro performs different activities such as extraction of the values of fields and records in the attribute table of the feature class, acquisition of the user input information for estimation of the soil heavy metals pollution. The overall goal of the CART analysis is to devise models that will use the predictor variables to predict the values of the response variable in a non-parametric way. A minimum node size of five or 1% of total number of dataset (records) was applied in the CART, the maximum tree depth or maximum purity also can be specified in the VBA macro. The data here is divided into two subsets, one for learning and the other for testing. The learning sample is used to split nodes, while the testing sample is used to compare the misclassification (see Fig. 3).

The developed VBA macro calculates classes that are homogeneous with respect to the features of the dependent variable (here the soil Pb content). It identifies the predictor variables with the highest correlation with soil Pb content by splitting the data set into the two most dissimilar groups. The splitting of the data set and tree development continue until the data in each group are sufficiently uniform. The method here partitions the data set into six discrete subgroups, based on the classification value of the dependent variable. The CART VBA macro finally produces a set of rules to allocate samples to predefined classes (see Fig. 4). The
rules are important in two ways. First, they are used to predict the values of the dependent (response) variable. Second, they contain a wealth of information about the relationship between the response and the predictor variables and the interactions among the predictors (Li, 2006).

4. Results and discussion

The ArcGIS tool with built-in macro programming languages and Component Object Model (COM) compliant protocols facilitated the integration of different data structures and programming logics (Sarangi, Madramootoo, & Singh, 2004). In ArcGIS tool, the Visual Basic for Applications (VBA) programming language was flexible enough to permit interoperability of different modules and data architectures within a multi task environment (Razavi, 2002).

The confusion matrix (see Fig. 5) shows the relationship between measured and estimated Pb classes. The total accuracy refers to the ratio of total number of correctly inferred Pb classes divided by the total number of samples (training, test data, respectively); and the Kappa Coefficient uses all of the information in the confusion matrix, ranging from 0 to 1. The overall CART accuracy of assigning samples to the right Pb classes is 89.62% and 85.71%, the Kappa coefficient is 0.8444 and 0.7575, respectively for training data and test data. The samples used in CART were also used in Kriging, Geostatistics module in ArcGIS to estimate Pb content spatially. Kriging estimates variable values at unknown locations from a semi-variogram model and appropriately sampled data set. Kriging uses the semi-variogram to quantify the spatial
variation. Although normality may not be strictly required in Krige- 
ging, serious violation of normality, such as high skewness and 
outliers, can impair the simulation results (McGrath et al., 2004).
The total accuracy of assigning kriged estimates of Pb classes to 
measured values is 42.65%, and the corresponding Kappa coeffi-
cient is 0.47. The main reason for increased accuracy might be 
that Pb content in this study area is greatly influenced by human activ-
ities leading to localized sharp variations and hotspots which are 
smoothed over by Kriging with a long range variogram.

Certainly, the method of CART decreases the measurement scale of 
the raw data to a lower level for the classification of the target. But 
these classifications are however useful when detailed concentra-
tions are not required. And the CART also can be used to explore 
the relationship between soil heavy metals pollution and environmen-
tal factors and identify the areas of soil heavy metals pollution risk. The 
results can be used to advise the local government in prioritizing the 
areas of soil heavy metals pollution risk. So the two methods have 
their own respective advantages and disadvantages in simulation 
the spatial distribution of soil heavy metals concentrations.

5. Conclusion

The developed CART VBA macro integrates all the components 
described in the methodology section to the user and allows the 
configuration, prediction, visualization and analysis of the model 
outcomes in the same environment thereby providing simplicity 
and flexibility. Code was written in VBA and deployed as a tem-
plate file from which any ArcGIS project can be derived. Also the 
CART model is easy to apply and the functions are accessible by 
users who do not have expert knowledge of modeling and 
programming.

The described VBA macro which implements the CART model is 
a useful tool for assessment of heavy metals pollution. The pre-
sentated VBA macro fills an important gap in the ArcGIS functional-
ity, since the decision-tree models does not belong to the standard 
functionality of this widely used GIS. The integrated approach al-
so allows for relatively easy, fast, and cost-effective estimation of 
spatially distributed soil heavy metals pollution. The methods 
and results described in this study are valuable for understanding 
the relationship between heavy metals pollution risk and environmen-
tal factors. Also the description of the macro provided a template 
not only for users who are working in the field of heavy metals 
pollution assessment but also in other fields of geosciences.

Acknowledgement

Funding for this research was provided by the National 
Technology Support Foundation (2006BAD10A07) and (2006 
BAJ05A02).

References

Agricultural Chemistry Committee of China (1983). Conventional methods of soil and 

Al-Khassah Omar, A. (2004). Heavy metal distribution in dust, street dust and 
slo Mrs from the work place in Karak Industrial Estate, Jordan. Atmospheric 
Environment, 38, 6803–6812.

comparison of indigenous and technical soil maps in a Mesoamerican 

of Pb and Zn bioavailable forms in metal polluted soils due to paper mill sludge 
addition Effects on Pb and Zn transferability to barley. Bioresource Technology, 
98, 2993–2999.

Bloemena, M.-L. (1995). The distribution of Cd, Cu, Pb and Zn in topsoils of 
Drenthe in relation to land use. The Science of the Total Environment, 166, 
137–148.

New Jersey: Pearson Education.


Prediction of enantioselectivity using chirality codes and classification and 

non-parametric classification tree techniques. Accident Analysis and 
Prevention, 38, 1019–1027.


session conservation evaluation and design optimisation system. Environmental 

Classification of drugs in absorption classes using the classification and 
regression trees (CART) methodology. Journal of Pharmaceutical and Biomedical 
Analysis, 39, 91–103.

system for the watershed assessment of agricultural nonpoint and point 

and estimation methods for the probability of archaeological site location. Journal 
of Cultural Heritage, 7, 47–155.

northern Namibia: The value of a local classification system. Agriculture, 

Jiang, L., Yang, X., Ye, H., Shi, W., & Jiang, Y. (2002). Effect of copper refining on 
spatial distribution of heavy metal in surrounding soils and crops. Journal of 

regression, classification and regression tree, and neural networks for 
predicting coronary artery disease. Expert Systems with Applications, 34, 
366–374.

Lee, T.-S., Chiu, C.-C., Chou, Y.-C., & Liu, C.-J. (2006). Mining the customer credit using 
classification and regression tree and multivariate adaptive regression splines. 
Computational Statistics and Data Analysis, 50, 1113–1130.

Li, Y. (2006). Predicting materials properties and behaviors using classification and 

591–603.

in ArcGIS. Computers and Geosciences, 30, 637–646.

assessment on soil lead in Silvermines area, Ireland Environmental Pollution, 127, 
239–248.


model of controls on dissolved inorganic nitrogen leaching from European 

Geostatistical estimation of chemical contamination in stream sediments: 
The case study of Vale das Gatas mine (northern Portugal). Journal of Geochemical 

assisted user interfaces for estimation of watershed morphologic parameters. 
Journal of Soil Water Conservation, 3(3–4), 139–149.


for data analysis and mapping in environmental monitoring and epidemiology. 
International Journal of Medical Microbiology, 296(51), 23–36.

Schroder, W., Pesch, R., Engler, C., Harmens, H., Suchara, I., Zechmeister, H. G., 
et al. (2008). Metal accumulation in mosses across national boundaries: 
Uncovering and ranking causes of spatial variation. Environmental Pollution, 
151, 377–388.

urban planning and decision making. Environmental Modelling and Software, 22, 
761–773.

in precision agriculture: CART – A decision tree approach. Agricultural Water 


Yao, T. (1999). Nonparametric cross-covariance modeling as exemplified by soil 
heavy metal concentrations from the Swiss Jura. Geoderma, 88, 13–38.

pasture productivity: integration of a decision tree and a geographical 
information system. Agricultural Systems, 87, 1–17.