A COM-GIS BASED DECISION TREE MODEL IN AGRICULTURAL APPLICATION

Wei Cheng[1,*], Ke Wang[1], Xiuying Zhang[1]

[1] *Institution of Remote Sensing & Information System Application, Zhejiang University, Hangzhou 310029, China.*

[*] *Corresponding author, Address: Institution of Remote Sensing & Information System Application, Zhejiang University, Hangzhou 310029, China, Tel: +86 571 86971272, Fax: +86 571 86971272, Email: 888888@zju.edu.cn (W. Cheng)*

Abstract: The problem of agricultural 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. 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.

Key words: ArcGIS; CART; Fuyang; heavy metals pollution; Pb; VBA macro

1. INTRODUCTION

Soils are critical environments where rock, air and water interface. 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 and Weil, 2002). 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 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 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 et al., 1995). 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)

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. (Schro¨der et al., 2008) CART handles both categorical and parametric data without data transformation and produces classification results that immediately indicate the variable that significantly discriminates between classes (Schro¨der, 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 et al., 2006). A testing sample will be used to provide estimates of future classification errors for each subtree. 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, $i(t)$, is defined as (Breiman et al., 1984):

$$i(t) = 1 - \sum_{i=1}^{n} (P_j(t))^2$$

The Gini criterion function for split s at node t is defined as (Kurt et al., 2008):

$$\Delta(s,t) = i(t) - (p_L i(t_L) + p_R i(t_R))$$

where *s* is the candidate split of a variable *v*, *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

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_α is defined as a linear combination of the cost of the tree and its complexity (Caetano et al., 2004):

$$R_\alpha = R(T) + \alpha\left|\tilde{T}\right| \Leftrightarrow \alpha = \frac{R_\alpha - R(T)}{\left|\tilde{T}\right|}$$

where *R*(*T*) is the resubstitution estimated error, which for a classification tree is given by the misclassification error, $\left|\tilde{T}\right|$ the size of the sub-tree (number of terminal nodes), and α the complexity parameter. During the pruning procedure α 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 et al., 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 and 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 et al., 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 and 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 et al., 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 et al., 2007)

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. 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 km^2. 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.

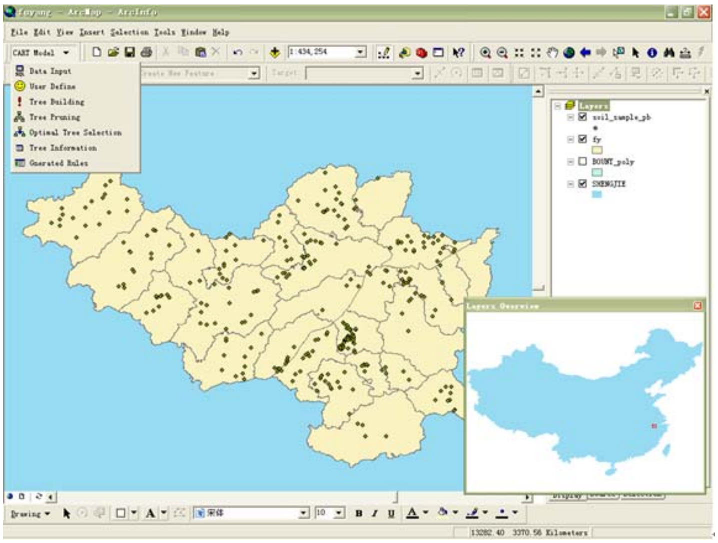


Fig. 1 Location of study area and of sampling points

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 0-20cm 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 (HNO_3) and perchloric acid ($HClO_4$) 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℃ with a mixture of potassium dichromate and sulfuric acid (Agricultural Chemistry Committee of China, 1983).

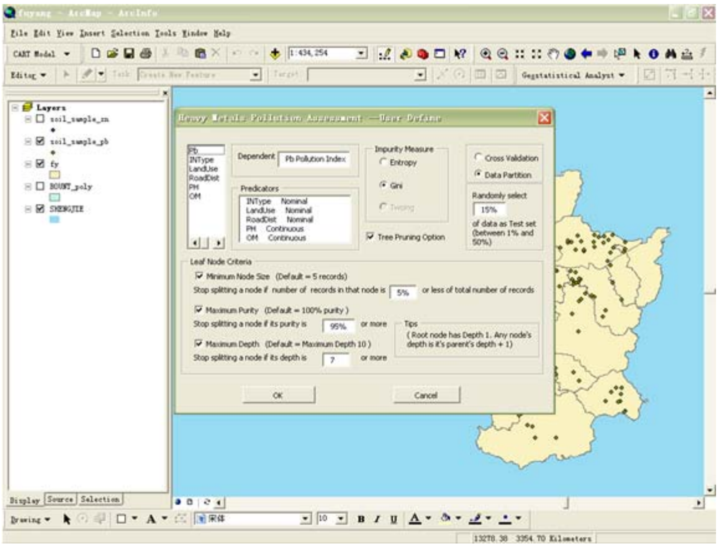
The CART model is implemented in VBA (Visual Basic for Applications) and integrated within the ArcObjects. The CART extension is automatically created in ArcMap during installation. When turned on, the extension operates as a dockable toolbar, similar to other ArcMap extensions. After selecting ‘Data Input’ from the ‘CART model’ pull down menu, a form appears.In our study area, the natural background soil Pb content was set to be 25mg/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 7 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.

The independent variable was named INType.

It has been noted that location close to roads are severally polluted by heavy metals such as Pb, Zn, Cu, Cd, etc. from traffic (Khashman, 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 100m, 200m, 300m, 500m,1000m and outside 1000m main roads buffer zone were respectively selected in this study. The independent variable was named RoadDist.



Heavy Metals Pollution Assessment —User Define

Pb
INType
LandUse
RoadDist
PH
OM

Dependent: Pb Pollution Index

Predicators
INType Nominal
LandUse Nominal
RoadDist Nominal
PH Continuous
OM Continuous

Impurity Measure
○ Entropy
● Gini
○ Twoing

☑ Tree Pruning Option

○ Cross Validation
● Data Partition

Randomly select 15% of data as Test set (between 1% and 50%)

Leaf Node Criteria

☑ Minimum Node Size (Default = 5 records)
Stop splitting a node if number of records in that node is 5% or less of total number of records

☑ Maximum Purity (Default = 100% purity)
Stop splitting a node if its purity is 95% or more

☑ Maximum Depth (Default = Maximum Depth 10)
Stop splitting a node if its depth is 7 or more

Tips
(Root node has Depth 1. Any node's depth is it's parent's depth + 1)

OK Cancel

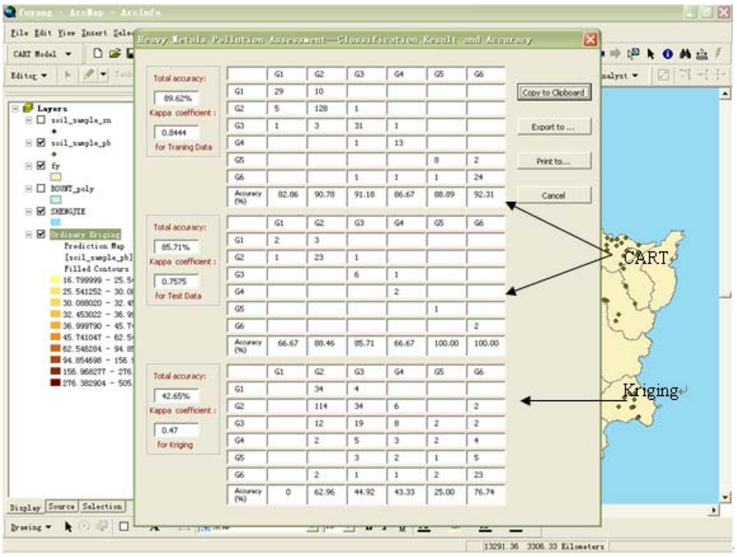
Fig. 2 Main interface of the User Input data of the developed CART VBA macro

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. 2).

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. 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 et al. 2004). The confusion matrix (see Fig. 3) 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..The total accuracy of assigning kriged estimates of Pb classes to measured values is 42.65%, and the corresponding Kappa coefficient is 0.47. The main reason for increased accuracy might be that Pb content in this study area is greatly influenced by human activities leading to localized sharp variations and hotspots which are smoothed over by Kriging with a long range variogram.



Heavy Metals Pollution Assessment—Classification Result and Accuracy

Total accuracy: 89.62%
Kappa coefficient : 0.8444
for Traning Data

	G1	G2	G3	G4	G5	G6
G1	29	10				
G2	5	128	1			
G3	1	3	31	1		
G4			1	13		
G5					8	2
G6			1	1	1	24
Accuracy (%)	82.86	90.78	91.18	86.67	88.89	92.31

Total accuracy: 85.71%
Kappa coefficient : 0.7575
for Test Data

	G1	G2	G3	G4	G5	G6
G1	2	3				
G2	1	23	1			
G3			6	1		
G4				2		
G5					1	
G6						2
Accuracy (%)	66.67	88.46	85.71	66.67	100.00	100.00

Total accuracy: 42.65%
Kappa coefficient : 0.47
for Kriging

	G1	G2	G3	G4	G5	G6
G1		34	4			
G2		114	34	6		2
G3		12	19	8	2	2
G4		2	5	3	2	4
G5			3	2	1	5
G6		2	1	1	2	23
Accuracy (%)	0	62.96	44.92	43.33	25.00	76.74

Fig.3 The result of the CART model and the kriging method

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 template 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 presented VBA macro fills an important gap in the ArcGIS functionality, since the decision tree models does not belong to the standard functionality of this widely used GIS. 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 valuable for understanding the relationship between heavy metals pollution risk and environmental 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.

ACKNOWLEDGEMENTS

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

REFERENCES

Agricultural Chemistry Committee of China, Conventional Methods of Soil and Agricultural Chemistry Analysis,Science Press, (Beijing, China), (in Chinese), 1983

Al-Khashman Omar. A. "Heavy metal distribution in dust, street dust and soils from the work place in Karak Industrial Estate, Jordan", Atmospheric Environment, 2004, 38, 6803–6812

Brady, N.C., Weil, R.R. The Nature and Properties of Soils. Pearson Education, New Jersey, 2002, 797–837

Breault Joseph L., Colin R. Goodall, Peter J. Fos, Data mining a diabetic data warehouse, Artificial Intelligence in Medicine, 26, 2002, 37–54

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. Classification and Regression Trees, Pacific Grove, CA Wadsworth, 1984

Caetano S., J. Aires-de-Sousa, M. Daszykowskia, Y. Vander Heyden, "Prediction of enantioselectivity using chirality codes and Classification and Regression Tree", Analytica Chimica Acta, 2005,544, 315–326

Crossman Neville D., Lyall M. Perry, Brett A. Bryan, Bertram Ostendorf, "CREDOS: A Conservation Reserve Evaluation And Design Optimisation System", Environmental Modelling & Software , 2007,22, 449-463

Di Luzio, M., Srinivasan, R., Arnold, J.G. "A GIS-coupled hydrological model system for the watershed assessment of agricultural nonpoint and point sources of pollution", Transactions in GIS 8 (1), 2004, 113–136.

Kurt Imran, Mevlut Ture, A. Turhan Kurum. "Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease", .Expert Systems with Applications, 2008, 34, 366–374

Lee Tian-Shyug, Chih-Chou Chiu,Yu-Chao Chou, Chi-Jie Lu. "Mining the customer credit using classification and regression tree and multivariate adaptive regression splines", Computational Statistics & Data Analysis, 2006, 50, 1113 – 1130

Li Yong, "Predicting materials properties and behavior using classification and regression trees", Materials Science and Engineering A, 2006, 433, 261–268

Liao, H., Tim, U.S. "An interactive modeling environment for nonpoint source pollution control", Journal of American Water Resources Association, 1997, 33 (3), 591–603.

Sarangi A, Madramootoo CA, Singh DK, "Development of ArcGIS assisted user interfaces for estimation of watershed morphologic parameters", J Soil Water Conserv, 3 (3, 4) (2004), 139–149

Satti Sudheer R., Jennifer M. Jacobs, "A GIS-based model to estimate the regionally distributed drought water demand", Agricultural Water Management, 2004, 66, 1–13

Schro¨ der Winfried, "GIS, geostatistics, metadata banking, and tree-based models for data analysis and mapping in environmental monitoring and epidemiology", International Journal of Medical Microbiology, 2006, 296 S1, 23–36

Schro¨der Winfried , Roland Pesch , Cordula Englerta, Harry Harmens , Ivan Suchara , Harald G. Zechmeister , Lotti Tho¨ni , Blanka Manˇkovska´ , Zvonka Jeran ,Krystyna Grodzinsk, Renate Alber, "Metal accumulation in mosses across national boundaries: Uncovering and ranking causes of spatial variation", Environmental Pollution, 2008, 151, 377-388

Stevens D., S. Dragicevic, K. Rothley, "ICity: A GIS-CA modelling tool for urban planning and decision making", Environmental Modelling & Software, 2007, 22, 761-773

Waheed T., R.B. Bonnell, S.O. Prasher, E. Paulet, "Measuring performance in precision agriculture: CART—A decision tree approach", Agricultural water management, 2006, 84, 173 – 185

Winterton Rose L. and Roy A. Livermore, "A Customised GIS to Aid Gondwana Research", Gondwana Research, V 7, No. 1, 287-292.

Yao Tingting, "Nonparametric cross-covariance modeling as exemplified by soil heavy metal concentrations from the Swiss Jura", Geoderma, 1999, 88, 13-38

Zhejiang Soil Survey Office, Zhejiang Soils. Zhejiang Technology Press, Hangzhou, China (In Chinese), (1994)