Ecological Indicators 77 (2017) 129–138

Contents lists available at ScienceDirect

Ecological Indicators

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

Original Article

Error index for additional sampling to map soil contaminant grades

Bingbo Gao[a,b], Yu Liu[a,b], Yuchun Pan[a,b,*], YunBing Gao[a,b], Ziyue Chen[c], Xiaolan Li[d,e], Yanbing Zhou[d,e]

[a] *Beijing Research Center for Information Technology in Agriculture, Beijing Academy of Agriculture and Forestry Sciences, 11 Shuguang Huayuan Middle Road, Haidian District, Beijing 100097, China*
[b] *National Engineering Research Center for Information Technology in Agriculture, Beijing, China*
[c] *College of Global Change and Earth System Science, Beijing Normal University, 19 Xinjiekouwai Street, Haidian, Beijing 100875, China*
[d] *Key Laboratory of Agri-informatics, Ministry of Agriculture, 11 Shuguang Huayuan Middle Road, Beijing 100097, China*
[e] *Beijing Engineering Research Center of Agricultural Internet of Things, 11 Shuguang Huayuan Middle Road, Beijing 100097, China*

ARTICLE INFO

Article history:
Received 25 October 2016
Received in revised form
17 December 2016
Accepted 9 February 2017

Keywords:
Error index
Additional sampling
Map
Soil contaminant grades

ABSTRACT

Soil contaminant grades classified by thresholds of concentrations are important for land use and management. Mapping of the soil contaminant grades only focuses on the precise prediction of the relationship between the contaminant concentration of each spatial units and a predefined threshold. Thus, unlike sampling for other purposes, additional sampling should add more sites in areas with a high possibility of misclassification and fewer or no sites in areas with a low possibility of misclassification. To guide additional sampling for mapping of the soil contaminant grades, an error index based on Indicator Kriging (IK) is proposed in this paper. By linear transformation of the predicted value of IK and summing the standard variance of the prediction error, the error index contains both the closeness of predicted value to the threshold and the uncertainty of the prediction, and can reflect the possibility of misclassification of the soil contaminant grades. Also, due to the adoption of IK, it can avoid the smoothing effect, remove or lower the unstationary variation of original data by indicator transform, and do not required normal distribution to model the error variance. Based on the error index, an optimization method for additional sampling to map soil contaminant grades is then put forward by defining an objective function and employing the Spatial Simulated Annealing optimization method. The chromium concentration data of the study area in central China were used as a case study. According to a comparison of the proposed additional sampling optimization method with spatially even sampling method and spatially random sampling method, the case study demonstrated that the proposed optimization method based on the error index is superior to the other methods in improving the prediction precision of soil contaminant grades and that its performance is stable. The results suggest that the error index proposed in this paper can be used to generate a design for additional sampling to improve the mapping precision of soil contaminant grades.

1. Introduction

Soil pollution becomes more and more severe and is posing a big threat to public health with the development of industry and increasing of agricultural inputs(D'Emilio et al., 2013; Yang et al., 2015). To protect environment and human health, land use activities should be regulated, and sometimes soil remediation actions should be applied in polluted areas. In order to provide a uniform guide on land use regulation or soil remediation, soil environmental specifications or standards that define different contaminant grades have been developed by many countries, such as The environmental quality standard for soils in China (GB 15618-1995), Guidelines for environmental management of EPA Victoria, Australia. By comparing the concentration values of soil contaminant with the thresholds defined in related soil environmental specifications or standards, the soil contaminant grades can be computed.

To obtain a map of soil contaminant grades, i.e. the exhaustive contaminant grade of every spatial units in the study area, spatial sampling and inference is still the most important method. Spatial inference is the goal of spatial sampling and serves as a guide to how the sampling sites should be selected (Gruijter et al., 2006; Wang et al., 2013). Ordinary Kriging (OK), Kriging with an

* Corresponding author at: Beijing Research Center for Information Technology in Agriculture, Beijing Academy of Agriculture and Forestry Sciences, 11 Shuguang Huayuan Middle Road, Haidian District, Beijing 100097, China.
E-mail address: panyc@nercita.org.cn (Y. Pan).

http://dx.doi.org/10.1016/j.ecolind.2017.02.011
1470-160X/

External Drift (KED) and CoKriging (CK) are frequently used to map the concentration of soil contaminant based on sampling data (Li and Heap, 2011, 2014; Minasny and McBratney, 2010; Najafian et al., 2012). However, the smoothing effect and the normal distribution requirement to model the error variance of prediction have limited their utilization in the prediction of soil contaminant grades (Isaaks and Srivastava, 1989). Due to the increase in small values and decrease in large values resulting from the smoothing effect, the corresponding spatial units may be misclassified into the wrong grades. In addition, the distribution of soil contaminant is usually highly skewed and not easily transformed into a normal distribution (Juang et al., 2001). Indicator Kriging (IK), which can estimate the probability of not exceeding a threshold, is suggested to be used in prediction of the soil contaminant grades(Antunes and Albuquerque, 2013; Chica-Olmo et al., 2014; Goovaerts, 1997; Journel, 1983).

In soil environment survey, additional sampling is needed in multi-stage sampling or supplementary investigation. To improve the precision of soil contaminant mapping, the additional sampling design should be optimized to reach higher precision with a limited sample size. Three types of purposive sampling, i.e., minimizing the variance of the estimation error, even coverage in geographical space, and even coverage in feature space, are frequently used to obtain samples for spatial mapping (Wang et al., 2012). However, these sampling approaches are not suitable for additional sampling for soil contaminant grades mapping, in which the sampling sites should be located densely in areas with a high possibility of misclassification and sparsely or not at all in areas with a low possibility of misclassification (Van Meirvenne and Goovaerts, 2001). Thus, the uncertainty of the soil contaminant grades misclassification can be employed in the additional sampling optimization for soil contaminant grades mapping. Based on the prediction result of IK, Garcia and Froidevaux (1997) proposed a method to classify the study area into low-risk regions, high-risk regions and medium-risk regions based on the minimum probability threshold (e.g., 0.2) and the maximum probability threshold (e.g., 0.8), and suggested that additional sampling sites can be selected from the medium-risk regions to better predict soil contamination grades. In addition to the predicted probability of exceeding the threshold defining the contaminant grades, the uncertainty of the prediction should also be taken into account in additional sampling optimization. Based on conditional simulation, Van Meirvenne and Goovaerts (2001) defined a sampling criterion ratio, which is calculated by dividing the standard deviation of the local cumulative distribution of the difference between the simulated concentrations and the threshold with the mean of that local cumulative distribution. This sampling criterion ratio contains both the proximity to the threshold and the variance to guide the additional sampling for soil contaminant grades mapping, but because thousands of simulations are needed, new uncertainty is introduced and it cannot be combined with optimization algorithms to design the sampling plan automatically. In addition, this method requires the full cumulative distribution function, which is not easy to obtain. Juang et al. (2008) proposed an index reflecting both type I error and type II error to guide additional sampling for the soil contamination grades classification based on rank-order geostatistics. This index is a direct measure of the probability of the misclassification of soil contaminant grades through the consideration of both the proximity of the predicted value to the threshold and the uncertainty of the prediction. However, this index depends heavily on the distribution of the sampling values; if the sample is biased from the population, the transformation between the concentration and standardized rank will be inaccurate. The index does not allow different criteria to be set for two types of misclassification, which is often required in practical applications. For example, to classify contaminants with high health risks in farmland, type II errors must be minimized to reduce the health risk at the expense of large type I error, while in the division of hazardous areas for prior remediation, type I errors must be minimized to improve the efficiency of investment at the expense of greater type II error. To improve the precision of the soil contaminant grades classification, in this paper, we propose an error index for additional sampling that measures the probability of misclassification of soil contaminant grades by considering the uncertainty of prediction and allows the adjustment of the two types of errors according to practical needs.

The remainder of this paper contains 4 sections. Section 2 presents the prediction method of soil contaminant grades, the error index and the optimization method for additional sampling that is based on it. Section 3 introduces case studies of the optimization method for additional sampling based on the error index. Section 4 discusses the error index. Section 5 presents conclusions.

2. Methods

2.1. Classification of soil contaminant grades with IK

The spatial units whose contaminant grades are to be predicted can be points or polygons. For point, the contaminant grades of each point must be classified, and these units are often used in soil environmental assessment of large spatial scale, for example, to divide hazardous areas for remediation. For polygon, only one contaminant grade is spatial for each polygon. Polygon may be parcels, administrative regions or other areas, and they are often used in soil environmental assessment of small spatial scale, for example, to predict soil contaminants grades for land use management. IK and Block Indicator Kriging (BIK) can be used to predict and classify the soil contaminants grades for point and polygon, respectively.

IK first transforms the concentration values into indicator values based on thresholds, as expressed in Eq. (1), and then predicates the cumulative probability for not exceeding the threshold of a location, as shown in Eq. (2), or of a region, as shown in Eq. (3) (BIK) (Goovaerts, 1997).

$$I\left(x;Z_c\right)=\begin{cases}1Z\left(x\right)\le Z_c\\0Z\left(x\right)>Z_c\end{cases}\qquad(1)$$

where $Z\left(x\right)$. is the concentration value at spatial unit x, Zc is a threshold, and $I\left(x;Z_c\right)$ is the corresponding indicator value.

$$\hat{F}(x_0;Z_c|(n))=\sum_{i=1}^{n}\lambda_i\mathrm{I}(x_i;Z_c)\qquad(2)$$

here $F\left(x_0;Z_c|\left(\mathrm{n}\right)\right)$ is the cumulative probability of not exceeding Zc at x0 and λi is the weight of the ith indicator value.

$$\hat{F}\left(x,Z_c|\left(\mathrm{n}\right)\right)=\sum_{i=1}^{n}\lambda_i I\left(x_i;Z_c\right)\qquad(3)$$

where $F\left(x,Z_c|\left(\mathrm{n}\right)\right)$ is the cumulative probability of not exceeding Zc of the polygon x.

The spatial correlation is the basis for solving the weights of Eqs. (1) and (2). In geostatistics, the spatial variogram which is defined in Eq. (4) is used to represent the spatial correlation.

$$\gamma_I(h)=\frac{1}{2}E([I(x)-I(x+h)]^2)=\frac{1}{2N(h)}\sum_{\partial=1}^{N(h)}\left[I(x_\partial)-I(x_\partial+h)\right]^2\qquad(4)$$

Thus before conducting Kriging, the spatial variogram should be modeled first. Then based on the spatial variogram model, the λi in