Journal of Geochemical Exploration 101 (2009) 265–282

Contents lists available at ScienceDirect

Journal of Geochemical Exploration

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

Evaluating and classifying contaminated areas based on loss functions using annealing simulations

Joaquim C.B. Queiroz [a], José R. Sturaro [b], Augusto C.F. Saraiva [c], Paulo M.B. Landim [b,*]

[a] Department of Statistics, Federal University of Para, Belém, Pará, Brazil
[b] Department of Applied Geology, São Paulo State University, Rio Claro, São Paulo, Brazil
[c] Central Laboratory Eletronorte, Belém, Pará, Brazil

ARTICLE INFO

Article history:
Received 18 November 2007
Accepted 25 September 2008
Available online 25 October 2008

Keywords:
Decision making process
Uncertainty modeling
Risk and loss functions
Annealing simulation
Geostatistics

ABSTRACT

This paper presents a methodology based on geostatistical theory for quantifying the risks associated with heavy-metal contamination in the harbor area of Santana, Amapá State, Northern Brazil. In this area there were activities related to the commercialization of manganese ore from Serra do Navio. Manganese and arsenic concentrations at unsampled sites were estimated by postprocessing results from stochastic annealing simulations; the simulations were used to test different criteria for optimization, including average, median, and quantiles. For classifying areas as contaminated or uncontaminated, estimated quantiles based on functions of asymmetric loss showed better results than did estimates based on symmetric loss, such as the average or the median. The use of specific loss functions in the decision-making process can reduce the costs of remediation and health maintenance. The highest global health costs were observed for manganese.

1. Introduction

Environmental contamination at industrial sites is a particular problem that requires a remediation approach plus monitoring of the remediation process itself. In environmental remediation, however, the inevitable presence of uncertainty makes decision making time-consuming and difficult. The uncertainty is caused by unknown concentrations of toxic elements at unsampled locations. The only way to reduce the uncertainty would be to collect additional samples but, because of the finite nature of sampling, even this would not altogether eliminate the uncertainty (Istok and Rautman, 1996). The error could, however, be minimized by using loss functions from an uncertainly model so as to minimize the difference between real and estimated values.

Quantifying the uncertainty of a random variable Z is done using a statistical distribution that describes the frequency of unsampled locations, $\boldsymbol{u}$. This can be represented by the conditional cumulative distribution function (ccdf), defined as

$$F(\boldsymbol{u};z\,|\,(n)) = \text{Prob}\{Z(\boldsymbol{u}) \leq Z(n)\} \tag{1}$$

The function $F(\boldsymbol{u};Z|(n))$ depends on the number of sampled points, n, the space configuration, the data values, and the specific space phenomenon in the study. The uncertainty is incorporated, therefore, starting from a model of $F(\boldsymbol{u};Z|(n))$ for the variables in the study, and the error is minimized by the loss functions used. The function $F(\boldsymbol{u};Z|(n))$ can be obtained from a stochastic simulation that corresponds to the variables in the study. Each simulation provides a concentration map of the contamination in the study area, consistent with known concentrations, the sample histogram, and the space continuity pattern exhibited by the data. In this case, each simulation is defined by a joint distribution at all grid locations taken at the same time. This guarantees that the space uncertainty is modeled starting from multiple fulfillments of joint distributions of values in the space.

Modeling the uncertainty of $z(\boldsymbol{u})$ is done with the function $F(\boldsymbol{u};Z|(n))$ in the sense that probability intervals can be derived, such as,

$$\text{Prob}\{Z(\boldsymbol{u}){\in}(a,b] \setminus (n)\} = F(\boldsymbol{u};b \setminus (n)) - F(\boldsymbol{u};a\,|\,(n)) \tag{2}$$

or in the sense that the probability of $Z(\boldsymbol{u})$ reaches a certain cutoff value. This is particularly important in many environmental applications and is given by

$$\text{Prob}\{Z(\boldsymbol{u}) > b\,|\,(n)\} = 1 - F(\boldsymbol{u};b\,|\,(n)) \tag{3}$$

In several decision-making processes, maps generated by the probability (3) are enough to outline areas in which remediation measures must be taken. Although the probabilistic evaluation of the uncertainty, using ccdf, represents a criterion for making a decision, the measures of local uncertainty, in general, should be supplemented by an estimate $z^*(\boldsymbol{u})$ of the unknown value. That is why the decision process is rarely based only on probability.

The estimate $z^*(\boldsymbol{u})$ of an unknown value $z(\boldsymbol{u})$ generally has a nonzero error given by $e(\boldsymbol{u}) = z^*(\boldsymbol{u}) - z(\boldsymbol{u})$. In estimates of a toxic concentration, underestimating a concentration may threaten health or result in complaints; overestimating a concentration may incur unnecessary costs and squander cleanup resources. This means that

* Corresponding author. 13506-900 Rio Claro, SP, Brazil. Tel.: +55 19 35262818.
E-mail address: plandim@rc.unesp.br (P.M.B. Landim).

0375-6742/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.gexplo.2008.09.005

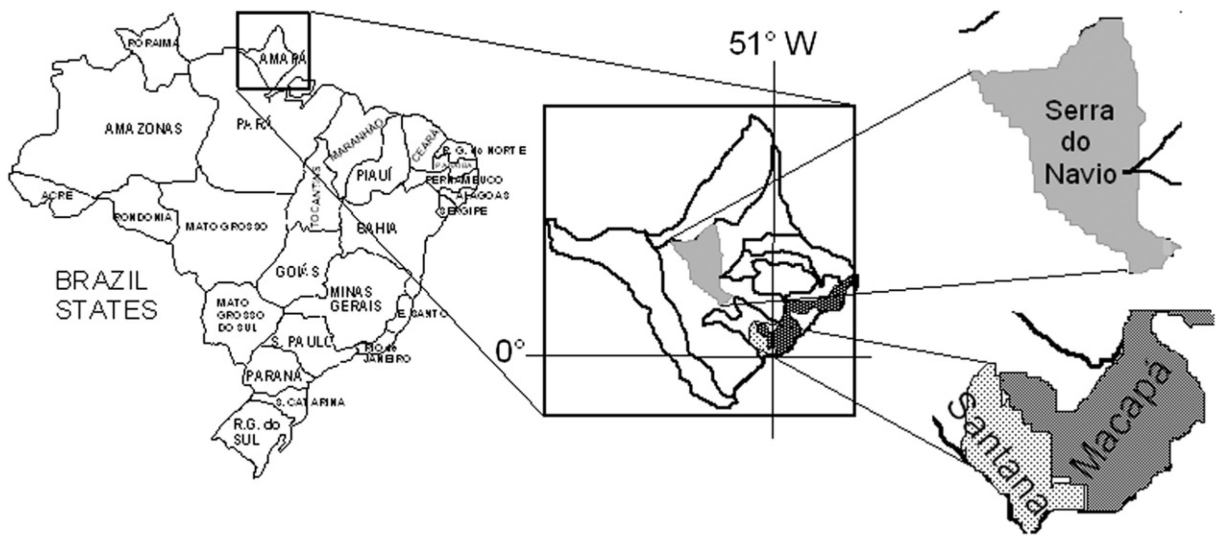

Fig. 1. (a) Amapá State in Northern Brazil; (b) and (c) Municipal districts in Amapá State showing Serra do Navio, where manganese was mined, and Santana, where manganese processing, commercialization, and overseas transport were performed.

an erroneous classification of toxic concentrations will always cause losses; therefore, an error function represents a loss.

To evaluate the impact or loss associated with an error we use a function $L(e(\boldsymbol{u}))$, called a loss function, that provides a criterion for minimizing errors in the classification. The classification is considered robust if the expected loss is at its minimum. Although the argument $[z^*(\boldsymbol{u})-z(\boldsymbol{u})]$ of the loss function is unknown, the uncertainty of $z(\boldsymbol{u})$ can be modeled by the ccdf, $F(\boldsymbol{u};z|(n))$, which is available.

The following uncertainty model can be used to determine the expected loss:

$$\varphi_L\left(z^*(\boldsymbol{u})|(n)\right) = E\left\{L\left[z^*(\boldsymbol{u}) - Z(\boldsymbol{u})\right]|(n)\right\} = \int_{-\infty}^{\infty} L\left(z^*(\boldsymbol{u}) - z\right)dF(\boldsymbol{u};z|(n)) \quad (4)$$

In practice, the following discrete sum is used:

$$\varphi_L\left(z^*(\mathbf{u})|(n)\right) \approx \sum_{k=1}^{K+1} L\left(z^*(\mathbf{u}) - \bar{z}_k\right)\left[F(\mathbf{u};z_k|(n)) - F(\mathbf{u};z_{k-1}|(n))\right] \quad (5)$$

where z_k, $k=1,\ldots,K$, are K cutoff values that discretize the variation values of z; by convention, $F(\boldsymbol{u};z_0|(n))=0$ and $F(\boldsymbol{u};z_{k+1}|(n))=1$. In addition, $\bar{z}_k$ is the class mean $(z_{k-1}, z_k]$, which depends on the interpolation model inside the classes; for example, in the simple case of the linear model, $\bar{z}_k=(z_{k-1}+z_k)/2$.

The optimum estimate for the loss function L is the value of $z^*(\boldsymbol{u})$ that minimizes the expected loss; that is, $z_L^*(\boldsymbol{u})$ is the value of $z^*(\boldsymbol{u})$ that minimizes $L(z^*(\boldsymbol{u})-z(\boldsymbol{u}))$. To calculate the optimum estimate, first the uncertainty of the unknown value, $z(\boldsymbol{u})$, is modeled by ccdf. Then we deduce an estimate for z^*_L according to a specific criterion for optimization. Thus, for a particular uncertainty model, different estimates are obtained depending on the loss function chosen. The quality of an estimate depends on its use; there is no one optimum estimate that is best for all purposes (Journel, 1987; Srivastava, 1987).

Optimum estimates for the criterion of the quadratic average deviation, or for least squares, are related to a function of quadratic symmetrical loss. In contrast, optimum estimates for the absolute average criterion are related to functions of symmetrical and asymmetric loss. All these criteria were used in the present work to extend estimates and to identify and classify areas that show indications of contamination from manganese and arsenic. Thus, it was possible to evaluate the optimum estimate and, consequently, to classify more adequately the variables in the study. Functions using specific loss allowed us to determine health costs and to monitor remediation in the studied area.

2. General aspects of the study area

The study area is located in the district of Santana, Amapá State, in the extreme north of Brazil, approximately between 50° and 55° W and 0° to 5° N (Fig. 1). Santana City has a population of 80,439, the second largest in the state (IBGE, 2000). In 1953, following the discovery of high-quality manganese in Serra do Navio, about 200 km from the state's capital Macapá, the Ore Trade Industry Inc (ICOMI) was established to exploit and commercialize the ore. To carry out the mining, ICOMI constructed a residential community near the manganese mines in Serra do Navio; the result was Santana, a community having a complete infrastructure including sanitation, a recreation facility, schools, supermarket, hospital, and housing for the company's employees and their families, in addition to the industrial installation.

The industrial Santana area covers approximately 129 ha and was planned to be strictly for industrial purposes. The area was basically used to stock manganese and iron ores, products (pellets/sinter and alloys), and raw materials (fuel, coke, etc.) that arrived and departed through the ICOMI port and rail terminal (PCA/Environmental Control Plan, 2001). Manganese and chromite ore were transported by railway from the Serra do Navio mines to the ICOMI industrial area in the Port of Santana, a distance of approximately 200 km.

Geologically, the studied area, which covers the perimeter of ICOMI, is over sediments of the Barreiras Formation consisting of silty organic clays, clay silts, and hard clay with occasional intercalations of fine and coarse sand. The hydrology is important to the physical landscape and the local economy. Santana Port fronts the Northern

Table 1
Maximum concentrations allowed (in mg/L) in drinking water (CONAMA), the minimum and maximum measured values, and the % of measured values above the allowed maximum.

Element	CONAMA	Min	Max	% Pollution
Iron (Fe)	0.300	0.000	85.41	36.6
Manganese (Mn)	0.100	0.002	51.44	27.0
Arsenic (As)	0.050	0.000	22.92	9.80
Aluminum (Al)	0.100	0.008	21.78	75.5
Selenium(Se)[2]	0.010	0.000	0.041	46.3
Lead (Pb)[2]	0.030	0.008	0.04	4.90
Copper (Cu)[2]	0.020	0.000	0.03	4.90
Cadmium (Cd)[1]	0.001	0.000	0.003	7.30

1 (×0.001 mg/L).
2 (×0.01 mg/L).