UNCERTAINTY ANALYSIS FOR GEOSPATIAL MODELLING IN ENVIRONMENTAL RESEARCH - A CASE STUDY OF PREDICTING EMISSION OF AIR POLLUTANT FROM NATURAL SOIL

Phuong N. Truong[1]

[1]Department of Environment and Natural Resource Management
College of Environment & Natural Resources, Campus II, 3/2 Street, Ninh Kieu, Can Tho, Viet Nam
Email: ngocphuong@ctu.edu.vn

ABSTRACT

Environmental research has focused on monitoring, assessing and predicting the impact of environmental agents on human beings and animals. The study of spatial distribution of the environmental agents requires the use of geospatial analysis and modelling. The outputs from the geospatial analysis and modelling are prone to possible errors in the model inputs and model parameters. The uncertainty in the outputs of the geospatial analysis and modelling should be quantified and provided for decision makers to effectively make choices or decisions related to, for example, mitigation or treatment that can have strong effects on human and animals.

In the context of this research, uncertainty is defined as an interval around a value such that any repetition of estimating this value will produce a new result that mainly lies within this interval. Different sources of uncertainties in geospatial modelling can be categorized into four main sources: (1) Input uncertainty; (2) Model parameter uncertainty; (3) Model structure uncertainty and (4) Model solution uncertainty. In this research, the questions of how to quantify model input, model parameter uncertainties and their propagation through environmental models were addressed.

The Monte Carlo uncertainty analysis method was used in this research. The idea of the Monte Carlo method is to repeatedly compute results of the model, with inputs that are randomly sampled from their probability distributions. These inputs can be the model inputs and/or the model parameters and/or error in the model structure. The model outputs form a random sample of the output probability distribution. Analysing this sample distribution by computing its mean and its standard deviation represents the level of uncertainty about model outputs, provided that the sample is large enough.

A case study of using a spatial linear regression model to predict the emission of air pollutant, i.e. soil nitrous oxide was used to illustrate the application of the Monte Carlo method to quantify uncertainty propagation to the prediction outcomes. The linear regression model calculates soil nitrous oxide emission as a function of many factors, including climate variables (e.g. monthly precipitation, minimum temperature), water-pH, soil organic carbon content, nitrogen deposition and vegetation types. The main results of this case study indicate that: (1) The developed statistical models are sufficient to quantify uncertainty about the model inputs and model parameters; (2) Uncertainty about nitrous oxide estimate is expressed by the standard deviation of the prediction outcomes that varies over the study area; (3) Uncertainty in the regression model is the most important source of error that propagates to the uncertainty in the prediction of soil nitrous oxide emission.

1. INTRODUCTION

Environmental research has focused on monitoring, assessing and predicting the impact of environmental agents on human beings and animals. The study of spatial distribution of the environmental agents requires the use of geospatial analysis and modelling. The outputs from the geospatial analysis and modelling are often prone to errors in the model inputs and model parameters. Errors of the model inputs and the model parameters can come from the combined effects of measurement errors, sampling, interpolation and rescaling errors (Brown & Heuvelink, 2007). The resulting errors in the outputs cause uncertainty about the spatial distribution and hence, about the adverse impact of a certain environmental agent. This uncertainty should be quantified and provided for decision makers to effectively make choices or decisions related to, for example, mitigation or treatment strategies that can have strong effects on human and animals.

Several approaches have been used for uncertainty quantification in geospatial analysis and modelling, depending on the aim of the study, the various types of spatial datasets and their attributes (Longley *et al.*, 2006; Attoh-Okine and Ayyub, 2005). For example, Linkov and Burmistrov (2003) investigate model uncertainty that includes the problem formulation, the model implementation and the parameter selection by comparing the outcomes from different models developed for the same environmental agent, i.e. radionuclide concentration. A probabilistic framework of which probability distribution function (pdf) is used for uncertainty representation about environmental variables is recommended by Heuvelink *et al.* (2007). Recently, experts' judgements have been also used for uncertainty quantification of spatial variation of soil properties (Truong and Heuvelink, 2013).

In this research, a probabilistic framework for uncertainty representation was applied, of which uncertainty is defined as an interval around a value such that any repetition of estimating this value will produce a new result that mainly lies within this interval. Different sources of uncertainties in geospatial modelling can be categorized into four main sources: (1) Input uncertainty; (2) Model parameter uncertainty; (3) Model structure uncertainty and (4) Model solution uncertainty (Heuvelink, 1998). The questions of how to quantify model input, model parameter uncertainties and their propagation through environmental models using Monte Carlo method were addressed in this study.

2. METHODS

The idea of the Monte Carlo method is to repeatedly compute results of the model, with inputs that are randomly sampled from their probability distributions. These inputs can be the model inputs and/or the model parameters and/or error in the model structure. The model outputs form a random sample of the output probability distribution. Analysing this sample distribution by computing its mean and its standard deviation (std) represents the level of uncertainty about model outputs, provided that the sample is large enough. Figure 1 outlines the three main steps that were followed in uncertainty propagation analysis using Monte Carlo method.

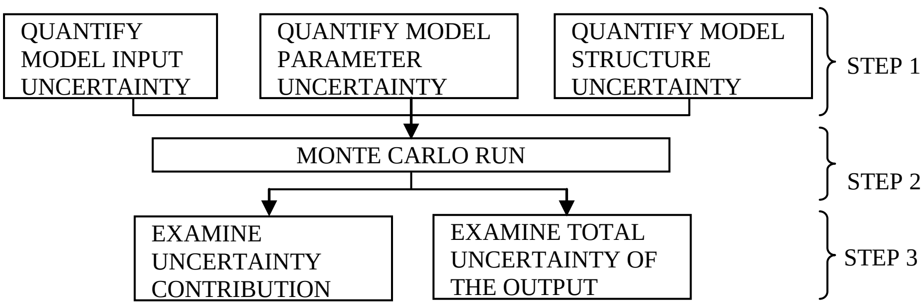

Figure 1. Three main steps of Monte Carlo uncertainty analysis

3. CASE STUDY

To illustrate the application of the Monte Carlo method, a case study of using spatial linear regression model to predict the emission of air pollutant, i.e. soil nitrous oxide (N_2O) from natural areas in Sweden was selected. A sub-model of the European INTEGRATOR model (Reis *et al.*, 2007) was used to predict N_2O soil emission (kgN/ha/yr) from those natural areas. The linear regression model (Equation 1) calculates soil nitrous oxide emission as a function of many environmental factors, including climate variables (e.g. monthly precipitation, minimum temperature), water-pH, soil organic carbon content, nitrogen deposition and vegetation types.

$$\mathrm{LogN_2O} = \beta_0 + \beta_1 {*} \mathrm{P} + \beta_2 {*} (\mathrm{Fraction\ T<0}) + \beta_3 {*} \log(\mathrm{Ndep}) + \beta_4 {*} (\mathrm{pH}) + \beta_5 {*} \mathrm{Organic\text{-}C} + \beta_6 {*} (\mathrm{vegetation{:}\ deciduous}) \quad (1)$$

Where: β_i are regression coefficients; climate variables: P is mean monthly precipitation over the measuring period (mm) and fraction T<0 = fraction of months with minimum T<0°C over the measuring period; Soil variables as derived from WISE & SPADE databases, averaged over the 0-20 cm layer: pH is pH-H_2O (all different pH values are converted to pH-H_2O using regressions) and organic C is OC content (g/kg soil); Deposition variable as derived from European Monitoring and Evaluation Programme (EMEP): Ndep is Nitrogen deposition (kgN/ha.yr); Vegetation types include: deciduous forest, coniferous forest, short vegetation (incl. heath and grass) and mixed woodland. Regression kriging (Hengl *et al.*, 2007) was use for mapping and quantifying uncertainty about the model inputs: pH and OC; and simple kriging was used for the regression residual. The uncertainty about the regression coefficients was quantified by the pdf. Uncertainty propagation analysis was done via Monte Carlo of which the uncertainty contribution of each parameter was calculated.

4. RESULTS AND DISCUSSION

4.1. pH uncertainty

Figure 2 shows the simulated map of pH and the std of pH value over the natural areas in Sweden. The pH value over those areas falls in the acidity range that reaches its maximum value of 6 at some small areas. The std value is higher (from about 0.7 to 1.0) in the southern areas.

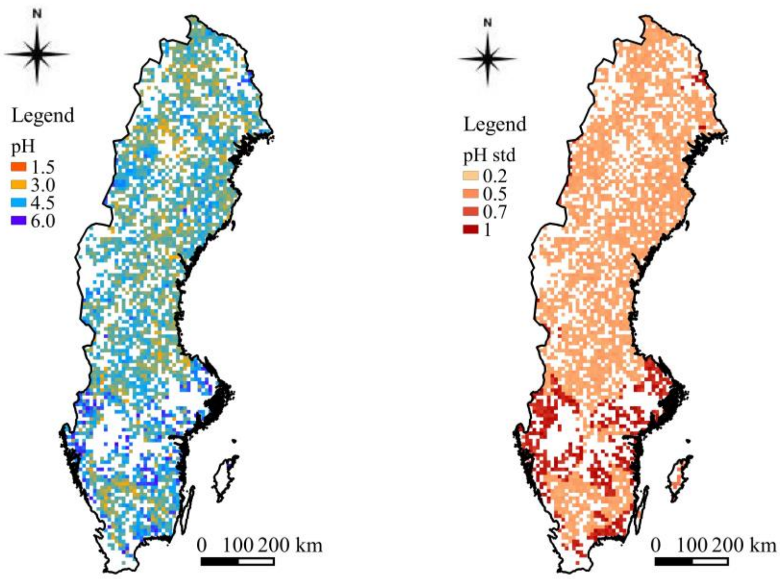

Figure 2. pH simulated map (left) and standard deviation (right)

4.2. OC uncertainty

Simulated map of organic C and the std of organic C values over the natural areas in Sweden are shown in Figure 3. The organic C is high in the western part of the areas, reaching its maximum value of about 55 g/kg soil. The std value that indicates the uncertain level about OC value fairly varies across the natural areas; its value ranges from about 7 to 32 g/kg soil.

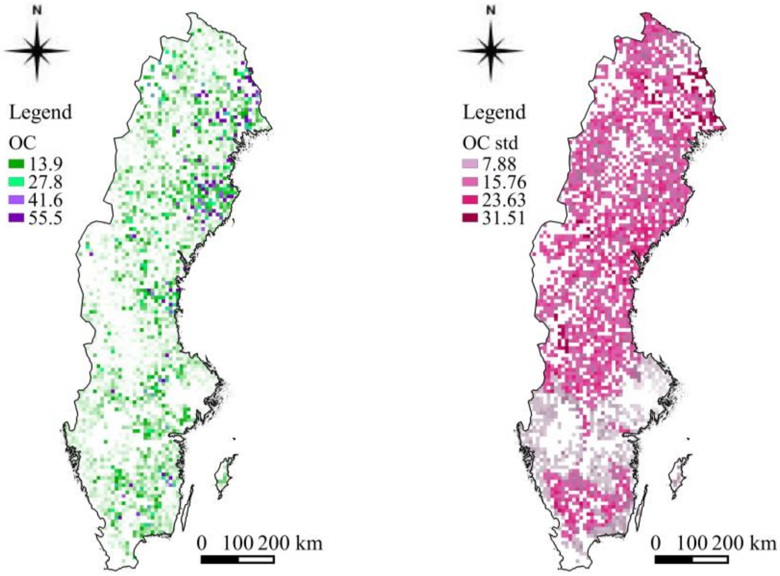

Figure 3. OC simulated map (left) and standard deviation (right)

4.3. Model parameter uncertainty

Equation 2 presents the fitted linear regression model that describes the relationship between $LogN_2O$ and other variables (Section 3). This model was calibrated by the data cover whole European member countries.

$$LogN_2O = 0.133 - 0.007*P + 1.537*(FractionT<0) + 0.497*log(Ndep) - 0.137*(pH) - 0.003*(Organic\ C) + 0.056*(vegetation: deciduous),\ Adjusted\ R^2: 0.2 \quad (2)$$

Table 1 indicates the estimated values of all regression coefficients and their std that measures the uncertainty about their estimated values. Amongst all of the coefficients, the intercept and the vegetation variable have high standard deviation in comparison with their mean.

Table 1. Estimated regression coefficients and parameters of their variation

Coefficients	Estimates	Std.dev from fitting	Coefficient of variation
Intercept	0.133	0.563	4.215
Precipitation	-0.007	0.003	0.429
FractionT<0	1.537	0.397	0.258
logNdep	0.497	0.290	0.584
pH	-0.137	0.062	0.448
OC	-0.003	0.001	0.292
Veg:dec	0.056	0.107	1.908

Figure 4 shows the value of the regression residual that measures the error in predicting N_2O emission using the model in equation 2. The std map shows the uncertain level of the regression residual, with its range from 0.1 to 0.5 kgN/ha/yr.

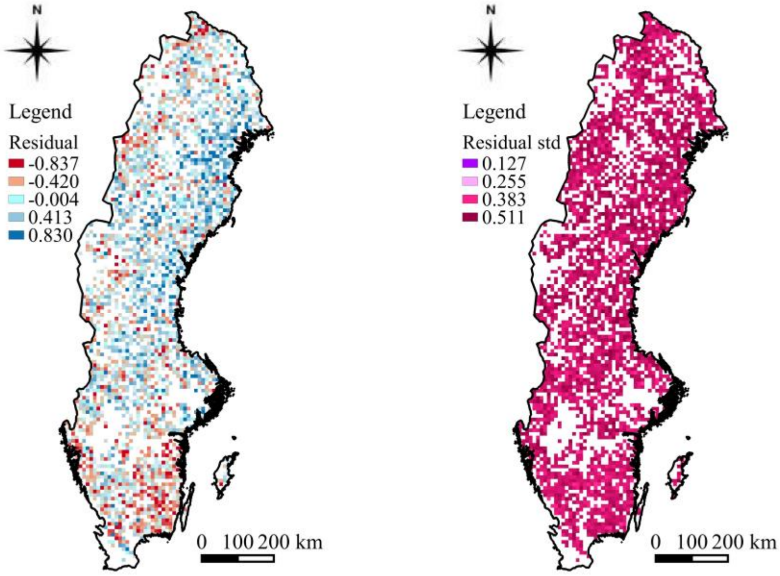

Figure 4. Simulated map of regression residual (left) and standard deviation (right)

4.4. Prediction uncertainty

The map of the prediction std (Figure 5 right) indicates high degree of dispersion of N_2O estimate outputs from its mean value (Figure 5 left). N_2O prediction varies from around 0.5 to around 3 kg N/ha/yr at all examined locations in Swedish natural areas. N_2O estimate generally has high uncertain level because of its high std value compared with its mean value.

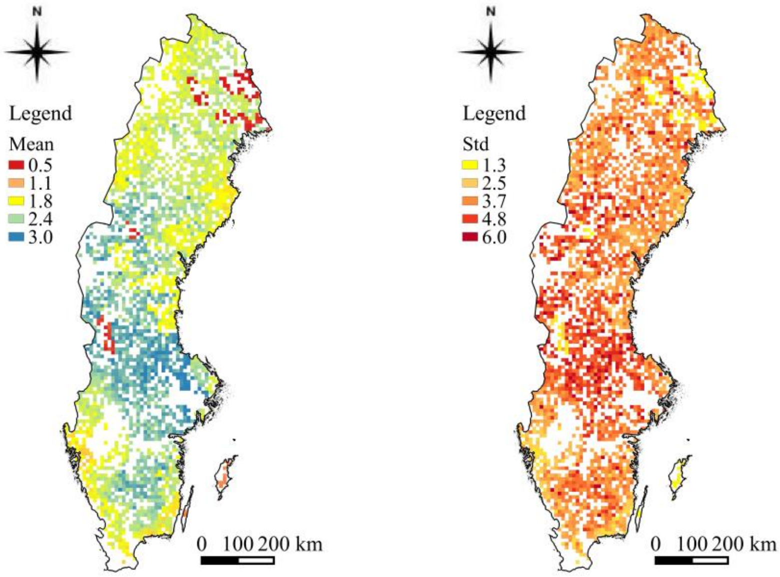

Figure 5. Simulated map of N_2O emission (left) and standard deviation (right)

In Figure 6, we can see that the regression model has the largest uncertainty contribution to the overall prediction uncertainty, compared with those of pH and OC.

Average contribution of the regression model is about 77%. pH had the smallest contribution of about 1% to the total uncertainty in N_2O predictions.

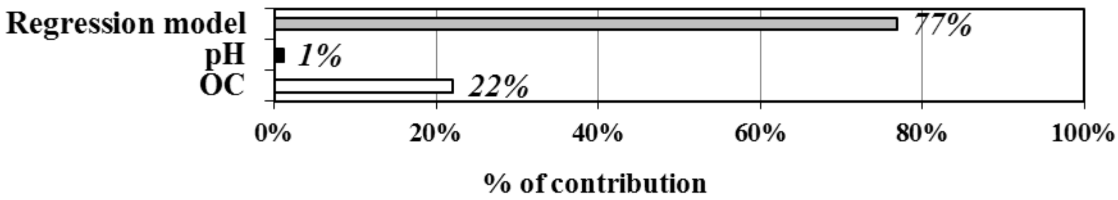

Figure 6: Average relative uncertainty contribution

5. CONLUSION

In this study, quantification of uncertainty about the model inputs and the model parameters was illustrated following the probabilistic framework of which the uncertainty is represented by a probabilistic interval. Uncertainty propagation analysis has been was done using the Monter Carlo method that is simple in its principle but very effective.

The main results of the case study indicate that: (1) The developed statistical models are sufficient to quantify uncertainty about the model inputs and model parameters; (2) Uncertainty about nitrous oxide estimate is expressed by the std of the prediction outcomes that varies over the study area; (3) Uncertainty in the linear regression model (i.e. the model parameter uncertainty) is the most important source of error that propagates to the uncertainty in the prediction of soil nitrous oxide emission.

6. REFERENCES

Attoh-Okine, N.O., Ayyub, B.M., 2005. *Applied Research in Uncertainty Modeling and Analysis*. Springer US, Boston, MA.

Brown, J. D., Heuvelink, G. B. M., 2007. The Data Uncertainty Engine (DUE): A software tool for assessing and simulating uncertain environmental variables. *Computers & Geosciences* 33, 172–190, doi:10.1016/j.cageo.2006.06.015.

Hengl, T., Heuvelink, G.B.M., David, G.R., 2007. About regression-kriging: From equations to case studies. *Computers & Geosciences* 33, 1301-1315.

Heuvelink, G.B.M. 1998. *Error propagation in environmental modelling with GIS*. London, Taylor & Francis.

Heuvelink, G.B.M., Brown, J.D., van Loon, E.E., 2007. A probabilistic framework for representing and simulating uncertain environmental variables. *International Journal of Geographical Information Science* 21, 497–513.

Linkov, I., Burmistrov, D., 2003. Model Uncertainty and Choices Made by Modelers: Lessons Learned from the International Atomic Energy Agency Model Intercomparisons. *Risk Analysis* 23, 1297–1308, doi:10.1111/j.0272-4332.2003.00402.x.

Longley, P.A., Goodchild, M.F., Maguire, D.J., Rhind, D.W., 2005. *Geographic Information systems and science*, 2nd edition. New York: Wiley, Chapter 6.

Reis, S., Sutton, M. A., de Vries, W., Kros, H., 2007. Modelling multidimensional problems – The case of integrated nitrogen management. In: Oxley, L. and Kulasiri, D. (eds). MODSIM 2007, International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand, December 2007, 2083-2089.

Truong, P.N., Heuvelink, G.B.M., 2013. Uncertainty quantification of soil property maps with statistical expert elicitation. *Geoderma* 202–203, 142–152.