Accounting for spatial sampling effects in regional uncertainty propagation analysis

Gerard B.M. Heuvelink, Dick J. Brus and **Gertjan Reinds**
Environmental Sciences Group
Wageningen University and Research Centre
Wageningen, the Netherlands
gerard.heuvelink@wur.nl

Abstract— **Spatial uncertainty propagation analysis (UPA) aims at analysing how uncertainties in model inputs propagate through spatial models. Monte Carlo methods are often used, which estimate the output uncertainty by repeatedly running the model with inputs that are sampled from their probability distribution. Regional application of UPA usually means that the model output must be aggregated to a larger spatial support. For instance, decision makers may want to know the uncertainty about the annual nitrate leaching averaged over an entire region, whereas a model typically predicts the leaching for small plots. For models without spatial interactions there is no need to run the model at all points within the region of interest. A sufficiently large sample of locations may represent the region sufficiently well. The reduction in computational load can then be used to increase the number of Monte Carlo runs, which decreases the Monte Carlo sampling error. In this paper we analyse how a combination of analytical and numerical methods can be used to evaluate the errors introduced by Monte Carlo and spatial sampling. This is important to be able to correct for the bias inflicted by the spatial sampling, to determine how many model runs are needed to reach sufficiently accurate results and to determine the optimum ratio of the Monte Carlo and spatial sample sizes. Results are briefly illustrated with an UPA of a linear regression model that predicts the terrestrial nitrous-oxide emission for Europe.**

Keywords: spatial uncertainty analysis, sampling error, Monte Carlo error, optimization, spatial aggregation

I. INTRODUCTION

When spatial data are inaccurate, the results of spatial analyses that use these data as input will be inaccurate too. The awareness that uncertainty propagates through spatial analyses and can lead to wrong decisions has triggered much research on spatial accuracy assessment (e.g. Heuvelink, 1998; Mowrer and Congalton, 2000; Heuvelink and Burrough, 2002; Shi et al., 2002; Saltelli et al., 2004; Zhang and Goodchild, 2008). The often used Monte Carlo method estimates the propagation of uncertainty by repeatedly running the model with inputs that are sampled from their probability distribution. The method has many appealing properties, among others that it can be easily implemented and can deal with any type of model. It can also reach an arbitrary level of accuracy, by using a sufficiently large number of Monte Carlo runs. The main disadvantage of the method is that it is computationally demanding. Particularly for complex spatial models, for which a single model run is computationally expensive, a Monte Carlo uncertainty propagation analysis (UPA) may become prohibitive. Efficiency can be improved by clever sampling from the input probability distribution using efficient sampling techniques, such as Latin hypercube sampling (LHS). However, the spatial extension of LHS involves approximation errors (Pebesma and Heuvelink, 1999) and the computational load remains large even with efficient implementations.

Many environmental models involve spatial interactions. Examples are erosion, groundwater flow and plant dispersal models. However, there are also many environmental models that are essentially point-based. For instance, models that predict crop growth, greenhouse gas emission, soil acidification or evapotranspiration at some location typically use soil, landuse, management and climate input data at that same location only (e.g. Kros et al., 1999; Earls and Dixon, 2008; Li et al., 2009; Qiu et al., 2009). In regional applications of point models, where the interest is in spatial averages of the model output, the computational load of the Monte Carlo method may be substantially reduced by applying the method to only a (small) sample of locations in the study area. This saves tremendously on computational resources, at the expense of introducing a sampling error. The aims of this paper are to assess the sampling error, correct for the associated sampling bias, and decide how large the spatial and Monte Carlo samples should be to obtain sufficiently accurate UPA results.

II. MONTE CARLO UNCERTAINTY PROPAGATION AND SPATIAL AGGREGATION

Regional application of an UPA typically includes a spatial aggregation step. This step is needed when models produce output at a spatial support that is smaller than the support at which the final result is required. For instance, decision makers may want to know the uncertainty about the annual greenhouse gas emission averaged over entire countries, whereas a model may predict the emission on a daily basis for plots that are smaller than one hectare. In such a case the model outputs of the individual Monte Carlo runs are aggregated to the target support before the uncertainty analysis continues. The example above involves both spatial and temporal aggregation, but in this paper we focus on spatial aggregation only. Thus, we address the case in which the model produces output at 'points' (i.e., areas that have

negligible support compared to the extent of the study area), while results are needed at the much larger 'block' support. The block might be a grid cell or region within the study area, or the study area itself. Let the ratio of the block and point support be given by M, where M can be extremely large. In fact, M will be infinite when the point support is infinitesimally small.

The Monte Carlo method estimates the uncertainty in the block-averaged model output as follows (Heuvelink and Pebesma, 1999):

- Repeat n times:

1. Use a (pseudo-)random number generator to generate a realization of the uncertain model inputs for all points in the block, while taking spatial and cross-correlations into account.
2. Run the model with the simulated inputs for all points, average the output over the block, and store the result.

- Analyse the n block-support outputs by computing summary statistics, such as the mean, standard deviation, percentiles and a histogram..

Note that the procedure above requires that the model is run $n \times M$ times. Here, n is the number of Monte Carlo runs, which must be chosen sufficiently large to reach sufficiently accurate results. However, there will be a Monte Carlo error because n is finite. The variance of the Monte Carlo error typically decreases proportional to the number of Monte Carlo runs (Heuvelink, 1998). In practice, n must often be chosen at least as large as 200, but in specific cases it may need to be much greater than that.

M equals the number of points within the block. To reduce computation time, it may be sensible to run the Monte Carlo analysis for only a subset (sample) of m points ($m << M$). Indeed, when the point support is effectively zero and M is infinite, a sample (such as the nodes of a dense spatial grid) must be used. Running the UPA for only a subset of m points will substantially reduce computing time and storage requirements, so that the number of Monte Carlo runs n may be increased. The price paid is a sampling error. The net result of introducing a sampling error and decreasing the Monte Carlo error may well be that a more accurate assessment of output uncertainty is achieved. Thus, ideally one chooses n and m such that the combined error is the smallest for a given maximum number of model runs $n \times m$.

Kros et al. (1999) analysed uncertainty propagation in a soil acidification model and used m=25 (5×5 km^2 blocks represented by 25 points located on a 1×1 km^2 grid) in combination with n=625 Monte Carlo runs. Heuvelink et al. (2010) represented the whole of the Netherlands with m=258 points, and executed a UPA for a pesticide leaching model using n=1,000 Monte Carlo runs. In neither of these two studies was a thorough assessment made of the trade-off between the sampling and Monte Carlo errors. In fact, the sampling error was not calculated and thus effectively ignored. In order to judge whether the sampling error is indeed small and has negligible bias, it must first be calculated. This will be done in the next section.

III. Evaluation of the Aggregated Output Variance

A. *Analytical Expression for the Output Variance*

Let the model input be denoted by $U(x)$ ($x \in B$), where x refers to location and where B is the block. Note that $U(x)$ is a vector in case the model has multiple inputs. Let the output be given by $Y(x)$, which is computed from the input $U(x)$ by running the model g:

$$Y(x) = g(U(x)) \tag{1}$$

To acknowledge that the model input is uncertain (hence stochastic) we write it in upper case. As a result, the output is also stochastic. Next the output is aggregated over B by defining its mean:

$$\overline{Y} = \frac{1}{M}\sum_{i=1}^{M} g(U(x_i)) \tag{2}$$

The goal of the UPA is to quantify the uncertainty about $\overline{Y}$. For this we take the variance as a measure:

$$V(\overline{Y}) = E[(\overline{Y} - \mu_{\overline{Y}})^2] \tag{3}$$

where $\mu_{\overline{Y}} = E[\overline{Y}]$ is the mean of $\overline{Y}$.

Both the mean and variance of $\overline{Y}$ can only be estimated because we use a finite number of Monte Carlo runs and a sample size m out of the total of M locations in B. Let us assume that the sample of m point locations in block B is chosen with simple random sampling. Thus, the sample mean is an unbiased predictor of $\overline{Y}$:

$$E_p[\hat{\overline{Y}}] = \overline{Y} \tag{4}$$

where E_p, the p-expectation, means averaging over a large number of spatial samples drawn according to the simple random spatial sampling design (De Gruijter et al., 2006, chapter 2), and where:

$$\hat{\overline{Y}} = \frac{1}{m}\sum_{i=1}^{m} g(U(X_i)) \tag{5}$$

Note that the locations are now random too and hence written in upper case.

With these results, Eq. (3) can be written as:

$$V(\overline{Y}) = V_\xi(E_p[\hat{\overline{Y}}]) \tag{6}$$

where we have introduced subscript ξ to clarify that the variance is taken over a large (infinite) number of realizations of the random function Y (De Gruijter et al., 2006, chapter 2). It is important to distinguish between the stochasticity introduced by the uncertain model input and that introduced by the spatial sampling.

Using a well-known decomposition result (Cochran, 1977, Eq. (10.2)), we can now derive:

$$V(\overline{Y}) = V_{\xi p}(\hat{\overline{Y}}) - E_{\xi}[V_p(\hat{\overline{Y}})] \quad (7)$$

This expression is useful because it transforms the variance of the unknown $\overline{Y}$ into means and variances of $\hat{\overline{Y}}$, which can be numerically evaluated. Note also that the second term on the right-hand side of Eq. (7) is the expected sampling variance, which quantifies the spatial sampling error. Ideally it is small relative to the variance of $\overline{Y}$. This can be achieved by choosing m sufficiently large.

B. *Numerical evaluation of the output variance*

The variance of $\overline{Y}$ can now be estimated by numerical evaluation of the two terms on the right-hand side of Eq. (7). The first term can be estimated as follows:

1. Select m sampling locations with simple random sampling.
2. Draw a realization u from the input U at the m locations (taking spatial and cross-correlations into account).
3. Compute the model outputs at the sampling locations.
4. Take the average of the m model outputs, yielding an estimate $\hat{\overline{y}}$.
5. Repeat steps 1 to 4 n times, yielding $\hat{\overline{y}}_i$, $i = 1...n$.
6. Compute the variance of the n estimates $\hat{\overline{y}}_i$.

The second term on the right-hand side of Eq. (7) can be estimated as follows:

1. Select m sampling locations with simple random sampling.
2. Draw a realization u from the input U at the m locations (taking spatial and cross-correlations into account).
3. Compute the model outputs at the sampling locations.
4. Compute the variance of the spatial sampling error by dividing the variance of the m model outputs by the sample size m.
5. Repeat steps 1 to 4 n times.
6. Compute the mean of the n sampling error variances.

The algorithms can partly be integrated to improve efficiency. It is important to note that the total number of model runs required is indeed reduced from $n \times M$ to $n \times m$. However, note also that the procedure only yields an estimate of the variance (i.e., uncertainty) of the model output $\overline{Y}$. This is usually the aim of an UPA and hence an UPA would stop after the estimate of $V(\overline{Y})$ is obtained, but the main aim of this work is to quantify the associated estimation error. Therefore, another iteration loop is needed to estimate the accuracy of the estimate of $V(\overline{Y})$.

C. *Quantifying the Accuracy of the Estimated Output Variance*

In order to assess the Monte Carlo and spatial sampling errors, the procedure presented above must be repeated many times. The variance of the so-obtained estimates of $V(\overline{Y})$ characterizes the accuracy of the estimated variance of the model output. Clearly, the accuracy depends on n and m. The larger n, the smaller the Monte Carlo estimation error. The larger m, the smaller the spatial sampling error. Given a restriction on the total number of model runs $n \times m$, there will be a trade-off between n and m. For some combination of n and m the smallest variance will be obtained. The next section calculates the optimum ratio of n and m for different values of $n \times m$ for a simple case study.

IV. Uncertainty Propagation with a Nitrous-Oxide Emission Model

Nitrous-oxide emission from soils in natural ecosystems in Europe was modelled by a multiple linear regression model (Bloemerts, 2007). Among others, the regression model uses the carbon content and pH of the topsoil as inputs. Both soil properties were considered uncertain. Geostatistical models were built and applied to create maps of these soil properties from point observations and auxiliary information (Truong, 2009). Interpolation errors were quantified and realizations of the soil property maps were generated using conditional sequential Gaussian simulation (Goovaerts, 1997).

The propagation of uncertainties was analysed using the Monte Carlo method, whereby the regression model was run n times at m randomly selected locations within the study area (i.e. the natural ecosystem areas within Europe). This was done for four values of $n \times m$ and for seven combinations of n and m for each value of $n \times m$. In addition, for each of the resulting 28 combinations the Monte Carlo analysis was done 1,000 times, in order to compute the accuracy of the estimated variance of the model output.

Results are presented in Figs. 1 and 2. Fig. 1 shows the means of the estimated output variances over the 1,000 repetitions. The estimated mean is centred around 0.0057 (kg N ha^{-1} year^{-1})^2 and is not systematically affected by the ratio of n and m. As expected, results are more stable for larger values of $n \times m$. Deviations from the mean are small in all cases, which is also as expected because these are averages of 1,000 estimated variances.

Fig. 2 shows the standard deviations of the estimated output variance. Three main observations can be made. First, standard deviations are smaller when the total number of model runs increases. Second, the standard deviations are of the same order of magnitude as the mean when the total number of model runs equals 100 or 200 (i.e. compare Fig. 1), indicating that these are too small numbers to obtain reliable estimates of the output variance. Acceptable estimates are obtained when $n \times m$=800. Third, the ratio of n and m has a substantial effect on the accuracy obtained. For

instance, taking $n \times m$=200 and n:m=1:1 yields more accurate results than taking $n \times m$=400 and n:m=10:1. Optimum ratios are obtained when n and m are equal, and accuracy steadily decreases as one moves away from the optimum. Interestingly, the optimum seems not to be influenced by the total number of model runs $n \times m$.

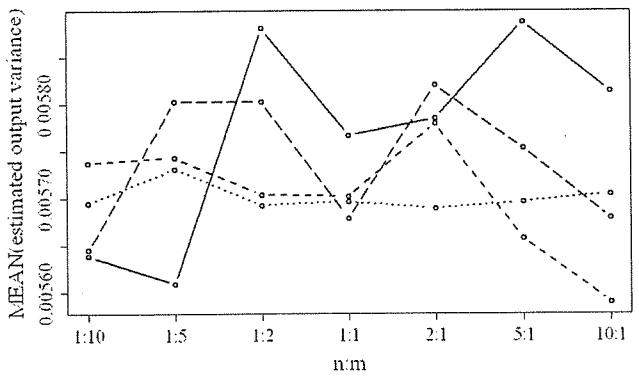

Figure 1. Mean of estimated output variance for the nitrous-oxide emission case for different values of the ratio of the number of Monte Carlo runs n and number of sampling locations m. Solid line: $n \times m$=100; long-dashed line: $n \times m$=200; dashed line: $n \times m$=400; dotted line: $n \times m$=800. Measurement units are (kg N ha^{-1} year^{-1})^2.

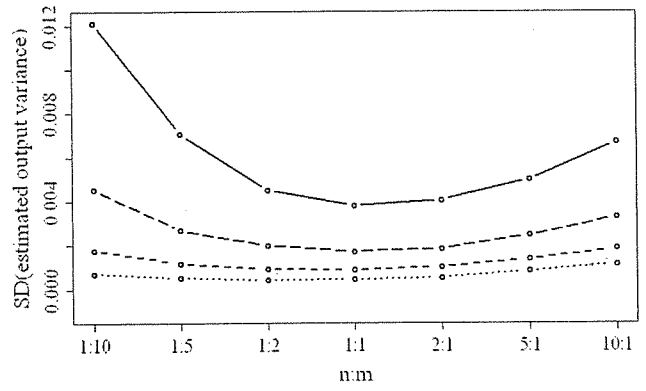

Figure 2. Standard deviation of estimated output variance for the nitrous-oxide emission case for different values of the ratio of the number of Monte Carlo runs n and number of sampling locations m. Solid line: $n \times m$=100; long-dashed line: $n \times m$=200; dashed line: $n \times m$=400; dotted line: $n \times m$=800. Measurement units are (kg N ha^{-1} year^{-1})^2.

V. CONCLUSIONS

This paper presented a method to analyse the propagation of input uncertainty to the spatial average of the output of a point model by running a Monte Carlo analysis at a limited sample of locations only. Unlike previous studies (e.g. Kros et al., 1999; Heuvelink et al., 2010), the method yields an unbiased estimate of the output variance because it corrects for the spatial sampling error. The sampling bias may be small in cases where the study area is represented by a large sample (e.g. a dense grid), but verification is important and can fairly easily be achieved with numerical evaluation of the expression given in Eq. (7). Moreover, the methodology presented here can help choose the optimum ratio of the Monte Carlo and spatial sample sizes and thus help avoid that a too large, inefficient spatial sample is used.

Theoretical results were illustrated with a simple case study. Many more case studies are needed to analyse how results vary in different cases.

ACKNOWLEDGEMENTS

This work was funded by the European Commission DG Research, integrated project NitroEurope (6th Framework nr 017841) and collaborative project iSOIL (7th Framework nr 211386).

REFERENCES

Bloemerts, M. (2007). *Nitrous oxide emissions from natural ecosystems on a European scale.* Internship report, Alterra, Wageningen, the Netherlands.

Cochran, W.G. (1977). *Sampling techniques.* Third edition. New York: Wiley.

De Gruijter, J.J., D.J. Brus, M.F.B. Bierkens and M. Knotters (2006). *Sampling for natural resource monitoring.* Berlin: Springer.

Earls, J. and B. Dixon (2008). A comparison of SWAT model-predicted potential evapotranspiration using real and modeled meteorological data. *Vadose Zone Journal* 7, 570–580.

Goovaerts, P. (1997). *Geostatistics for natural resources evaluation.* New York: Oxford University Press.

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

Heuvelink, G.B.M. and P.A. Burrough (Eds.) (2002). Developments in statistical approaches to spatial uncertainty and its propagation. *International Journal of Geographical Information Science*, 16(2).

Heuvelink, G.B.M. and E.J. Pebesma (1999). Spatial aggregation and soil process modelling. *Geoderma* 89, 47–65.

Heuvelink, G.B.M., F. Van Den Berg, S.L.G.E. Burgers and A. Tiktak (2010). Uncertainty and stochastic sensitivity analysis of the GeoPEARL pesticide leaching model. *Geoderma* 155, 186–192.

Kros, J., E.J. Pebesma, G.J. Reinds and P.F. Finke (1999). Uncertainty assessment in modelling soil acidification at the European scale: a case study. *Journal of Environmental Quality* 28, 366–377.

Li, T., Y.S. Feng and X.M. Li (2009). Predicting crop growth under different cropping and fertilizing management practices. *Agricultural and Forest Meteorology* 149, 985–998.

Mowrer, H.T. and R.G. Congalton (Eds.) (2000). *Quantifying spatial uncertainty in natural resources: theory and applications for GIS and remote sensing.* Ann Arbor: Ann Arbor Press.

Truong, P. (2009). *Uncertainty analysis of predicting terrestrial nitrous-oxide emissions in Europe with INTEGRATOR.* MSc thesis Wageningen University, Wageningen, the Netherlands.

Pebesma, E.J. and G.B.M. Heuvelink (1999). Latin hypercube sampling of Gaussian random fields. *Technometrics* 41, 303–312.

Qiu, J.J., C.S. Li, L.G. Wang, H.J. Tang, K. Li and E. Van Ranst (2009). Modeling impacts of carbon sequestration on net greenhouse gas emissions from agricultural soils in China. *Global Biogeochemical Cycles* 23:GB1007.

Saltelli, A., S. Tarantola, F. Campolongo and M. Ratto (2004). *Sensitivity analysis in practice, a guide to assessing scientific models.* New Yoprk: Wiley.

Shi, W., P.F. Fisher and M.F. Goodchild (Eds.) (2002). *Spatial data quality.* London: Taylor & Francis.

Zhang, J. and M.F. Goodchild (Eds.) (2008). *Spatial uncertainty.* Liverpool: World Academic Press.