Monte Carlo and spatial sampling effects in regional uncertainty propagation analyses

Gerard B.M. Heuvelink, Dick. J. Brus

Wageningen University and Research Centre
gerard.heuvelink@wur.nl
dick.brus@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. In this paper we explore 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 accurate results and to determine the optimum ratio of the Monte Carlo and spatial sample sizes.

1 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., [1-6]). 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 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 [7] 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. [8-11]). 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.

2 MONTE CARLO UNCERTAINTY PROPAGATION ANALYSIS 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 [12]:

- Repeat n times:
  Generate a possible reality from the uncertain model inputs for all points in the block, while taking spatial and cross-correlations into account.
  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 or cumulative frequency distribution.

Note that the procedure above requires that the model is run n×M times. Here, n is the number of Monte Carlo runs, which must be chosen sufficiently large to reach sufficiently accurate results. The Monte Carlo error variance typically decreases proportional

with the number of Monte Carlo runs [1]. In practice, n must often be chosen at least as large as 200, but in specific cases it may need to be much larger 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) will always 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 adding 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 would choose m and n such that the combined error is smallest for the given maximum number of model runs n×m.

[10] 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. [13] represented the whole of the Netherlands with m=258 points, and ran an UPA for a pesticide leaching model using n=1,000 Monte Carlo runs. In neither of these two cases 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 negligible, it must first be calculated. This will be done in the next section.

3 EVALUATION OF THE AGGREGATED OUTPUT VARIANCE

3.1 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. Let the model 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 and 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. 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 [14, 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 [14, 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 [15, Eq. (10.2)], we can now derive:

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

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 sampling error. Ideally it is small relative to the variance of $\overline{Y}$. This can be achieved by choosing m sufficiently large.

3.2 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 correlation into account)
3. compute the model outputs at the sampling locations and take their average, yielding an estimate $\hat{\overline{y}}$
4. repeat steps 1 to 3 n times, yielding $\hat{\overline{y}}_i$, i = 1...n
5. 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. draw a spatially exhaustive realization u from the input U
2. select m sampling locations with simple random sampling

compute the model outputs at the m sampling locations and take their average

3. repeat steps 2 to 3 many times and compute the variance of the so-obtained averages
4. repeat steps 1 to 4 n times and compute the mean of the so-obtained variances

The algorithms can be integrated to improve efficiency. However, even then the numerical load can be quite involved (particularly because another iteration loop is needed to analyse the accuracy of the estimate, see next section). Analytical solution of Eq. (7) seems not possible because in general it may not be assumed that Y is stationarity.

4 DISCUSSION AND CONCLUSIONS

Section 3 presented a procedure to estimate the variance of the spatially averaged output of a point model using a Monte Carlo analysis at only a sample of locations. The procedure yields an unbiased estimate of the output variance. In previous studies (e.g. [10, 13]), the sampling error was typically not computed and resulting output variances were therefore biased. 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 the methodology for doing this has been presented in section 3.

The algorithms presented in section 3 yield only an estimate of the variance of the aggregated model output, because only a finite number of Monte Carlo runs are executed. In order to assess the Monte Carlo error, the procedure needs to be repeated many times and the variance of the resulting output variances computed. The larger n, the smaller the Monte Carlo estimation error. However, m plays a role too. Given n, a large m will result in a smaller estimation error than a small m. Thus, there will be an optimum balance between m and n, where the smallest estimation error is achieved given a constraint on the maximum size of the product of m and n. However, computing the optimum balance is extremely computationally demanding, because it requires an extra loop. Alternatively, one may compare the expected sampling variance with the estimated variance of the aggregated model output using Eq. (7). A proper strategy might be to choose m sufficiently large to ensure that the first is small relative to the second, but not make it any greater than that. This needs to be investigated.

The next step obviously is to apply the methodology to practical cases. Currently, we are running the analysis with a terrestrial N_2O emission model for Europe.

5 ACKNOWLEDGEMENTS

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

6 REFERENCES

[1] G.B.M. Heuvelink. Error propagation in environmental modelling with GIS. Taylor & Francis, 1998.

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

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

[4] W. Shi, P.F. Fisher and M.F. Goodchild (Eds.). Spatial data quality. Taylor & Francis, 2002.

[5] A. Saltelli, S. Tarantola, F. Campolongo and M. Ratto. Sensitivity analysis in practice, a guide to assessing scientific models. Wiley, 2004.

[6] J. Zhang and M.F. Goodchild (Eds.). Spatial Uncertainty. World Academic Press, 2008.

[7] E.J. Pebesma and G.B.M. Heuvelink. Latin hypercube sampling of Gaussian random fields. Technometrics 41:303-312, 1999.

[8] T. Li, Y.S. Feng and X.M. Li. Predicting crop growth under different cropping and fertilizing management practices. Agricultural and Forest Meteorology, 149:985-998, 2009.

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

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

[11] J. Earls and B. Dixon. A comparison of SWAT model-predicted potential evapotranspiration using real and modeled meteorological data. Vadose Zone Journal, 7:570-580, 2008.

[12] G.B.M. Heuvelink and E.J. Pebesma. Spatial aggregation and soil process modelling. Geoderma, 89: 47-65, 1999.

[13] G.B.M. Heuvelink, F. Van Den Berg, S.L.G.E. Burgers and A. Tiktak. Uncertainty and stochastic sensitivity analysis of the GeoPEARL pesticide leaching model. Geoderma (accepted), 2009.

[14] J.J. De Gruijter, D.J. Brus, M.F.B. Bierkens and M. Knotters. Sampling for natural resource monitoring. Springer, 2006.

[15] W.G. Cochran. Sampling techniques. Third edition. Wiley, 1977.