The stochastic dimension in a dynamic GIS

Edzer J. Pebesma, Derek Karssenberg and Kor de Jong

Utrecht Centre for Environment and Landscape Dynamics, Faculty of Geographical Sciences, Universiteit Utrecht, P.O. Box 80.115, 3508 TC Utrecht; e.pebesma@geog.uu.nl

Abstract. Coping with random fields in a time-dynamic geographic information system (GIS) increases the computational burden and storage requirements with a large amount, and calls for a number of custom functions to enable easy analysis of the resulting random components, as well as specialised output reporting functions. This paper addresses the computational and implementation issues when a Monte Carlo approach is taken, and shows some results from a rainfall-runoff model running within a GIS.

Keywords. Geographical information systems, Monte Carlo, temporal GIS, stochastic modelling, geostatistics

1 Introduction

Geographical information systems (GIS, Burrough and McDonnel, 1998) liberate the end-user from worrying about looping over all spatial entities by providing a set of functions that operate on maps as a whole. Such functions operate either point-wise, meaning that a result value exclusively depends on the values in a set of maps at the same location, or they operate in a spatial neighbourhood, meaning that values at other spatial locations contribute as well. A point-wise operation can be the addition of two maps,

```
sum.map = a.map + b.map
```

a neighbourhood operation the calculation of distances to a point location:

```
distances.map = distanceto(point.map)
```

Writing GIS operations on maps as algebraic expressions was first proposed by Tomlin and Berry (1979), and they named it 'map algebra'.

Beyond the standard map algebra, PCRaster (Wesseling et al., 1996) extends the set of spatial-only, static functions with a set of functions that have a temporal dimension, by adding a loop over a (discretised) time axis: spatial time series (stacks of maps) as well as aggregates over the iteration period can now be processed and analysed. The thus obtained dynamic GIS (implemented as a concise modelling language) has shown to be an easy tool for fast implementation of a wide range of spatio-temporal processes, ranging from distributed hydrological models to ecological dispersion or urban growth models (cellular automata). Easy control over the modelling process enables users with little expertise in GIS or computer programming to evaluate many similar but alternative models, by simply trying them out.

Although some of the functions provided by PCRaster may contain some form of randomness (e.g. a function returning a map with standard normal *i.i.d.* variates), all functionality regards deterministic computation, meaning that uncertainty with respect to model input variables cannot be handled. In the practice of environmental modelling however, users are often confronted with highly uncertain model inputs, and want to be able to assess how input errors propagate to the output of their GIS models.

Today, freely available geostatistical software tools (e.g. Deutsch and Journel, 1998; Pebesma and Wesseling, 1998) provide the means for a fairly straightforward modelling and simulation of spatial fields for both discrete (nominal) and continuous variables, using the indicator simulation formalism for the former and Gaussian random fields for the latter. Using these maps as input to GIS models is far from easy though, and it may prevent large groups from doing so for that reason.

This paper discusses the concept of a stochastic (dynamic) GIS, and proposes a set of functions that should be added to the GIS to provide error estimates of model output, coming from model input error. In a case study regarding the analysis of stream flow and surface runoff during a rain storm, we analyse how uncertainty on infiltration capacity influences the model output, being spatial distribution of runoff and temporal variation of river flow at the catchment outlet.

2 Stochastic GIS

2.1 A Monte Carlo approach

Spatial operations of a map algebra GIS can be generalised as follows:

$$\{r\} = f(A, B, C, ...) \tag{1}$$

with $\{r\}$ a set of one or more maps where the output is written to, with $f(\cdot)$ a simple or compound operation, and with input maps A, B, C,... (possibly including some constant maps). In a dynamic GIS (1) is evaluated every time step, and the operation may involve an update of some maps for each time step. The challenge in a stochastic GIS is to characterise the joint (that is, multivariable) distribution function of $\{r\}$ over space and time, given the joint distributions of A, B, C,... . Analytic approaches (Heuvelink, 1998) have addressed analytical solutions to (1) for the case where only non-spatials were stochastic. The only feasible approach to cases that include random spatial or spatio-temporal fields appears to be a full spatio-temporal Monte Carlo (MC) simulation (Heuvelink, 1998).

The most general implementation for an MC approach considers evaluation over the stochastic, temporal and spatial domain in the order:

```
for m in MC-Sample
 for t in TimeSteps
   < evaluate r >
```

(evaluation of r trivially involves the looping over all spatial locations). For a full analysis of the results, all output of each MC run has to be stored. This can be a full spatial time series for each MC run, or a single map (e.g. a state variable at the last time step, or a map aggregate over time) or a non-spatial time series (read from a location in a map, or aggregated over space), or an aggregated (scalar) value over either of the latter two.

Moving the MC loop inside the time or space loop may be more efficient in some respects, but only works in highly simplified cases, e.g. in absence of temporal and spatial correlation of the model input.

2.2 The 'stochastic' dimension

Taking the 'randomness' as a new dimension, in the sense of spatial or temporal dimensions, can be justified from both a conceptual and an operational point of view. From a conceptual point of view, geographical space, or the

three-dimensional space where we live in, is captured by traditional GIS (although most emphasis is traditionally put to the two 'horizontal' dimensions). A temporal GIS adds to this the temporal dimension, to capture changes of the spatial settings over time. These four dimensions can thus be used to represent everything we know. To add to this a representation of things we do not know, for instance by probability density functions (PDFs), we need at least one more dimension. It may be argued that *every* point in space/time that needs a PDF adds one dimension, but usually easier representations are sought (for instance by assuming stationarity).

From an operational point of view, the obvious regular discretisation of a 3D-space+time 'block' is a four-dimensional array. Using MC analysis, this array is replicated a number of times equal to the MC sample size. The obvious discretisation of such a stochastic 3D-space/time block is a five-dimensional array: The added index identifies the MC sample element.

The consequence of adding a stochastic dimension is that the modeller has to take one more aspect into account when choosing the spatial discretisation (raster map cell size) and temporal discretisation (time step): the MC sample size. The total computational burden (and, as we will see in the worst case, storage requirements) is of order:

$$\text{<}\textit{map cells}\text{>} \times \text{<}\textit{time steps}\text{>} \times \text{<MC } \textit{sample size}\text{>}$$

A real danger will be that naive users choose a small sample MC size because they prefer high-resolution maps, and end up with highly inaccurate *estimates* of model output distributions.

3 Analysis of the Monte Carlo output

To obtain the complete results from the MC analysis, the full output has to be retained: for instance the cumulative probability density function of r at a given time and location is estimated from the ranked MC sample values at this time and location. Also, this output may serve as input to another function, say $g(\cdot)$:

$$p = g(r, ...)$$

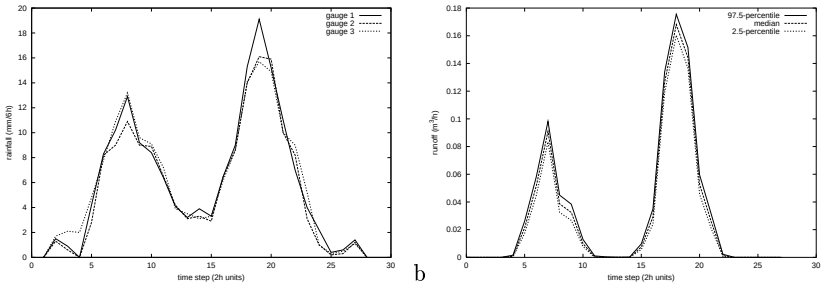
An example of $g(\cdot)$ is spatial aggregation of output of the function $f(\cdot)$, which should be applied to every MC sample element (Heuvelink and Pebesma, 1998; Pebesma and Heuvelink 1999).

To summarise all marginal cumulative probability density functions, at the cost of loosing information of the joint density of r, one can use two approaches that differ strongly with respect to storage requirements:

1. collect a set of percentiles, e.g. the 5-,10-,25-,50-,75-,90-,and 95-percentile,
2. collect the sum, sum of squares, sum of cube squares, ... and the frequency of sample elements being above a set of pre-defined thresholds.

The second option gives the possibility of estimating for each location and time the mean, variance, skewness and kurtosis, and probabilities of exceeding thresholds.

Estimation of the percentiles of a distribution requires that the full set of MC sample elements is stored, because they need to be ranked before percentiles can be estimated. It should be noted here that for very high or low percentiles some gain is obtained by only storing the n tail values, with $n \approx qN$, with q the quantile estimated and with N the MC sample size. In general however, the full output distribution is of interest, eliminating this advantage.



a b

Fig. 1. (a) Rainfall, measured at the three gauges (see Fig. 2b) in mm/6h; (b) runoff at the catchment outlet: 2.5- and 97.5-percentiles and median value in m^3/h

The second group of summary variables has the attractive property that they can be collected *without* storing the complete set of MC sample elements: they can be obtained by tracking all the necessary sums during the MC simulation, and after each MC run the previous results can be discarded.

4 Case study: rainfall-runoff simulation

Rainfall-runoff modelling can be useful for predicting river floods, soil moisture contents, soil erosion, and eventually landscape evolution, in undulating or hilly catchments. A simple rainfall-runoff simulation model for the 0.42 km^2 Catsop catchment (Limburg, The Netherlands) was implemented in PCRaster. It evaluates for each time step: (1) when rainfall intensity exceeds the soil infiltration capacity, the soil becomes saturated and excess rainfall will run off downhill; (2) surface runoff is routed through a drainage network and will either re-infiltrate in subsequent non-saturated cells, or leave the catchment at its outlet.

Rainfall at each location is obtained by reading the rainfall time series data (Fig. 1a) from the nearest rainfall gauge (Fig. 2b) at that moment. The local drain direction map defining the runoff network is derived from the elevation map (Fig. 2a); and infiltration is modelled as a function of soil texture (Fig. 2b). Mean values for infiltration were 2.8 (clay), 8.3 (loam) and 19.0 (sand) [mm/6h]. Standard deviations were taken as one third of the mean value, and the semivariance between infiltration at two sites a distance h ($h > 0$) apart within the same soil texture class was modelled as an exponential variogram

$$\gamma(h) = \sigma^2(s)(1 - 0.9\exp(-h/50)),$$

with $\sigma^2(s)$ the soil-dependent variance (variance and spatial correlation information were obtained from Loague and Kyriakidis, 1997).

Simulation of infiltration capacity was done by sequential Gaussian simulation (Deutsch and Journel, 1998). Negative simulated values were reset to a zero value. To increase the MC sampling efficiency, Latin hypercube sampling (Pebesma and Heuvelink, 1999) was applied. A sample of 1000 infiltration capacity maps was obtained to study the effect of unknown infiltration capacity on spatial patterns of runoff (Fig. 3) and time series of runoff at the catchment outlet (Fig. 1b).

The results show that the uncertainty increases with runoff levels, but that the errors are small compared to the modelled runoff levels. Actual runoff

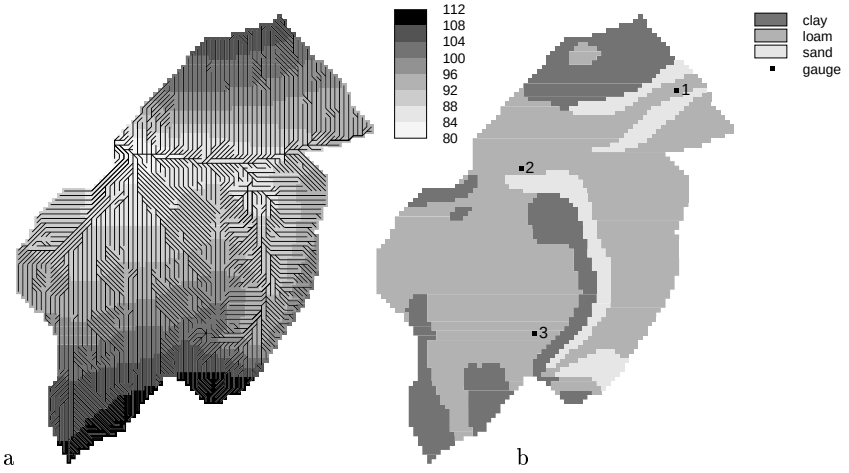


Fig. 2. (a) map of altitude (grey shades; units m above sea level) and local drain directions (drawn lines); (b) map of soil texture and location of rain gauges; mapped area is 1 km × 0.8 km

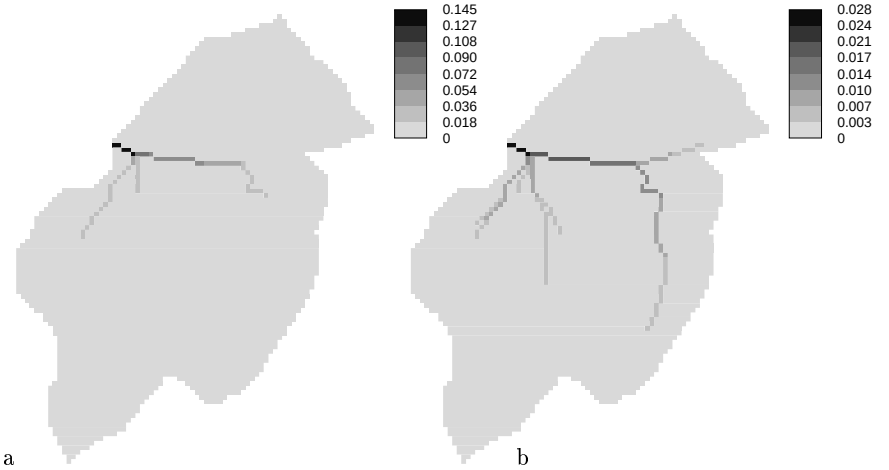


Fig. 3. (a) Median runoff at time step 20; (b) width of the empirical 95% confidence interval at time step 20

measurements should be used to assess whether the errors in infiltration as they are modelled here can account for the model prediction errors. In addition, infiltration measurements should have been used for a more realistic assessment of spatial errors in infiltration.

5 Discussion

A stochastic dynamic GIS can be obtained when one dimension is added to a GIS that provides a temporal dimension. We do not yet provide such an

environment in the PCRaster project, but all components required (a map algebra package; stochastic spatio-temporal simulation of discrete or continuous random fields, and stochastic output analysis functions) are available. The bottleneck is a convenient data structure that allows higher dimensional mapping.

When a stochastic GIS environment is provided, the question arises how a suitable MC sample size can be recommended to non-expert users. This depends trivially on which results (e.g., means or tail percentiles) are needed, but general guidelines such as they exist for bootstrapping methods still have to be developed. In addition, simple methods to experimentally determine the accuracy (sampling error) of estimated MC sample statistics, such as repeated Latin hypercube sampling (Pebesma and Heuvelink, 1999) should be provided.

Bigger challenges than handling random fields in a GIS are the comparison and critical evaluation of alternative GIS *model structures*, by analysing model residuals (observed minus predicted values). One of the questions a stochastic dynamic GIS can help to answer is whether the uncertainty with respect to GIS model input can completely account for the variation in model residuals.

Acknowledgments
This paper was written while the first author was a visiting scholar at the Department of Geological and Environmental Sciences at Stanford University. The Netherlands Organisation for Scientific Research (NWO) supported this visit with a travel stipend.

References

Burrough, P.A., and McDonnell, R.A. (1998). *Principles of Geographical Information Systems*. Oxford: Oxford University Press.

Deutsch, C.V. and Journel, A.G. (1998) *GSLIB Geostatistical Software Library and User's guide, second edition*. 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 E.J. Pebesma (1999) *Spatial aggregation and soil process modelling*. Geoderma 89, 47-65.

Kros, J., Pebesma, E.J., Reinds, G.J., Finke, P.A. (1999) *Uncertainty assessment in modelling soil acidification at the European scale: a case study*. Journal of Environmental Quality 28 (2), 366-377.

Loague, K., and Kyriakidis, P.C. (1997) *Spatial and temporal variability in the R-5 infiltration data set: Déjà vu and rainfall-runoff simulations*. Water Resources Research 33 (12), 2883-2895.

Pebesma, E.J., and Wesseling, C.G. (1998) *Gstat, a program for geostatistical modelling, prediction and simulation*. Computers & Geosciences, 24(1), 17-31. Software at `http://www.geog.uu.nl/gstat/`

Pebesma, E.J., Heuvelink, G.B.M. (1999) *Latin hypercube sampling of Gaussian random fields*, Technometrics 41(4), 303-312.

Tomlin, C.D., and Berry, J. (1979) *A Mathematical Structure for Cartographic Modeling and Environmental Analysis*. In: Proceedings of the ACSM, 269-283.

Wesseling, C.G., Karssenberg, D., Van Deursen, W.P.A. and Burrough, P.A. (1996) *Integrating dynamic environmental models in GIS: the development of a Dynamic Modelling language*. Transactions in GIS 1, pp 40-48. Software at `http://www.geog.uu.nl/pcraster/`