Caveat Emptor: Random Number Generators In Geospatial Analysis

Kimberly P. Van Niel and Shawn W. Laffan
School of Resources, Environment and Society, The Australian National University
Kimberly.VanNiel@anu.edu.au

Abstract. Many analyses within the field of GIS apply stochastic methods and systems, such as Monte Carlo techniques, dynamic modelling, stochastic simulation, artificial life and simulated data development. A pseudo-random number generator (PRNG) is employed within all these analyses, which can affect the validity of any results, yet GIS articles rarely report on the PRNG being used or on its settings. Not only does this make the research irreproducible, it also indicates that GIS researchers rarely, if ever, check the PRNG being employed for suitability for their analysis or simulation. Exacerbating the problem is that many geospatial and Monte Carlo software are not explicit about the PRNG used. Critical aspects of PRNGs from a geospatial standpoint need to be explored, especially before they are routinely used in the wider spatial analysis community.

1. INTRODUCTION

Analyses within the field of GIS are increasingly using methods and systems that apply pseudo-random number generators (PRNGs). However, PRNGs create deterministic approximations of random number sequences, and these can contain biases and correlation structures. It is therefore important, not only that GIS researchers are familiar with the concepts of PRNGs and communicate that knowledge to the wider spatial analysis community, but also that problems and issues specific to the application of PRNGs in GIS research are investigated and addressed.

2. THE USE OF PRNGS IN GIS

Techniques based on PRNGs are being used for a number of applications in GIS. This includes assessing the impact of error (Holmes et al. 2000, Næsset 1999, Huevelink and Burrough 1993, Openshaw 1989), aggregation within data sets (Bian and Butler 1999) and developing simulated data (Deutsh and Journel 1992). In addition, dynamic modelling (Burrough et al. 2000), stochastic simulation (Goovaerts 1999, 2000), artificial life (cellular automata), and flow algorithm perturbation (Fairfield and Leymarie 1991). However, it is likely that few GIS researchers assess the applicability of a PRNG before commencing with a research application. This is not unique to GIS research (see Barry, 1996). Even if GIS researchers are aware of the history, literature, testing, and nature of PRNGs, it is important that any PRNGs employed are tested for their suitability and are reported in the literature, so that less knowledgeable users do not unwittingly commit serious mistakes.

Although there are a number of different classes of PRNGs and many relatively "good" and "bad" PRNGs, no PRNG can perform well under all circumstances (L'Ecuyer and Simard 2001). Given that PRNGs are deterministic, all will have non-random characteristics and will fail under certain conditions (Hellakalek 1998). For example, good PRNGs can exhibit poor local randomness and still have good global randomness scores (Barry 1996). Thus, GIS researchers need to ensure not only that they are employing a relatively "good" PRNG, but also that they have investigated and tested its appropriateness in their specific research scenario. Unfortunately, very little is known about how certain characteristics of PRNGs will affect spatial analyses. A rich literature exists on PRNGs, consisting of theoretical discussions and testing methods in the mathematical literature (for example, Hellekalek 1998, Press et al. 1992), and programming and theoretical advice in the computer science literature (for example, Knuth 1998). Most information about using PRNGs in applications exists in the physics literature, and therefore many of the concerns about PRNGs and tests have been designed for problems in physics such as atomic particle movement. However, the user of PRNGs in spatial analyses may have very different concerns and requirements.

3. ISSUES IN GIS APPLICATIONS

There are two areas to consider in the application of PRNGs in GIS. The first is the nature of spatial data, such as its size and structure. The second is the nature and goals of the types of GIS analyses.

3.1 The Nature of Spatial Data and PRNGs

3.1.1 Dataset Size

Geographic datasets are typically large enough to reach or exceed maximum sample size restictions for many commonly used PRNGs, and certainly in the case of multiple runs without reseeding. All PRNGs have a period length, which is the number of values generated in a sequence before the sequence begins to repeat itself. Ideally, the entire period length could be used, but this is not normally the case due to increasing global correlation as the length of the sequence increases.

For one common class of PRNGs, linear congruential generators (LCGs), Hellekalek (1998) suggests the square root of the period length is the maximum sample size that should be used, while L'Ecuyer and Sinnard

(2001) suggest the cubed root. This can have serious implications for GIS analysis.

Consider an analysis of a GIS raster or image simulated using the RAND algorithm, as is implemented for the GRID.MakeRandom command in ArcView on a UNIX platform. The period length for RAND is 2^{31}, so the maximum raster size that could be used before the cycle repeats is 46341 by 46341 cells. This is quite large. However, given that correlations begin to occur at distances as short as the square root (Hellakalek, 1998) or even the cubed root (L'Ecuyer and Simard, 2001) of the period length, the maximum usable sequence may actually be approximately 46341 values (or 215 by 215 pixels) for the square root, and approximately 1290 values (36 by 36 pixels) for the cubed root.

3.1.2 Grid Analyses

In addition to often being large in size and therefore exceeding maximum sample sizes for PRNGs, raster grids are particularly suceptible to problems with global correlation and correlation structures within random sequences. Hidden correlations generally exist between numbers in a pseudo-random sequence due to the regularities generated by using a deterministic method. One of the implications of correlation in the sequence is the order in which pseudo-random numbers are assigned to dataset elements. The potential impact of correlation structures is different if each dataset element (cell) is assessed individually for the number of iterations, or if entire datasets are filled (for the creation of simulated data) or perturbed (for error analyses) before reiteration.

The manner in which a grid is filled or perturbed by a random sequence, combined with the size of the dataset and the nature of the correlation will have different effects. When the column size of the grid is a multiple of the correlation sequence, vertical structures develop within the grid. Their effect on a Moran's I using a rook's case sample will be significant and related to the correlation scale. Vertical structures result in indications of similarity at some sample sizes while diagonal structures in the other grids can create an impression of contrast.

The correlation structures can also have differing effects on results depending on the shape of the spatial analysis window. Some windows will be more susceptible to correlation structures than others, especially where they correspond to the sequence in which the random numbers are assigned to dataset elements. This has potentially large implications for spatial analyses using simulated datasets (eg. Bian and Butler 1999) and when testing the significance of local spatial analyses (Fotheringham et al. 2000).

3.1.3 Node and Line Analyses

Node and line analyses, on the other hand, may be strongly affected by low order serial correlation, especially if they are perturbed linearly. Minimum sample sizes should also be a concern for small-scale operations like the perturbation of arcs and nodes (Næsset 1999) when using a non-uniform distribution. To generate sequences with non-uniform distributions, a PRNG is used to create a uniform distribution, which is then transformed to the target distribution, for example Gaussian. The validity of the transformation depends on the assumption that the sequence is random (L'Ecuyer 1998). Also, the sequence should be checked for stability of the target distribution for a given sample size. LCGs, for example, are known to produce sequences which fall into "ruts", generating substrings well below or above the means (Barry 1996). Stability of the distribution generally increases with sample size (for example, Bang et al. 1998). In a sequence created using the ARC/INFO Grid function NORMAL on a Unix Solaris System with the system seed, the mean and standard deviation required up to 20,000 values before stabilising.

Most GIS error propagation studies assume a normal distribution for the error model (Openshaw 1989, Huevelink and Burrough 1993, Næsset 1999). Stability of the mean and standard deviation of the random sequence is particularly an issue where the primary interest of the application lies in the tails of the distribution, for example in significance testing (Barry 1996).

3.2 Types of GIS Analyses and PRNGs

A number of different types of GIS analyses have been listed in the introduction. The implications of the creation of simulated grid data sets and the perturbation of grid data sets and lines and nodes have also been discussed. One other method, Monte Carlo, which employs PRNGs, is becoming commonly used in GIS. Monte Carlo simulations have been relatively well studied (as a sample of the literature see Ferrenberg et al. 1992, Niederreiter 1992, and Barry 1996), although not specifically in a GIS context, whereas other uses of PRNGs in GIS have not been studied at all. Nonetheless, users of Monte Carlos should still be knowledgeable in the software and thus the PRNG employed and its potential problems, especially in regard to spatial analyses. Such analyses are highly reliant on the PRNG employed and its settings (Niederreiter 1992). Poor PRNGs can therefore lead to systematic errors in the analyses (Ferrenberg et al 1992). To test the stability of results, several non-overlapping sequences should be generated (Barry 1996). The result of the tests should determine whether the generated sequence can give unbiased or reliable answers to the problem at hand (Hammersley and Handscombe 1965). One approach is to solve a test problem similar to the application, but with a known answer, or that can be solved by another method (Deák 1990).

4. CONCLUSIONS

PRNGs are a valuable tool in GIS analysis, but they need to be used properly. PRNGs employed within a GIS context need to be analysed and tested given the

problems, assumptions, and specific requirements of each application.

The authors recommend that GIS researchers consider the following when employing a PRNG: know which PRNG is being used in the analysis, test it for suitability to the application, use more than one PRNG to test the stability of the results, and report both the PRNG and the starting seed in any publications. GIS and related software vendors, including risk assessment packages providing Monte Carlo analyses (Barry 1996), should be expected to disclose in the help files the PRNG used for specific commands, software, and platforms.

Each application of a PRNG should be tested against the needs, structure and constraints of each application (Shchur et al 1997). We also recommend the application of at least two different PRNGs for each analysis (for example see Press et al. 1992). For example, one could apply an LCG and then an inverse generator, which have very different structure and correlation properties (Hellekalek 1998). If it appears that the selection of the PRNG has a critical effect on the result, then further investigation would be necessary.

Finally, it is critical that GIS researchers report both the PRNG used and the starting seed so that other researchers can assess the reliability of the analysis and repeat it if necessary. In addition, this increases awareness within the GIS community that PRNGs underlie many GIS functions, PRNG testing might be a requisite part of an analysis, and that potential problems in the application of PRNGs may exist. It also provides a measure of accountability for the application of PRNGs.

This paper has presented some broad considerations for the use of PRNGs in spatial analyses, but more work is needed, especially considering the effects of the development and perturbation of raster data in dynamic modeling and for cellular automata. In the case of grid interactions with PRNGs, research from lattice structures in physics may be useful (for example Ferrenberg et al. 1992). It is most important that GIS researchers understand the nature and effects of the PRNGs employed, so that research and results are not based upon a problematic and poorly understood foundation.

REFERENCES

Bang, J., Schumacker, R.E., and Schlieve, P.L. (1998) "Random number generator validity in simulation studies: An investigation of normality," *Educational and Psychological Measurement*, vol.58 no.3, pp.430-450.

Barry, T.M. (1996) "Recommendations on the testing and use of pseudo-random number generators used in Monte Carlo analyses for risk assessment," *Risk Analysis*, vol.16 no.1, pp.93-105.

Bian, L. and Butler, R. (1999) "Comparing effects of aggregation methods on statistical and spatial properties of simulated spatial data," *Photogrammetric Engineering and Remote Sensing*, vol.65 no.1, pp.73-84.

Burrough, P.A., Van Gaans, P.F.M., and Macmillan, R.A. (2000) "High resolution landform classification using fuzzy k-means," *Fuzzy Sets and Systems*, vol.113 no.1, pp. 37-52.

Deak, I. (1990) *Random Number Generators and Simulation*, Akadémiai Kiadó: Budapest, Hungary.

Deutsch, C., and Journel, A.G. (1992) *GSLIB Geostatistical Handbook*. Oxford University Press: New York.

Emmi, P.C. and Horton, C.A. (1995) "A Monte Carlo simulation of error propagation in GIS-based assessment of seismic risk," *International Journal of Geographical Information Science*, vol.9 no.4, pp. 447-461.

Fairfield, J. and Leymarie, P. (1991) "Drainage networks from grid digital elevation models," *Water Resources Research*, vol.27, pp. 709-717.

Ferrenberg, A.M., Landau, D.P., and Wong, Y.J. (1992) "Monte Carlo simulations: Hidden errors in "good" random number generators," *Physical Review Letters*, vol. 69 no. 23, pp.3382-3384.

Fotheringham, A.S., Brunsdon, C. and Charlton, M. (2000) *Quantitative Geography, Perspectives on Spatial Data Analysis*, Sage.

Goovaerts, P. (1999) "Geostatistics in soil science: state of the art and perspectives," *Geoderma*, vol. 89, pp. 1-45.

Goovaerts, P. (2000) "Estimation or simulation of soil properties? An optimization problem with conflicting criteria," *Geoderma*, vol. 97, pp. 165-186.

Hammersley, J.M., and Handscomb, D.C. (1965) *Monte Carlo Methods*, Methuen & Co.

Hellekalek, P. (1998) "Good random number generators are (not so) easy to find," *Mathematics and Computers in Simulation*, vol. 46 no. 5-6, pp. 485-505.

Heuvelink, G.B., and Burrough, P.A. (1993) "Error propagation in cartographic modelling using Boolean logic and continuous classification," *International Journal of Geographical Information Science*, vol. 7 no. 3, pp. 231-246.

Holmes, K.W., Chadwick, O.A., and Kyriakidis, P.C. (2000) "Error in a USGS 30-meter digital elevation model and its impact on terrain modeling," *Journal of Hydrology*, vol. 233, pp. 154-173.

Knuth, D.E. (1998) *The Art of Computer Programming, Third Edition, Volume 2: Seminumerical Algorithms*, Addison-Wesley.

L'Ecuyer, P., and Simard, R. (2001) "On the performance of birthday spacing tests with certain families of random number generators," *Mathematics and Computers in Simulation*, vol. 55, pp. 131-137.

Næsset, E. (1999) "Effects of delineation errros in forest stand boundaries on estimated area and timber volumes," *Scandinavian Journal of Forest Resources*, vol. 14, pp. 558-566.

Niederreiter, H. (1992) *Random Number Generation and Quasi-Monte Carlo Methods*, Society for Industrial and Applied Mathematics.

Openshaw, S. (1989) "Learning to live with errors in spatial databases," In Goodchild, M. F., and Gopal, S., (eds.), *The Accuracy of Spatial Databases*, Taylor and Francis: London, pp. 263-276.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P. (1992) *Numerical Recipes in C: The Art of Scientific Computing*, Second Edition, Cambridge University Press.

Shchur, L.N., Heringa, J.R., and Blöte, H.W.J. (1997) "Simulation of a directed random-walk model: The effect of pseudo-random-number correlations," *Physica A*, vol. 241, pp. 579-592.