‘spup’ – an R package for uncertainty propagation in spatial environmental modelling

K. Sawicka*, G.B.M. Heuvelink

Soil Geography and Landscape Group, Wageningen University, PO Box 47, 6700 AA, Wageningen, The Netherlands

*Corresponding author: kasia.sawicka@wur.nl

Abstract

Computer models are crucial tools in engineering and environmental sciences for simulating the behaviour of complex systems. While many models are deterministic, the uncertainty in their predictions needs to be estimated before they are used for decision support. Advances in uncertainty analysis have been paralleled by a growing number of software tools, but none has gained recognition for universal applicability, including case studies with spatial models and spatial model inputs. We develop an R package that facilitates uncertainty propagation analysis in spatial environmental modelling. The ‘spup’ package includes functions for uncertainty model specification, propagation of uncertainty using Monte Carlo (MC) techniques, and uncertainty visualization functions. Uncertain variables are represented as objects which uncertainty is described by probability distributions. Spatial auto-correlation within a variable and cross-correlation between variables is also accommodated for. The package has implemented the MC approach with efficient sampling algorithms, i.e. stratified random sampling and Latin hypercube sampling. The MC realizations may be used as an input to the environmental models called from R, or externally. Selected static and interactive visualization methods that are understandable by non-statisticians can be used to visualize uncertainty about the measured input, model parameters and output of the uncertainty propagation.

Key words

R language, uncertainty analysis, uncertainty propagation, spatial models, Monte Carlo

I INTRODUCTION

Computer models have become a crucial tool in engineering and environmental sciences for simulating the behaviour of complex static and dynamic systems. However, while many models are deterministic, the uncertainty in their predictions needs to be estimated before they are used for decision support. Currently, advances in uncertainty propagation and assessment have been paralleled by a growing number of software tools for uncertainty analysis, but none has gained recognition for a universal applicability, including case studies with spatial models and spatial model inputs. Due to the growing popularity and applicability of the open source R programming language we undertook a project to develop an R package that facilitates uncertainty propagation analysis in spatial environmental modelling. The tool is intended for researchers and practitioners who understand the problems of uncertainty in data and models, and are looking for a simple, accessible implementation of the universal

methodology for uncertainty assessment. At the same time, it is designed to enable more experienced users to easily understand, customise, and possibly further develop the code.

A number of computational tools are readily available to tackle the uncertainty quantification problem to different degrees. These include both free software, like OpenTURNS (Andrianov et al., 2007), DACOTA (Adams et al., 2009) and DUE (Brown and Heuvelink, 2007), commercial, like COSSAN (Schuëller and Pradlwarter, 2006), or free, but written for a licenced software, e.g. SAFE (Pianosi et al., 2015) or UQLab (Marelli and Sudret, 2014) toolboxes for MATLAB. A broad review of existing software packages is available in Bastin et al. (2013). To the best of our knowledge, however, none of the existent software is specifically designed to be extended by the environmental science community. The use of powerful but complex languages like C++ (e.g. Dakota), Python (e.g. OpenTURNS) or Java (e.g. DUE) often discourages relevant portions of the non-highly-IT trained scientific community from the adoption of otherwise powerful tools.

The R programming language is an important tool for development in numerical and statistical analysis. R has advantages through its advanced statistical capabilities and high-quality graphical output (Ripley, 2001), and is gaining widespread use in science and education. Furthermore, through the use of R packages, the software can be used for a variety of geoscience analyses and visualisations. It has grown tremendously over the last 20 years, with over 8000 packages at the time of preparation of this paper. There is a number of R packages invoking uncertainty analysis through sensitivity analysis or use of a Bayesian framework for model calibration. We have found only one package named 'propagate' that deals with uncertainty propagation explicitly, using similar approaches as described in this paper. The package 'propagate', however, does not provide functionality for spatial models and variables.

II EMPLOYED (SPATIAL) UNCERTAINTY PROPAGATION ANALYSIS APPROACH

Uncertainty propagation aims to analyse how uncertainties in data (e.g. from measurement error, sampling, interpolation), combined with model uncertainties (e.g. in the model parameters and structure) propagate through the model (Heuvelink et al., 2007). Many environmental phenomena of interest are spatial, temporal or spatio-temporal in nature and often have strong correlations imposed by the physics and dynamics of the natural systems, making uncertainty evaluation difficult. The most frequently used approach represents uncertainty with probability distribution functions (pdfs). The pdf describes the relative likelihood for the random variable to take on a given value and typically it is viewed as a shape of the distribution, for example normal, uniform, lognormal or exponential. It is common for the pdf to be parametrized, i.e. to be characterized by distribution parameters. For example, the normal distribution is parametrized in terms of the mean and the variance, or uniform distribution is parametrized by minimum and maximum values. For situations in which pdfs can be estimated reliably, they have a number of advantages over non-probabilistic techniques. They include methods for describing cross- and auto- correlation between uncertainties, methods for propagating uncertainties through simple algebras or more complex environmental

models, and methods for tracing the sources of uncertainty in environmental data and models (Heuvelink 1998).

A frequently used method for the analysis of uncertainty propagation is the Monte Carlo (MC) method (Hammersley and Handscomb, 1979, Lewis and Orav, 1989). It is very flexible and can reach an arbitrary level of accuracy, and therefore it is generally preferred over analytical methods such as the Taylor series method (Heuvelink, 1998). The idea of the MC method is to compute the output of the model repeatedly, with input values that are randomly sampled from their marginal or joint pdf. The set of model outputs forms a random sample from the output pdf, so that the parameters of the distribution, such as the mean, variance and quantiles, can be estimated from the sample. The method thus consists of the following steps:

1. Characterise uncertain model inputs with pdfs.
2. Repeatedly sample from (spatial) pdfs of uncertain inputs.
3. Run model with sampled inputs and store model outputs.
4. Compute summary statistics of model outputs.

Note that the above ignores uncertainty in model parameters and model structure, but these can easily be included if available as pdfs. A random sample from the model inputs can be obtained using an appropriate pseudo-random number generator (Lewis and Orav, 1989). Note that a conditioning step will have to be included when the model inputs are correlated. Application of the MC method to uncertainty propagation with operations that involve spatial interactions requires the simultaneous generation of realisations from the spatially distributed inputs implying that spatial correlation will have to be accounted for (Heuvelink et al., 1989). For uncertain spatially distributed continuous variables, such as elevation, rainfall and soil organic carbon content, we assume the following geostatistical model:

$$Z(x) = \mu(x) + \sigma(x) \cdot \varepsilon(x) \quad (1)$$

where μ is the (deterministic) mean of Z, σ is a spatially variable standard deviation of the prediction of μ (spatial variability of σ reflects that in some parts of study area the uncertainty is greater than in other parts), and ε is a standardized, zero-mean, spatially auto-correlated residual modelled with a semivariogram or a correlogram (Diggle and Ribeiro, 2007, Webster and Oliver, 2007, Plant, 2012). The random sample is drawn from the pdf of ε to further calculate the realizations of Z.

The drawback of the MC method is that the accuracy of the uncertainty assessment is inversely related to the square root of the number of runs N. This means that to double the accuracy, four times as many runs are needed. In complex, multi-variable systems high accuracies are obtained only when the number of runs is very large (i.e. $N \geq 1{,}000$), which may cause the method to become extremely time consuming. The improvement on MC efficiency can be made by employing efficient sampling techniques (e.g. Latin hypercube sampling) and parallel computing.

III 'spup' (SPATIAL UNCERTAINTY PROPAGATION) PACKAGE DESIGN

The adopted approach for uncertainty propagation analysis dictates the general package design. The 'spup' package provides functions for examining the uncertainty propagation starting from input data and model parameters, via the environmental model onto model outputs (Fig. 1). The functions include uncertainty model specification, stochastic simulation and propagation of uncertainty using MC techniques, as well as several uncertainty visualization functions.

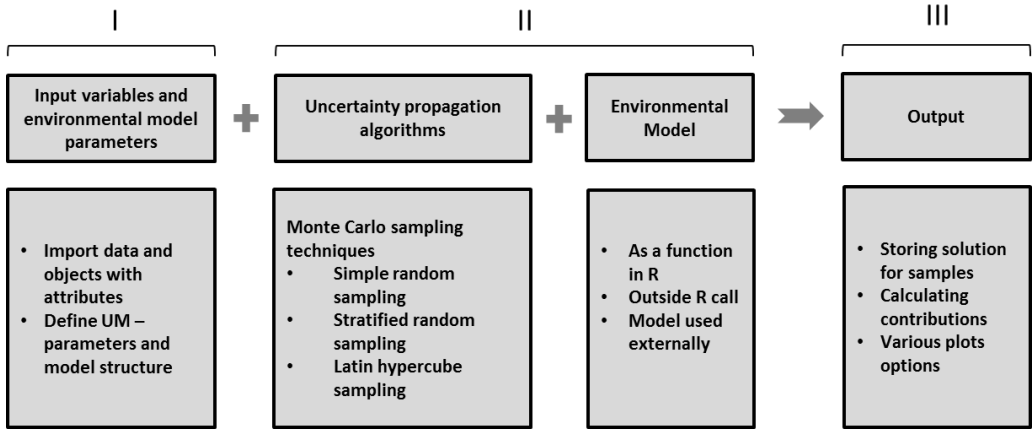

Figure 1 . The 'spup' package design. 'spup' comprises of functions for defining uncertainty model (I), quantifying uncertainty propagation (II) and storing output in a format of data or images.

Uncertain environmental variables are represented in the package as objects whose attribute values may be uncertain and described by probability distributions. Uncertainty assumption may also be ignored, in which case, during the model run the user works with μ (Eq. 1) as the model input that best represent the reality. Both numerical (e.g. air humidity) and categorical data (e.g. land cover) types are handled. Spatial auto-correlation within an attribute and cross-correlation between attributes is also accommodated for. The attributes may be independent in space, for which a marginal probability density function (mpdf) is defined at each point in space, or may co-vary in space, for which a joint probability density function (jpdf) is defined. Different shapes of marginal pdfs are supported, whereas joint pdfs may be defined for groups of attributes characterized with the normal distribution only. The specification of correlations between errors in space and cross-correlations between objects or attributes is made under the assumption that the correlations depend only on the distance between locations.

For spatially correlated variables the package relies on the unconditional Gaussian simulation implemented in the 'gstat' package (Pebesma, 2004). For drawing realizations of uncertain variables without assumed correlations the package has implemented the MC approach with efficient sampling algorithms, i.e. stratified random sampling and Latin hypercube sampling. The design includes facilitation of parallel computing to speed up MC computation. The MC realizations for uncertainty propagation quantification may be used as an input to the environmental models called from R, or externally.

Selected static (adjacent maps and glyphs) and interactive visualization methods that are understandable by non-experts with limited background in statistics can be used to summarize and visualize uncertainty about the measured input, model parameters and output of the uncertainty propagation.

IV APPLICATION EXAMPLE – MAPPING SOIL MOISTURE CONTENT FOR THE ALLIER CATCHMENT

As part of a research study in quantitative land evaluation, the World Food Studies (WOFOST) crop simulation model (van Diepen et al., 1989) was used to calculate potential crop yields for floodplain soils of the Allier river in the Limagne rift valley, central France. The moisture content at wilting point (Θ_{wp}) is an important input attribute for the WOFOST model. Because Θ_{wp} varies considerably over the area in a way that is not linked directly with soil type, it was necessary to map its variation separately to see how moisture limitations affect the calculated crop yield.

Unfortunately, because Θ_{wp} must be measured on samples in the laboratory, it is expensive and time-consuming to determine it for a sufficiently large number of data points for creating the prediction map by kriging. An alternative and cheaper way is to calculate Θ_{wp} from other indicators which are cheaper to measure. Because the moisture content at wilting point is often strongly correlated with the moisture content at field capacity (Θ_{fc}) and the soil porosity (Φ), both of which can be measured more easily, it was decided to investigate how errors in measuring and mapping these would work through to a map of calculated Θ_{wp}. Calculation of Θ_{wp} can be done using a pedo-transfer function, which in this case takes the form of multiple linear regression:

$$\Theta'_{wp} = \beta_0 + \beta_1 \cdot \Theta_{fc} + \beta_2 \cdot \Phi + \delta \qquad (2)$$

where Θ'_{wp} denotes measured moisture content at wilting point, β_0 , β_1 , and β_2 are the regression coefficients and δ denote residuals attributed to lack of model fit and measurement error. The regression coefficients were estimated using standard ordinary least squares regression, ignoring spatial correlation between the observations at the locations. The maps of Θ_{fc} and Φ were derived using co-kriging and accounted for spatial cross-correlation between Θ_{fc} and Φ. Each component on the right hand side of Eq. 2 is subject to uncertainty, which will propagate to uncertainty about Θ_{wp}. Following the adopted MC approach, for each variable and parameter the uncertainty model is defined and 1000 MC samples are drawn. For the spatial variables a linear model of co-regionalization (Wackernagel, 2003) is fitted with use of the 'gstat' package and possible realities are simulated. The joint pdf of the model parameters and structural error δ was estimated using Bayesian calibration (Van Oijen et al., 2005) (note, this is not included in the 'spup' package) and a random sample was drawn from their joint posterior distribution. 1000 realizations of Θ_{wp} was then calculated using Eq. 2 and summary statistics such as mean of prediction and standard deviation were derived.

If an uncertainty analysis with WOFOST would show that the errors in Θ_{wp} cause errors in the output of WOFOST that are unacceptably large, then the accuracy of the map of Θ_{wp}

would have to be improved. In order to decide how to proceed in such a situation, the contribution of each individual error source to the overall uncertainty in Θ_{wp} was determined as well. Figure 2 presents the results and these show that both Θ_{fc} and Φ, rather than model parameters and model structural error, form the main source of uncertainty. Thus, the main source of error in Θ_{wp} is the one associated with the kriging errors of Θ_{fc} and Φ.

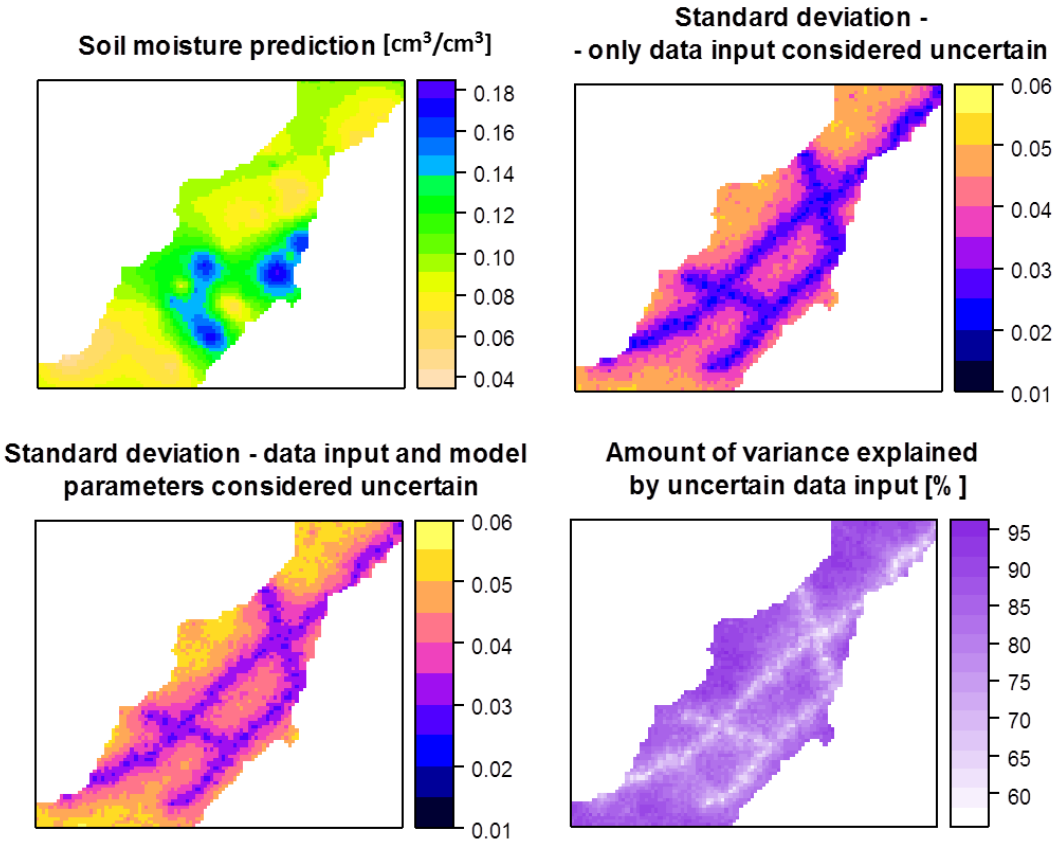

Figure 2 Results of uncertainty propagation for soil moisture prediction in the Allier catchment.

V CONCLUSIONS AND FURTHER WORK

We present a tool for uncertainty propagation assessment based on the uncertainty quantification framework described in e.g. Heuvelink et al. (2007). As the theoretical framework and implementation of the package progress, its application to real cases will be necessary, both to test the algorithms and usability of the tool, and to demonstrate the importance of assessing uncertainty in environmental data. The 'spup' package is being developed and used within the project "Quantifying Uncertainty in Integrated Catchment Studies (QUICS)". QUICS aims to carry out research in order to take the implementation of the Water Framework Directive (WFD) to the next level and improve water quality management by assessing the uncertainty of integrated catchment model water quality predictions. Currently, the potential case studies for the 'spup' application include uncertainty propagation analysis

with the LandscapeDNDC model (Haas et al., 2012) and German Schwingbach catchment data, and Metaldehyde Prediction model developed currently for the Severn Tent Water, water provider in the Midlands, UK. Finally, 'spup' will be introduced to the wider scientific community through CRAN (The Comprehensive R Archive Network), where many more challenges will be faced, including the time and resources required to implement an uncertainty assessment and the need to make uncertainty analyses understandable to non-statisticians.

Acknowledgements

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 607000. We would like to thank Sytze de Bruin, Damiano Luzzi, Stefan van Dam and Dennis Walvoort for valuable contributions to the development of *spup.*

References

ADAMS, B. M., BAUMAN, L. E., BOHNHOFF, W. J., DALBEY, K. R., EBEIDA, M. S., EDDY, J. P., ELDRED, M. S., HOUGH, P. D., HU, K. T., JAKEMAN, J. D., SWILER, L. P. & VIGIL, D. M. 2009. DAKOTA, A Multilevel Parallel Object-Oriented Framework for Design Optimization, Parameter Estimation, Uncertainty Quantification, and Sensitivity Analysis: Version 5.4 User's Manual.

ANDRIANOV, G., BURRIEL, S., CAMBIER, S., DUTFOY, A., DUTKA-MALEN, I., DE ROCQUIGNY, E., SUDRET, B., BENJAMIN, P., LEBRUN, R., MANGEANT, F. & PENDOLA, M. OpenTURNS, an open source initiative to Treat Uncertainties, Risks'N Statistics in 520 a structured industrial approach. ESREL'2007 Safety and Reliability Conference, 2007 Stavenger, Norway.

BASTIN, L., CORNFORD, D., JONES, R., HEUVELINK, G. B. M., PEBESMA, E., STASCH, C., NATIVI, S., MAZZETTI, P. & WILLIAMS, M. 2013. Managing uncertainty in integrated environmental modelling: The UncertWeb framework. *Environmental Modelling & Software,* 39**,** 116-134.

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.

DIGGLE, P. & RIBEIRO, P. J. 2007. *Model-based geostatistics*, Springer.

HAAS, E., KLATT, S., FRÖHLICH, A., KRAFT, P., WERNER, C., KIESE, R., GROTE, R., BREUER, L. & BUTTERBACH-BAHL, K. 2012. LandscapeDNDC: a process model for simulation of biosphere–atmosphere–hydrosphere exchange processes at site and regional scale. *Landscape Ecology,* 28**,** 615-636.

HAMMERSLEY, J. M. & HANDSCOMB, D. C. 1979. *Monte Carlo methods,* London, Chapman and Hall.

HEUVELINK, G. B. 1998. *Error Propagation in Environmental Modelling with GIS,* London, Taylor and 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.

HEUVELINK, G. B. M., BURROUGH, P. A. & STEIN, A. 1989. Propagation of errors in spatial modelling with GIS. *International Journal of Geographical Information Systems,* 3**,** 303-322.

LEWIS, P. A. W. & ORAV, E. J. 1989. *Simulation methodology for statisticians, operations analysts, and engineers* Pacific Grove, Wadsworth & Brooks/Cole.

MARELLI, S. & SUDRET, B. 2014. UQLab: A Framework for Uncertainty Quantification in Matlab. *Vulnerability, Uncertainty, and Risk.* American Society of Civil Engineers.

PEBESMA, E. J. 2004. Multivariable geostatistics in S: the gstat package. *Computers & Geosciences,* 30**,** 683-691.
PIANOSI, F., SARRAZIN, F. & WAGENER, T. 2015. A Matlab toolbox for Global Sensitivity Analysis. *Environmental Modelling & Software,* 70**,** 80-85.
PLANT, R. E. 2012. *Spatial data analysis in ecology and agriculture using R*, CRC Press.
RIPLEY, B. D. 2001. The R project in statistical computing. *MSOR Connections. The newsletter of the LTSN Maths, Stats \textbackslash{\&} OR Network***,** 23--25.
SCHUËLLER, G. I. & PRADLWARTER, H. J. 2006. Computational stochastic structural analysis (COSSAN) – a software tool. *Structural Safety,* 28**,** 68-82.
VAN DIEPEN, C. A., WOLF, J., VAN KEULEN, H. & RAPPOLDT, C. 1989. WOFOST: a simulation model of crop production. *Soil Use and Management,* 5**,** 16-24.
VAN OIJEN, M., ROUGIER, J. & SMITH, R. 2005. Bayesian calibration of process-based forest models: bridging the gap between models and data. *Tree Physiol,* 25**,** 915-27.
WACKERNAGEL, H. 2003. *Multivariate Geostatistics: An Introduction with Applications*, Springer.
WEBSTER, R. & OLIVER, M. A. 2007. *Geostatistics for environmental scientists*, John Wiley & Sons.