Copula metadata est

Didier G. Leibovici and Mike J. Jackson
University of Nottingham
Nottingham Geospatial Institute
Nottingham, UK
didier.leibovici@nottingham.ac.uk
mike.jackson@nottingham.ac.uk

Abstract

Assessing the errors in model prediction from running a geospatial workflow (combining geospatial data and geo-processes from various sources) is becoming an essential component and requirement in geospatial modelling. Metadata such as the spatial data quality play a crucial role in bounding the decision-making made from the results of the geospatial workflow run, with uncertainty qualifying the outcome. The error propagation due to the succession of geocomputational operations is the major method to estimate this uncertainty. Recently the meta-propagation of uncertainty has been described as an alternative to the computationally intensive error propagation approach based on running many times the whole scientific workflow model. In short, meta-propagation uses spatial data quality information standardised by the ISO19157 and matching equivalent principles derived for the geo-processing tasks, that allows embedding the error propagation problem within an "uncertainty distributional space". Deriving these metadata for geo-processes becomes as much important as deriving the spatial data quality information to be able to evaluate the quality associated to an instantiated scientific workflow. As this gives rise to consider multivariate issues, the paper explores the role that copula distribution estimation can play in this matter. Seen as an intrinsic modelling of the dependence of a set of random variables, the copula information can be integrated in different ways as metadata related to a geospatial workflow. Is it a quality metadata linked the inputs, to the outputs, to the geo-processing tasks or all of them? For each case the paper explores its potential uses in meta-propagation.

Keywords: uncertainty, error propagation, metadata, copula distribution, scientific workflow.

1 Introduction

Spatial data quality, its derivation, its use, its impact and even its quality (meta-quality) have made important progress in research and methodologies allowing its implementation in spatial data infrastructure [1]. Maybe the biggest progress of all has been its outreach and consideration as a major issue in geospatial data management, usability and particularly in decision-making [2, 3, 4, 8]. This decision-making process can derive from simple use of a data layer or a data mash-up (seen here as just an overlay) maybe involving a decision rule or part of geocomputational workflow process: a geoworklfow.

The different descriptions of the data accuracy, recorded within the spatial data quality metadata, play a crucial role in bounding the decision-making with uncertainty qualifying the outcome. Error propagation, is a method to estimate this uncertainty propagated from the input data to the output datasets via the geocomputational operations of the workflow and eventually when applying the decision rules [1, 5]. Recently the meta-propagation of uncertainty [6, 7] has been described as an alternative to the computationally intensive error propagation running many times the whole scientific workflow (the computer model) or the less intensive method using a stochastic approximation emulating the workflow [8]. Meta-propagation uses, and combines within a meta-workflow computation, the spatial data quality information standardised by the ISO19157 and their equivalents using matching principles for the geo-processing tasks [7].

As part of an uncertainty analysis these metadata can be derived for geo-processes but, like any uncertainty analysis, simple or more complex approaches can be taken. For example in the case of quantitative variable (*e.g.,* position, numerical attribute) a relatively straightforward variance transfer function can be obtained within a classical univariate OAT (one-at-a-time) sampling within a Monte-Carlo simulation [9]. In order to get an uncertainty information propagated more comprehensive, either because of the form of the uncertainty measure, or its support (spatial information) or the simultaneous assessment for a combination of inputs, then more advanced meta-propagation (or error propagation in general) is needed. The concept of meta-propagation appeared to be easily extendable to these situations in principle but maybe at the price of complex and tedious uncertainty experiments [7].

A cornerstone facilitating further research for this topic could come from using a copula distribution seen as an intrinsic modelling of the dependence of a set of random variable [10]. Copulae distributions have been used already in the context of uncertainty and sensitivity analysis [11, 12] and the paper investigates the different ways of integrating this information in the metadata for quality information: metadata linked to sets of inputs or metadata to geo-processing tasks or both? This integration is dependent on its use in meta-propagation and for different purposes it may be used differently, but still as a metadata. A few numerical experiments, using existing libraries such as the copula R packages [10], are set to illustrate the potentials of this approach using a geoworkflow (Figure 2.) aiming at producing risk of land degradation maps within an infrastructure for desertification monitoring [14].

2 Meta-propagation of uncertainties

Meta-propagation is based on the principle that, if classical error propagation and possibly error propagation by emulation, can give much more accurate results, they are computationally expensive and particularly when it comes to have a more comprehensive approach as described in the introduction. The possibility of keeping a low computational cost even when the workflow becomes large (a workflow with a lot of tasks and lots of data inputs at different levels) has been driving the idea of combining existing information already recorded in the metadata [6]. Nonetheless, going beyond basic use of the meta-propagation for any quality element of the ISO19157 and any quality data type offers some challenges.

2.1 Basic meta-propagation

On Figure 1, an atomic workflow, consisting of a single task, supposedly indivisible at least in term of management, is represented with, for the sake of simplicity, 3 input map datasets of a single variable each and 2 output maps of a single variable each as well. The meta-propagated uncertainty as recorded in the quantitative attribute accuracy, for $o1$ and due to $i1$, here measured by the variance, is then defined by:

$$v_{o1} = Tf(v_{i1}) \qquad (1)$$

where the transfer function Tf is a metadata of the geoprocessing task here with the scope from $i1$ to $o1$. The Tf is obtained from an initial (say before publishing the atomic geoworkflow as an OGC Web Processing Service) uncertainty analysis experiment, *i.e.,* a classical error propagation, to fit the function for a specific range a variances, potentially with different datasets within a determined usability requirement setting for this geoprocessing task.

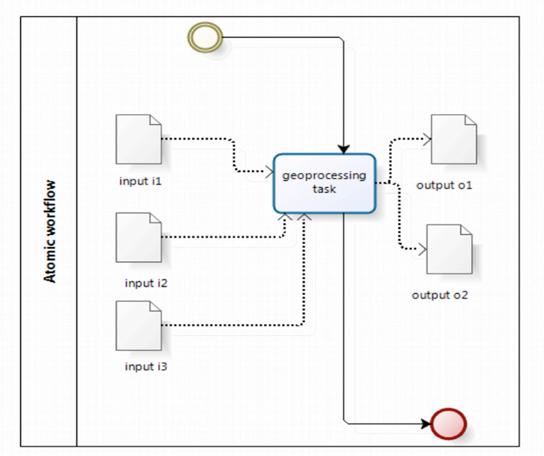
Following this principle other quality elements and different measures of the ISO1957 have been investigated in [7] using potential other transfer functions.

2.2 Advanced issues

As mentioned in the introduction it may be desirable or necessary to extend the error propagation to more comprehensive uncertainty information such as replacing the simple uncertainty measurements of the variance by the full probability density function (pdf_{i1}). Also, in equation (1) is this variance attached to the whole map? to each single pixel/point? and if a variance map is available, does the transfer function takes into account the autocorrelation of the uncertainties [15]?

This spatiality issue can be seen as a specific case of a more general non-separable issue, that is considering the input uncertainty as a whole, such as for the quantitative attribute accuracy a multivariate density (pdf_I) which can be propagated only to one output or to all. These were briefly discussed in [7], particularly the geostatisticaly flavoured issue of a variance map transfer function taking into account the autocorrelation, building on the potential use of a geostatistical sampling [15]. Nonetheless, acknowledging the need to recover a multivariate pdf in the case of non-separable (multivariate) uncertainty assessments with the potential use of a copula modelling was introduced in [7] as to be part of the transfer function.

Figure 1: Atomic geo-processing workflow.



I is the set of inputs $i1, i2, i3$ and O is the set of outputs $o1, o2$.

3 Copulae based approach

Let us remind the basics of copulae distributions. A copula of dimension p (or p-copula) is defined as a multivariate distribution on the unit hypercube in $\mathbb{R}^p$, $[0,1]^p$, with uniform univariate margins. Sklar's theorem stipulates that for any multivariate cumulative distribution function (*cdf*) F with margins $F_1, F_2, \dots, F_p$, there exists a copula C such that (2) holds. Then if the margins $F_1 , F_2, \dots, F_p$, are continuous (3) holds that is $C = C_F$ is unique, and if the distributions have absolute continuity properties (2) becomes (4) when considering the *pdf*s.

$$F(x_1, x_2, \dots, x_p) = C(F_1(x_1) , F_2(x_2), \dots, F_p(x_p)) \qquad (2)$$

$$C_F(p_1, p_2, \dots, p_p) = F(F_1^{-1}(p_1) , F_2^{-1}(p_2), \dots, F_p^{-1}(p_p)) \quad (3)$$

$$f(x_1, , \dots, x_p) = f_1(x_1) \dots f_p(x_p) c_f(F_1(x_1) , \dots, F_p(x_p)) \quad (4)$$

The simplicity of these equations has probably led to the success of copula modelling: any multivariate distribution can be fitted by separately fitting the margins and "its" dependence structure. Obviously great care must be taken, as this modelling can be critical due, for example, to a possible semantic confusion between a sample correlation between two variables say (x_1, x_2) and between the images of the samples on the hypercube $(F_1(x_1), F_2(x_2))$ [13]. Nonetheless as a copula can be estimated from the empirical distribution of the ranks (rescaled in [0,1]), this confusion is like comparing Pearson correlation and Spearman correlation coefficients.

Copulae have been applied in different context, lots of recent utilisations and references can be found in [10, 11, 12] which doesn't reflect the abundant literature even in geostatistics and uncertainty analysis.

3.1 Metadata of inputs

For a given atomic workflow as in Figure 1, if one is able to perform an uncertainty analysis that draw samples from the knowledge of a joint *cdf* of a set of inputs (or all the inputs), its copula can be estimated and kept as a metadata associated to this set. When reusing the (geo)processing task, it can be expected the set of inputs to behave in similar ways as during say the "calibration". Then, under the assumption that the dependence structure should be similar for any other semantically similar set of inputs but with different margins, one is able to recover an approximation of the joint *cdf* that could be used for uncertainty propagation under a multivariate approach:

$$\widehat{F_I}(i_1, i_2, \dots, i_l) = C_I(F_1(i_1), F_2(i_2), \dots, F_p(i_I)) \quad (5)$$

where the margins are the uncertainty metadata associated to each input variable of the newly selected set of inputs. Here, the copula C_I is seen as metadata of a set of inputs but also somehow attached to the geoprocessing task as resulting from the uncertainty analysis made under this dependence structure. For the new sets of inputs, C_I is more a borrowed metadata, which may also have been built as a "compromise" between different sets of inputs during the original uncertainty analysis.

Nonetheless, its use in meta-propagation still means that the producer of the geoprocessing task has been able to provide an appropriate metadata as a transfer function of a *cdf (or of a pdf*) [7]:

$$\widehat{F}_O = Tf(\widehat{F_I}) \quad (6)$$

The approximation of $\widehat{F_I}$ in equation (5) can of course also be used in a traditional error propagation experiment or using an emulator (also called a metamodel in the engineering sensitivity analysis community [9]).

3.2 Metadata of outputs

In the same way as for the inputs, it is possible to imagine that, during the original uncertainty experiment performed by the producer of the geoprocessing task with initial uncertainties within certain requirements, the joint *cdf* of the ouputs (or a subset of them) had been fitted using a copula distribution C_O. Again this metadata attached to the outputs is borrowed from the fitting during the geoprocessing task uncertainty analysis.

Nonetheless, if it is expected that under a new use of this geoprocessing task, the ouputs would expose similar dependence structure, and if we can recover each marginal *cdfs* of each outputs, say by other meta-propagations:

$$\widehat{F}_{oy} = Tf_{xy}(F_{ix}) \text{ or } \widehat{F}_{oy} = Tf_{Iy}(\widehat{F}_I) \quad (7)$$

then

$$\widehat{F}_O = C_O(\widehat{F}_{o1}, \widehat{F}_{o2}, \dots, \widehat{F}_{oO}) \quad (8)$$

reconstructs an estimation of the joint output *cdf.*

3.3 Metadata of geoprocesses

In the two previous sections the wording borrowed metadata, to qualify the copula fitted, has been used as if it was seen as dependent of the uncertainty experiment involving the geoprocessing task. It is evident for the copula C_O as a metadata of the outputs, but to a lesser extent it is also the case for the copula of the inputs C_I as it is linked to the transfer function fitted at the same time. Now this can be entirely the case when a copula is fitted for the joint input and output *cdf,* say for example a bivariate copula for one output and one input: C_{oi}. Then, keeping as metadata for the geoprocessing task, the copula but also the original output *cdf*, a new use of the geoprocessing task will lead to the uncertainty meta-propagation using:

$$\widehat{F}_{oi} = C_{oi}(\widehat{F}_o, F_i) \quad (9)$$

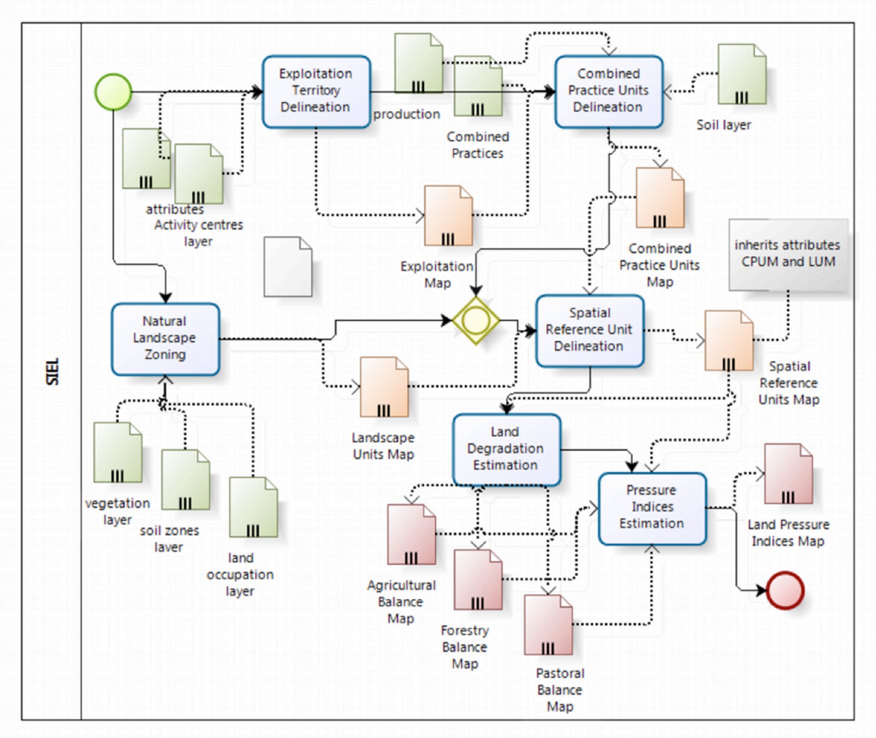
$$\widehat{F}_o = \lim_{i(.)\to\infty} \widehat{F}_{oi} \quad (10)$$

A meta-propagation algorithm, representing the transfer function Tf_{io}can then be implemented by alternating equation (9) and (10). This can be extended to the all set of inputs and outputs, $I \quad O,$ if feasible in practice. Notice here for the multivariate inputs or outputs the potentiality of using in combination sections 3.1 and 3.2 with section 3.3.

4 Experiment and spatial sampling

As an exploratory and feasibility experiment we used the geoprocessing workflow depicted on Figure 2. It has the

Figure 2: Geoworkflow of the SIEL model [14], BPMN representation.


SIEL
Exploitation Territory Delineation
production
Combined Practices
Combined Practice Units Delineation
Soil layer
attributes
Activity centres layer
Exploitation Map
Combined Practice Units Map
inherits attributes CPUM and LUM
Natural Landscape Zoning
Spatial Reference Unit Delineation
Spatial Reference Units Map
Landscape Units Map
Land Degradation Estimation
Pressure Indices Estimation
Land Pressure Indices Map
vegetation layer
soil zones layer
land occupation layer
Agricultural Balance Map
Forestry Balance Map
Pastoral Balance Map

advantage of having its atomic workflows to be of various categories when looking at the uncertainties involved. The “Natural Landscape Zoning” works mainly on geometries but also on classification attributes (*e.g.*, soil classes) whilst the “Combined Practices Units Delineation” uses both geometries and quantitative attributes, and the “Land Degradation Estimation” is a polygon-wise operation on quantitative attributes.

Here, we concentrate only on the “Land Degradation Estimation” dealing with quantitative attribute accuracy. Univariate uncertainties are obtained from expert elicitation, then estimates of needed copula are made using a spatial sampling under assumption of ergodicity (in the hypercube space), and using mainly the non-parametric estimation for the copula. The sampling can be fitted to depart from potential auto-correlation or in the contrary can account for it (under assumption of stationarity).

All the computations and simulations are made in R using for example dedicated packages to copula [10]. Results for the three different situations and a all copulae metadata (inputs, outputs, geoprocess) can then be derived. Besides the fact that these metadata can be derived, then allowing the meta-propagation based on copulae to be applied, a complete example implies reusing the same geoprocessing task in a different context. This illustrative part of the example is still in progress but thanks to the ROSELT project [14] this is possible in the best situation possible as the same whole geoworkflow of Figure 2 is applied in different observatories within circum-Saharan countries. We fitted the metatdata using a Tunisian site and we are planning to reuse the information for a site in Senegal, then allowing a comparison between meta-propagation and classical error propagation at least on a univariate case.

5 Conclusion

As geospatial data and geoprocessing services consuming them are to become the building blocks for future large scientific geoworkflows, it is natural to question on how the current error propagation techniques are going to cope. Alike a cascade method, meta-propagation, putting the difficult task up-front at the metadata level, before assembling these large workflows, propose an alternative to the computationally intensive error propagation. Nonetheless some difficult tasks still remain at the atomic level if comprehensive uncertainty has to be used. Copulae distributions as potential metadata encapsulating the expected dependence structures for the inputs or the outputs or for both are seen as leveling down the pain of estimating the appropriate metadata for the geoprocesses. An experiment looking at the feasibility and benefits of the three metadata approaches described in the paper using copula fitting is built on an agro-ecological model for desertification monitoring. Non-parametric copula estimation is used but this could be coupled with expert elicitation at least for the univariate margins [8].

References

[1] D. Li, J. Zhang, and H. Wu. Spatial data quality and beyond. *International Journal of Geographical Information Science*. 26 (12) : 2277–2290, 2012.

[2] M. Brown, S. Sharples, J. Harding, C. J. Parker, N. Bearman, M. Maguire, D. Forrest, M. Haklay and M.J Jackson. Usability of Geographic Information: Current Challenges and Future Directions. *Applied Ergonomics*, (in press) 2013.

[3] S.A.B. Cruz, A.M.V. Monteiro and R. Santos. Automated Geospatial Web Services Composition Based on Geodata Quality Requirements. *Computers and Geosciences*, 47:60-74, 2012.

[4] X. Yang, J. D. Blower, L. Bastin, V. Lush, A. Zabala, J. Masó, D. Cornford, P. Díaz and J. Lumsden. An Integrated View of Data Quality in Earth Observation. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1983):online, 2013

[5] A.U Frank. Analysis of dependence of decision quality on data quality. *Journal of geographical systems*, 10(1): 71-88, 2008.

[6] D.G. Leibovici, A Pourabdollah and M.J. Jackson. Meta-propagation of uncertainties for scientific workflow management in Interoperable spatial data infrastructures. In *Proceedings of the European Geosciences Union, General Assembly*, Vienna, Austria, 2011.

[7] D.G. Leibovici, A. Pourabdollah and M.J. Jackson. Which spatial data quality can be meta-propagated? *Journal of Spatial Science*, (to appear), 2013.

[8] L. Bastin, D. Cornford, R. Jones, G.B.M. Heuvelink, E. Pebesma, C. Stasch, S. Nativi, P. Mazzetti and P. Williams. Managing uncertainty in integrated environmental modelling: The UncertWeb framework. *Environmental Modelling and Software*, 39:116-134, 2013.

[9] A. Saltelli, M. Ratto, A. Terry, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, S. Tarantola. *Global Sensitivity Analysis. The Primer.* John Wiley & sons, Chichester, 2008.

[10] I. Kojadinovic and J. Yan. Modeling Multivariate Distributions with Continuous Margins Using the copula R Package, *Journal of Statistical Software*, 34(9):1-20, 2010.

[11] A. Possolo. Copulas for uncertainty analysis. *Metrologia*, 47(3):262–271, 2010.

[12] D. Kurowicka and RM. Cooke *Uncertainty Analysis with High Dimensional Dependence Modelling*. John Wiley & Sons, 2006

[13] T. Mikosch. Copulas: Tales and facts. *Extremes*, 9(1):. 3-20, 2006 (with interesting discussions)

[14] M. Loireau, M. Sghaier, M. Fetoui, M. Ba, M. Abdelrazik, J-M. d’Herbès, J-C. Desconnets, D. Leibovici, S. Debard, and E. Delaître. Système d’Information sur l’Environnement à l’échelle locale (SIEL) pour évaluer le risque de désertification : situations comparées circum-sahariennes (réseau ROSELT). *Sècheresse*, 18(4): 328-35, 2007.

[15] G.B.M. Heuvelink. J.D. Brown, and E.E. van Loonvan. A probabilistic framework for representing and simulating uncertain environmental variables. *International Journal of Geographical Information Science*, vol. 21, no. 5, pp. 497–513, 2007.