Quantifying the uncertainty in digital soil class maps developed using model-based approaches

M. A. Nelson[A,C], I. O. A. Odeh[A], T. F. A. Bishop [A] and N. Weber[B]

[A]Faculty of Agriculture, Food and Natural Resources, The University of Sydney, Sydney, NSW, Australia.
[B]School of Mathematics and Statistics, The University of Sydney, Sydney, NSW, Australia.
[C]Corresponding author. Email michael.n@usyd.edu.au

Abstract
Digital soil assessments use digital soil class maps as inputs for modelling soil processes. The accuracy of the digital soil class maps is required for the outputs digital soil assessments to be most useful. Few studies have considered the quantification of uncertainty in digital soil class maps. The use of model based approaches means the uncertainty in parameter estimates may be quantified. This study considers the effect of this model parameter uncertainty on digital soil class map, and provides a methodology to perform conditional simulation for a digital soil class map when model based approaches are used. We consider a multinomial logistic regression model and a Poisson multinomial generalised linear spatial model. The results of both the logistic regression and generalised linear spatial models show uncertainty is greatest near the boundary of soil classes. The simulated soil class maps can be used as input for digital soil assessments. The computational burden of deriving the required Fisher information matrices from the generalized linear spatial models is high, with further study required into methods of approximation.

Key Words
Geostatistics, accuracy, maximum likelihood, Monte Carlo Markov Chain.

Introduction
Digital soil class maps provide an important an important source of soil information from which harder to measure soil properties can be inferred. As such digital soil class maps form an essential component of the *soil attribute space inference system* thus forms the foundations of the digital soil assessment framework (Carrè *et al.* 2007). A significant advantage of the digital soil mapping framework is that predictions made should be accompanied by estimates of associated uncertainty (McBratney *et al.* 2003). For digital soil assessment, both the predictions contained within a digital soil map, and the associated uncertainty, are vital (Carrè *et al.* 2007).

The four main sources of uncertainty in a digital soil maps are input, model, positional and analytical (Bishop *et al.* under review). In practice, the enumeration and quantification of the uncertainty has lagged behind the development of digital soil mapping techniques, in particular for digital soil class mapping. While numerous papers have considered each of the four main sources of uncertainty four soil property mapping and Bishop *et al.* (under review) have considered relative contributions when all sources are considered simultaneously, few papers have investigated the uncertainty associated with digital soil class mapping predictions.

Model-based geostatistics (Diggle *et al.* 1998) provides a unified generic approach that merges geostatistical methods with generalized linear mixed models (GLMMs). The generalized linear spatial model (GLSM, Diggle *et al.* 1998; Zhang 2002) is a GLMM where the random effects are spatially correlated. Nelson *et al.* (2009) developed a GLSM for digital soil class mapping. The use of model-based geostatistics for digital soil class mapping allows the evaluation of prediction variance whilst also providing estimates of parameter variance which can be used to quantify uncertainty from the model parameters (Dowd and Pardo-Igùzquiza 2002).

The assessment of model uncertainty for a digital soil map has usually taken the form of conditional simulation by sequential Gaussian simulation given a set of model parameters (Goovaerts 1997), in this study, however, we will use the alternative approach suggested by Dowd and Pardo-Igùzquiza (2002). We investigate the effect of parameter estimate uncertainty on predictions in a digital soil class map developed using model-based approaches – both a Poisson multinomial generalized linear spatial model and a multinomial logistic regression model.

Methods

Parameter estimation uncertainty under model-based approaches

Under a model-based approach to digital soil class mapping, inference by maximum likelihood allows the estimation of the uncertainty in the model parameters. The main premise of this approach is the asymptotic multivariate normality of the parameter estimates, whose variance-covariance matrix is approximated by the Fisher information matrix (Dowd and Pardo-Igùzquiza 2002, McCulloch and Searle 2001)

$$\theta \sim \mathrm{A}N\left(\hat{\theta}, \mathrm{I}\left(\hat{\theta}\right)^{-1}\right)$$
$$\mathrm{I}(\theta) = -E\left[\frac{\partial^2 \ln L_c}{\partial\theta\partial\theta'}\right]. \qquad (1)$$

where θ are the model parameters and is $\ln L_c$ is the likelihood function. For generalized linear models (GLMs) and GLMMs the estimation of the Fisher information matrix (Equation 1) is computationally burdensome, however numerical methods can be used. For a GLSM, the variance parameters and trend parameters are independent, thus the information matrix is block diagonal (Zhang 2002).
By taking numerous realizations of the model parameters given these distributional assumptions, simulations of a digital soil class map can be produced.

Study area

As a case study, we map a small area of the Upper Namoi river catchment in NSW, Australia. The soil data is from a small section of the Curlewis Soil-Landscape Map (Banks 1994) with four mapping units present. Four environmental covariates were considered for use in the predictive model, SRTM DEM, γ radiometric %K, NDVI and the multi-resolution valley bottom flatness index (MRVBF, Gallant and Dowling 2003).

Model specifications

For the multinomial logistic regression model, a stepwise procedure using Akaike Information Criteria (AIC) was used to select the most parsimonious model, for which the observed information matrix for the parameters was estimated, and 200 realizations drawn. The Poisson multinomial GLSM models each soil class as an independent Poisson variable, with spatial random effects introduced for each soil class (Nelson *et al.* 2009). As with the multinomial logistic regression model, the observed information matrix was approximated for the AIC selected most parsimonious model, and from which 50 realizations of the parameters were drawn. We consider the effect of this uncertainty on the most likely soil class at each location on a digital soil class map. All statistical analyses were performed in R (R Development Core Team 2008), with the GLSM implemented using the geoRglm package (Christensen and Ribeiro 2002).

Results

Multinomial logistic regression

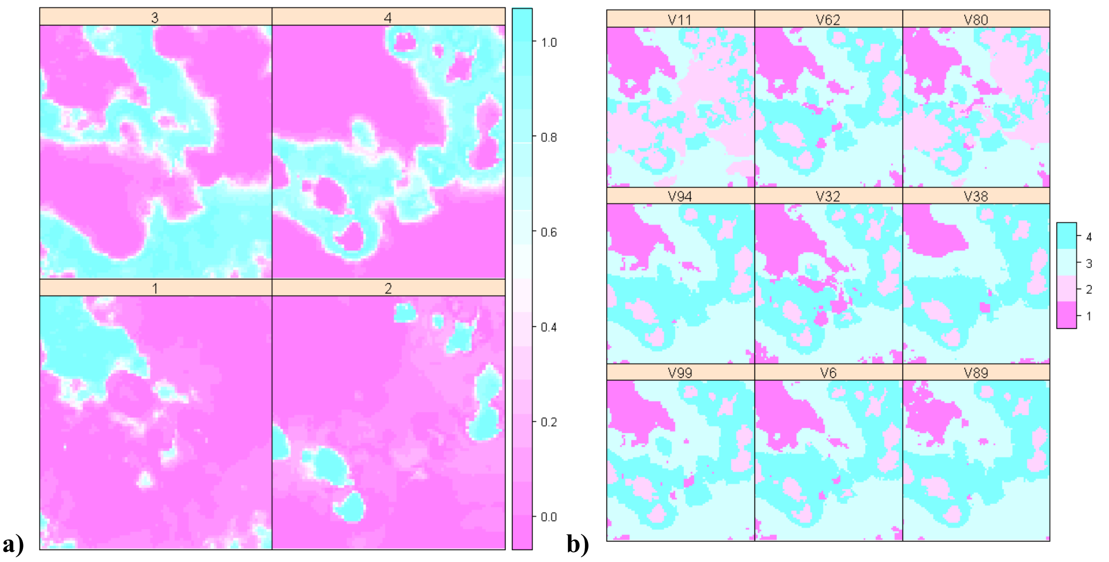
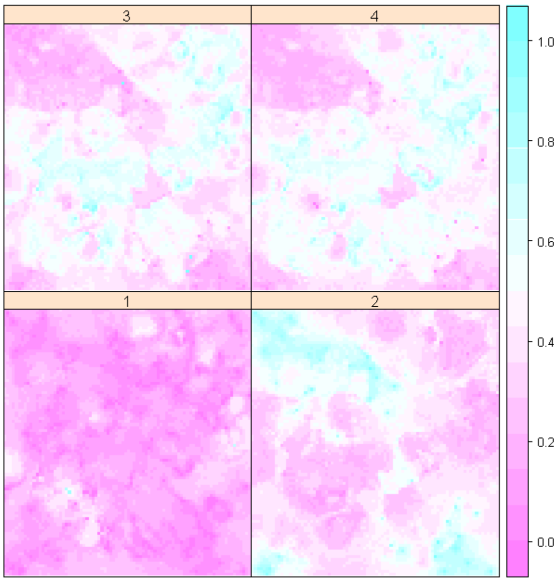


Figure 1. a) Proportion of simulations at each prediction location where each soil class is most likely for a digital soil class map produced using a multinomial logistic regression model; b) Nine simulations of the digital soil class map using realizations drawn from the estimated distribution of the parameter estimates.

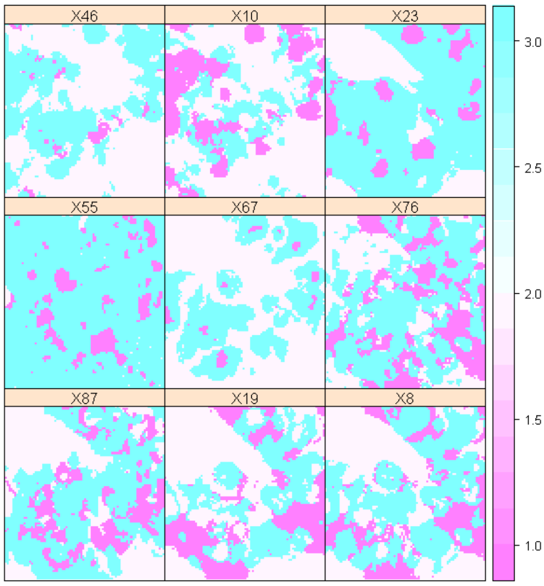
Figure 1 (a) shows the effect of parameter estimate uncertainty on the most likely soil class at each location of the digital soil class map. This shows the greatest effect nearest the boundary of the soil classes. Figure 1 (b) shows 9 simulations of the digital soil class map using different realizations of the parameter estimates. The main feature of Figure 1 (b) is the variation in the extents of classes 2 and 4 in the middle section of the map, highlighting the difficulty in differentiating between these classes using this model, and the uncertainty in the resultant digital soil class map.

Poisson multinomial GLSM

Figure 2 (a) shows the effect of parameter estimate uncertainty on the most likely soil class at each location of the digital soil class map. The main feature of this is the lack of prediction of class 1. Figure 2 (b) shows 9 simulations of the digital soil class map using different realizations of the parameter estimates. Again the lack of class 1 predictions is evident. These are a result of the difficulty in approximating the Fisher information matrix under a GLSM. Figures 1 (a) and 2(a) show similar results for all classes except class 1.



a)



b)

Figure 1. a) Proportion of simulations at each prediction location where each soil class is most likely for a digital soil class map produced using Poisson multinomial GLSM; b) Nine simulations of the digital soil class map using realizations drawn from the estimated distribution of the parameter estimates.

Discussion and Conclusion

The distributional assumptions of parameter estimates are valid as large sample asymptotes (McCulloch and Searle 2001). The sample size for this study (100) may not be large enough. Parameter estimation and the derivation of the Fisher information matrix under a GLMM are not trivial computations. Further study on the various approaches and methods is required to find a method that is both accurate and computationally reasonable. We have shown that for a digital soil class map produced using model based approaches, conditional simulation using the uncertainty in the model parameter estimation provides a useful assessment of the map uncertainty that may be used in digital soil assessments.

References

Banks R (1994b) 'Soil Landscapes of the Curlewis 1:100 000 Sheet: Map'. (Department of Conservation and Land Management: Sydney).

Bishop TFA, Nelson MA, Triantifilis J, Odeh IOA (2010) Enumeration of all sources of uncertainty in a digital soil map. *European Journal of Soil Science* (under review)

Carrè F, McBratney AB, Mayr T, Montanarella L (2007) Digital soil assessments: Beyond DSM. *Georderma* **142,** 69 - 79.

Christensen OF, Ribeiro Jr PJ (2002) geoRglm: A package for generalized linear spatial models. *R-News* **2,** 26-28.

Diggle PJ, Tawn JA, Moyeed RA (1998) Model-based geostatistics. *Journal of the Royal Statistical Society Series C-Applied Statistics* **47,** 299-326.

Dowd PA, Pardo Igùzquiza E (2002) The Incorporation of Model Uncertainty in Geostatistical Simulation.

Geographical and Environmental Modelling **6,** 147-169.
Gallant JC, Dowling TI (2003) A multiresolution index of valley bottom flatness for mapping depositional areas. *Water Resources Research* **39,** ESG41-ESG413.
Goovaerts P (1997) 'Geostatistics for natural resource evaluation'. (Oxford University Press: New York).
McBratney AB, Santos MLM, Minasny B (2003) On digital soil mapping. *Geoderma* **117,** 3- 52.
McCulloch CE, Searle SR (2001) 'Generalized, linear and mixed models' (John Wiley and Sons: New York)
Nelson MA, Bishop TFA, Odeh IOA, Weber N (2009) A generalized linear spatial model for digital soil class mapping [Digital soil class mapping using model-based geostatistics]. In 'Proceedings of Pedometrics 2009 Conference, Beijing, 60-61'.
R Development Core Team (2008) 'R: A Language and Environment for Statistical Computing'. *(*R Foundation for Statistical Computing).
Zhang H (2002) On estimation and prediction for spatial generalized linear mixed models. *Biometrics* **58,** 129-136.