ELSEVIER

Ecological Modelling 154 (2002) 93–101

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecolmodel

Uncertainties in spatially aggregated predictions from a logistic regression model

P.W. van Horssen, E.J. Pebesma *, P.P. Schot

Department of Geography, The Netherlands Centre for Geo-ecological Research, Utrecht University, P.O. Box 80.115, 3508 TC Utrecht, The Netherlands

Received 7 January 2000; received in revised form 15 January 2002; accepted 4 March 2002

Abstract

This paper presents a method to assess the uncertainty of an ecological spatial prediction model which is based on logistic regression models, using data from the interpolation of explanatory predictor variables. The spatial predictions are presented as approximate 95% prediction intervals. The prediction model is based on logistic regression analysis of field data of a wetland area in the central parts of the Netherlands. The model predicts block average probability of occurrences of 78 wetland plant species for 500 m × 500 m blocks. The explanatory variables comprise groundwater chemistry, hydrological characteristics, and land use management. The uncertainty of the spatial model output is assumed to be a function of the uncertainty in the estimated regression coefficients and uncertainty in the interpolated values of explanatory variables. Monte Carlo analysis was used to assess the model output error due to uncertainty in both the regression coefficients and the explanatory variables. Correlation between errors in regression coefficients and spatial autocorrelation in explanatory variables are accounted for in the Monte Carlo analysis. Spatial patterns of the relative contribution of uncertainty of the regression coefficients to the total model uncertainty are presented. The patterns of the relative contributions of uncertainty to the total model uncertainty give information on the most effective way to reduce error, i.e. either by reducing uncertainty in the regression coefficients or in the interpolated input patterns. The spatial patterns and values of the 95% prediction intervals vary widely between species but are in general large and the relative contribution of the uncertainty of the regression coefficients is in general large (over 80%).

Keywords: Spatial prediction; Block kriging; Error propagation; Monte Carlo; Plant species

1. Introduction

There are scientific and practical needs to be able to predict the occurrence of species in ecosystems. Often, multiple logistic regression models are used for this. These models are based on empirical data sets of species abundance and environmental conditions (explanatory variables) and on the assumption of an unimodal symmetric response curve in relation to site conditions (e.g. Jongman et al., 1995; McCullagh and Nelder, 1989). Data are usually available for a limited set

* Corresponding author. Tel.: +31-30-253-3051; fax: +31-30-253-1145.

E-mail address: e.pebesma@geog.uu.nl (E.J. Pebesma).

0304-3800/02/$ - see front matter
PII: S0304-3800(02)00060-1

of sample sites and do not cover the study area completely. Therefore the estimated regression coefficients in the logistic regression models have associated estimation errors. Also the explanatory variables themselves have associated levels of uncertainty (Elston and Buckland, 1993; Turner et al., 1995; Dunning and Stewart, 1995).

Most often, the use of logistic regression models for predicting the effect of management scenarios at a given location—i.e. change of response to change of environmental conditions—assumes zero uncertainty in the explanatory variables; so that when the values of explanatory variables at the location are changed the model predicts an exact change in response.

Calculating predictions for every location in an area implies exact knowledge of explanatory variables every location. However, limited knowledge (limited number of sample points) combined with natural spatial variability may cause uncertainty in values of explanatory variables. Also, regression coefficients are subject to estimation error.

In our application of a prediction model we are not concerned with predicting occurrence probabilities for point locations (i.e. areas the size of individual observations, a few tens of squared meters) but rather with predicting average probabilities of occurrence for 500 m × 500 m square blocks. The reason for this is that in contrast to 'point locations', for these blocks statistically meaningful estimates may be obtained while retaining sufficient spatial resolution in the resulting maps for policy making.

Assessing uncertainty in predictions of logistic regression models when using spatially distributed explanatory variables as input, means considering the uncertainty in the regression coefficients of the logistic regression model as well as the uncertainty in the explanatory variables. In order to depict this uncertainty, the model output is presented as approximate 95% prediction intervals for block mean values.

The aim of this paper is to quantify the spatial distribution of 95% prediction intervals resulting from predictions with a logistic regression model, when uncertainties in explanatory variables and uncertainties in regression coefficients are taken into account.

2. Methods

2.1. Multiple logistic regression modeling

In logistic regression, the logit of the probability of occurrence $p(s_i)$ of an observation $y(s_i)$ (y at the spatial location s_i) is modeled as a linear function in p known explanatory variables x_j (s_i)

$$E(y(s_i)) = p(s_i),\ \text{logit}(p(s_i)) = \beta_0 + \sum_{j=1}^{p} x_j(s_i)\beta_j \tag{1}$$

with $\text{logit}(u) = \log(u/(u-1))$ and β_j, $j = 0,\ldots, p$ the $p+1$ unknown regression coefficients. Using the logit transform is one way to ensure that a back-transformed prediction (i.e. a species probability of occurrence) is constrained to its physical boundaries between 0 and 1.

Not all species respond to the same explanatory variables and usually a stepwise variable selection procedure is carried out to select essential variables. If it is assumed that species response to explanatory variables is unimodal, on the logit scale only first and second order linear models are considered for inclusion (Jongman et al., 1995). For instance, consider a species y that seems to respond only to changes in salinity S and nutrient content N, a second order logistic regression model for the observation at location s_i of this species would be:

$$\text{logit}(p(s_i)) = \beta_0 + \beta_1 S(s_i) + \beta_2 S^2(s_i) + \beta_3 N(s_i) + \beta_4 N^2(s_i)$$

with $S(s_i)$ and $N(s_i)$ the salinity and nutrient content respectively at location s_i.

2.2. Spatial prediction

At any location s_0, given the values of the explanatory variables x_j (s_0) and the estimates $\hat{\beta}_j$ of β_j, the (logit of the) occurrence of a plant species is predicted by:

$$\text{logit}(\hat{p}(s_0)) = \hat{\eta}(s_0) = \hat{\beta}_0 + \sum_{j=1}^{p} x_j(s_0)\hat{\beta}_j \tag{2}$$

and this value can be back-transformed using the inverse logit transform:

$$\hat{p}(s_0) = \text{logit}^{-1}(\hat{\eta}(s_0)) = \frac{\exp(\hat{\eta}(s_0))}{1 + \exp(\hat{\eta}(s_0))} \quad (3)$$

In standard applications, predictions are made for given (i.e. known) values of the explanatory variables and the prediction error of logit($\hat{p}(s_0)$) can be derived from the estimation error variances and covariances of $\hat{\beta}_j$.

For the prediction of probabilities of occurrences at 500 m × 500 m blocks B_0 we need to estimate the explanatory variables by interpolation from known data. Assuming that the spatial variation of these variables can be modeled as intrinsic random functions, their predicted values and associated prediction variances can be obtained for each location by using ordinary block kriging (Journel and Huijbregts, 1978; Cressie, 1993). Kriging has been proven to be a robust and useful method for spatial interpolation (Burrough and McDonnell, 1998). Block kriging (Journel and Huijbregts, 1978; Pebesma and Wesseling, 1998), as opposed to point kriging, is used to estimate 500 m × 500 m block average values of the explanatory variables

$$x_j(B_0) = \frac{1}{|B_0|}\int_{B_0} x_j(s)\,\mathrm{d}s$$

with $|B_0|$ the area of the block. When these estimated values of explanatory variables are used for predictions $\hat{p}(B_0)$ with Eq. (2) and Eq. (3), the uncertainty of these spatial predictions not only depends on the variances and covariances of the estimates of the regression coefficients but also on the variance of the estimates of the explanatory variables.

2.3. Estimation of uncertainty

Because standard regression software does not allow for prediction with random explanatory variables, a Monte Carlo simulation was performed for this. All error distributions were assumed to be normal on the logit scale. Mean and variances of explanatory variables and regression coefficients were used to simulate a simple random sample.

For each block, a prediction of $\hat{\eta}(B_0)$ was obtained by Eq. (2) using this simulated set of regression coefficients and explanatory variables as input. This ensemble of predictions was used to estimate the prediction variance, $\sigma^2(B_0)$, at each location. For a given block B_0, uncertainty is expressed as an approximate 95% prediction interval by:

$$[\text{logit}^{-1}(\hat{\eta}(B_0) - 2\sigma(B_0)), \text{logit}^{-1}(\hat{\eta}(B_0) + 2\sigma(B_0))] \quad (4)$$

The prediction variance at B_0, $\sigma^2(B_0)$, comprises the prediction variance as a result of uncertainty in the regression coefficients, $\sigma_r^2(B_0)$, and the variance as a results of uncertainty in explanatory variables, $\sigma_e^2(B_0)$. Assuming independence between the two variance parts, in our model the prediction variance can be decomposed as:

$$\sigma^2(B_0) = \sigma_r^2(B_0) + \sigma_e^2(B_0). \quad (5)$$

The relative variance contribution due to regression (RVC_r) is then calculated by:

$$\text{RVC}_r = \frac{\sigma_r^2(B_0)}{\sigma^2(B_0)} \times 100\% \quad (6)$$

3. Case study

The analysis of uncertainty of logistic regression models with spatially interpolated input was carried out for an area with wetlands in the central parts of the Netherlands. The data for this case study were taken from fieldwork carried out for the development of regional vegetation models (Barendregt and Wassen, 1989; Barendregt and Nieuwenhuis, 1993) and for the description of regional hydrology (Schot, 1989; Schot and Molenaar, 1992; Schot and Van Der Wal, 1992).

The regional vegetation models are based on a database with 306 sample locations in the area. At each location the abundance of 78 wetland plant species were recorded together with 21 environmental characteristics (Barendregt and Wassen, 1989). Thirteen of these 21 environmental variables describe chemical concentrations in groundwater: pH, HCO_3^-, Cl^-, SO_4^{2-}, PO_4^{3-}, NO_3^-, NH_4^+, Na^+, Mg^{2+}, Ca^{2+}, K^+, SiO_2 and Fe_{total}.

Hydrological characteristics are described by three variables: phreatic groundwater level, groundwater flux and seepage versus infiltration.

Soil texture on four depths (0–30, 30–60, 60–90, and 90–120 cm below surface level) is described and finally land use management is described in three classes (Barendregt and Wassen 1989). Sample locations in the area are given in Fig. 1. A stepwise multiple logistic regression analysis on this data set resulted in a unique regression model for each of the 78 plant species. Regression coefficients and covariances were estimated using generalized linear model theory (McCullagh and Nelder, 1989) and S-Plus software (Chambers and Hastie, 1993). For the regression analysis the chemical concentrations, phreatic levels and seepage or infiltration fluxes were log transformed

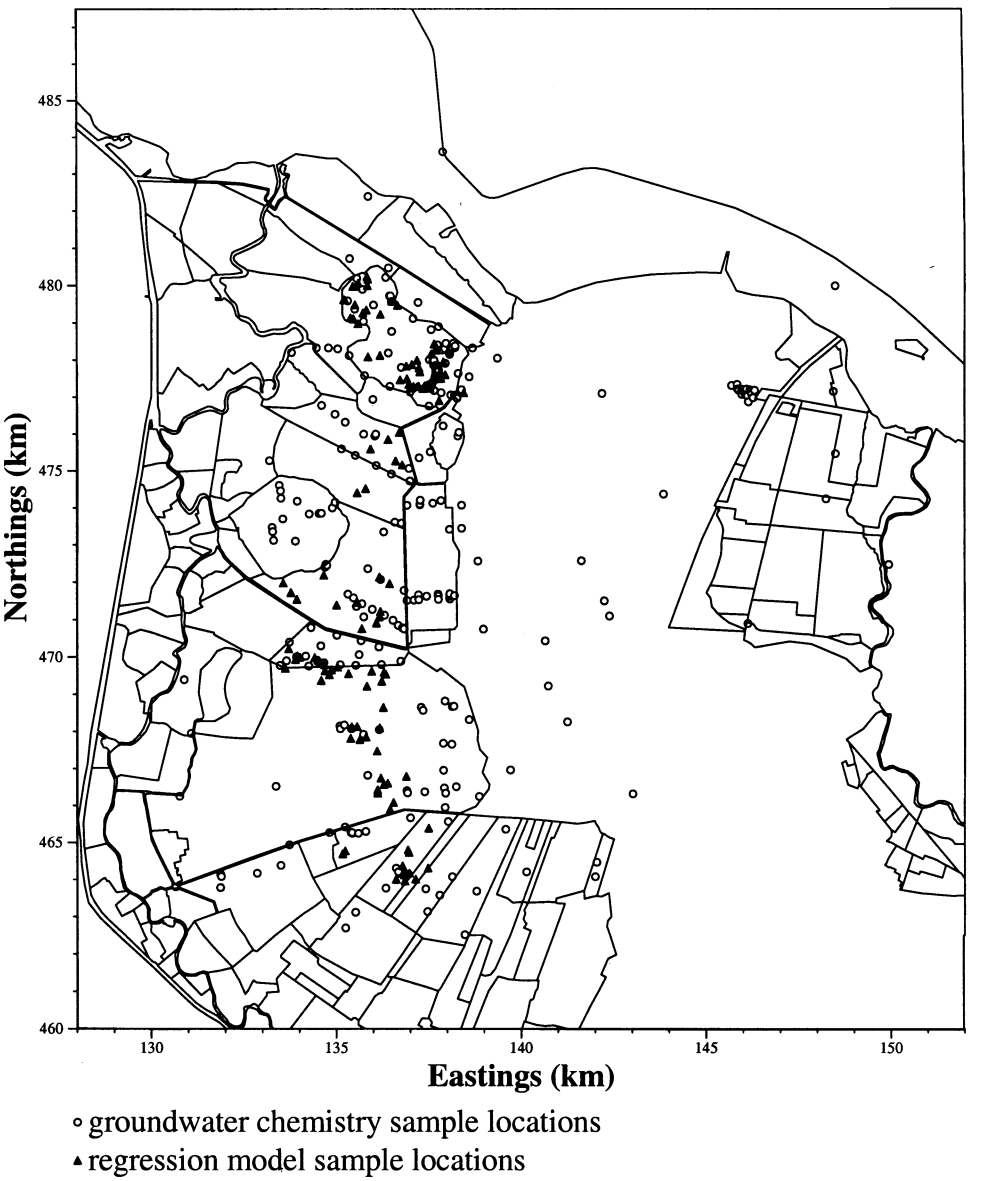

Fig. 1. Map with sample locations for plant species data and groundwater chemistry survey data.

while the other variables were treated as class variables (Barendregt and Wassen, 1989).

The regional patterns of the groundwater chemistry variables, used as explanatory variable for the regression models, are based on an independent set of groundwater chemical data (Schot and Van Der Wal, 1992). Fig. 1 shows locations of the groundwater chemical sample points with an observation well depths between 0 and 10 m minus surface level. Apart from pH, each of the groundwater chemical variables are log transformed to obtain symmetric distributions. Variogram analysis was conducted on each of the 13 groundwater chemical variables. These variograms were used to construct spatial predictions of 500 m × 500 m block mean concentrations using block kriging (Journel and Huijbregts, 1978). This method has been applied to similar problems (e.g. Myers et al., 1982; Pebesma and De Kwaadsteniet, 1997). Information on land use and soil type were taken from topographical and soil maps, while information on phreatic levels, infiltration and seepage were taken from Schot (1989), and was treated as known variables.

The mean and variance of estimated regression coefficients, their correlation matrix and the mean and variance of the 13 groundwater chemical variables were used to obtain a simple random sample of size 100. For each of the 78 plant species (regression models) Monte Carlo analysis was carried out and maps of the spatial patterns of the approximate 95% prediction limits and RVC_r were made.

The upper en lower prediction limits are presented in separated maps to facilitate comparison on an absolute scale. The RVC_r is presented in one map.

4. Results

4.1. Prediction intervals

For two species, *Calamagrostis canescens* and *Sphagnum palustre*, the approximate 95% prediction limits are given in Fig. 2. In case of negligible uncertainty, the upper and lower prediction limits would show identical values. The legend of Fig. 2 is deliberately made coarse to present large differences in prediction limits. If a finer legend had been chosen, more precise prediction limits could be distinguished in the maps, but then the general picture would be blurred in numerous gray tones.

The prediction intervals of *C. canescens* are wide, having an interval width of nearly 0.9–1.0 throughout the whole area, indicating large uncertainties in the spatial predictions. The prediction intervals of *Sphagnum palustre* are narrower, having a width between 0.1 and 0.5 for large parts of the study area, indicating less uncertain predictions. Two alternative approaches for merging the information of upper and lower prediction limits in a single map are given in Pebesma and De Kwaadsteniet (1997).

4.2. Uncertainty contributions

Spatial patterns of relative variance contributions of the regression model (RVC_r) to the model output of *C. canescens* and *S. palustre* are given in Fig. 3. If both sources of uncertainty contribute equally to the total uncertainty of the model, the maps in Fig. 3 should have a constant value of 50% throughout the whole area. Apparently this is not the case. Areas with percentages much lower or much higher appear in the maps of both species. Areas where the RVC_r is lower than 50% indicate locations where the uncertainty of the interpolated input data dominates the total uncertainty of the prediction, especially in areas where RVC_r is between 0 and 25%. Areas where the RVC_r is higher than 50% indicate locations where uncertainty due to regression coefficients dominates the total uncertainty, especially in areas where RVC_r is between 75 and 100%.

5. Discussion and conclusions

In this paper the uncertainty of a spatial logistic regression model has been quantified on the basis of two contributing uncertainty components. The error in the spatial predictions is presented as maps with approximate 95% prediction limits. It can be concluded that spatial predictions with logistic regression models and explanatory variables with some margin of uncertainty can lead to

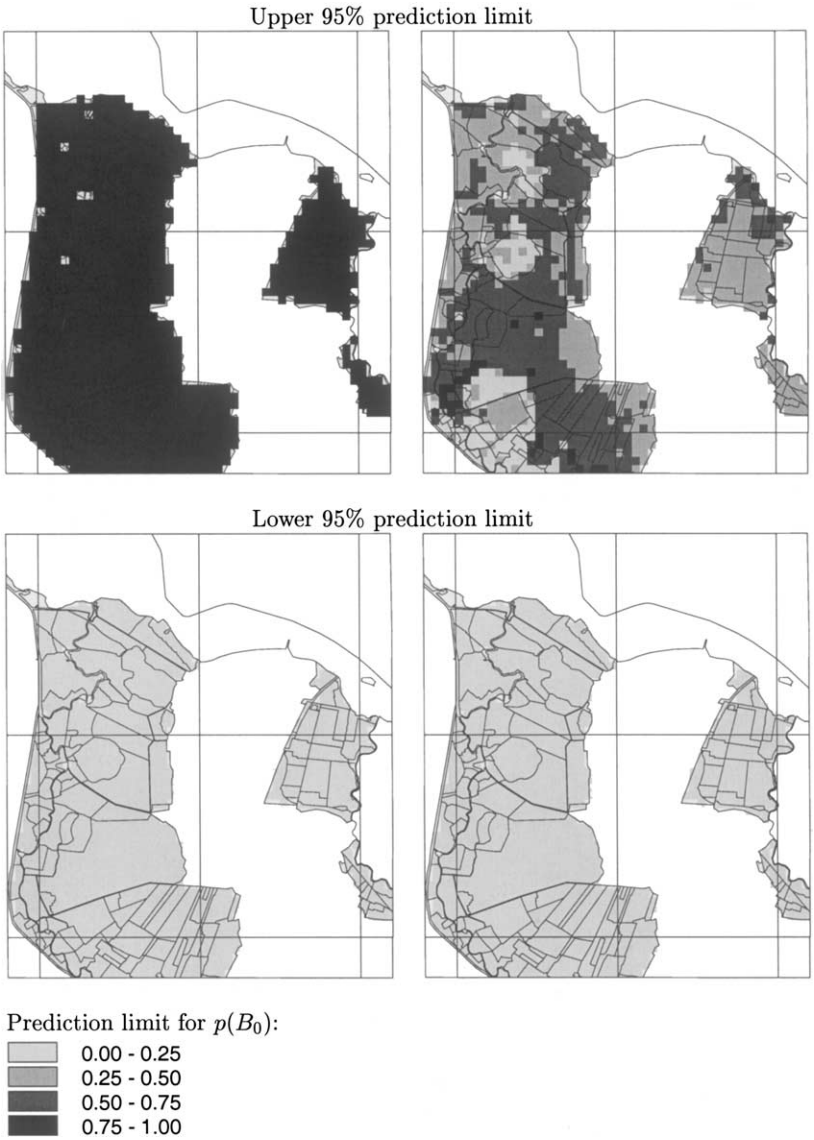

Fig. 2. Maps of approximate 95% prediction limits for *Calamagrostis canesces* (left) and *Sphagnum palustris* (right). Top row: upper prediction limit, bottom row: lower prediction limit.

considerably large 95% prediction limits. These confidence limits vary considerably, both spatially as well as for different logistic regression models (species prediction models). The relative contributions of uncertainty in both regression coefficients and explanatory variables are quantified and presented as maps. These maps tell us where the main source of uncertainty is located and as such indicate where improvements may be most cost effective. The uncertainty in the estimated regression coefficients will decrease by collecting more plant abundance data; the estimation error in the (interpolated) explanatory data will decrease by collecting more groundwater chemistry data. Fig. 3 shows that the dominant source of uncertainty varies spatially.

Presenting model results as 95% prediction intervals results in maps that are harder to read than the single map with the mean model result. However if model results do have a considerable margin of error, showing the mean value $\hat{p}(B_0)$ may give a false impression of the knowledge we have about possible occurrence of plants.

The uncertainty of model predictions was presented as block mean values. The choice of 500 m × 500 m blocks for interpolation was a trade-off: choosing smaller blocks would result in higher resolution of spatial patterns but in more inaccurate predictions (larger prediction variances). Choosing larger blocks would result in more accurate predictions (smaller prediction variances) on a low spatial resolution: much of the spatial pattern would average out.

The 95% prediction limits are approximate confidence intervals because several assumptions remain unverified. The most important of these are:

- prediction intervals were calculated using Eq. (2) assuming normality of prediction error and known prediction variances
- cross-correlations between explanatory variables were assumed to be zero
- regression residual errors were assumed to be spatially uncorrelated (resulting in zero predicted block mean regression residuals)
- the measurement error of explanatory variables is assumed to be zero at the plant measurement locations
- error contributions were quantified by assuming independence between uncertainty of explanatory variables and regression coefficients
- the regression models selected were assumed to be the 'true' models

These assumptions suggest that the error analysis given in this paper underestimates the true uncertainty. True prediction intervals will therefor be even wider than those presented here.

In addition to addressing the unverified assumptions above, a number of possible methodological enhancements would be:

- The Monte Carlo analysis should have been done for point locations on a very dense grid (e.g. 100 m × 100 m) and the back transformed point values $\hat{p}(s_0)$ should be aggregated to predicted block mean occurrences to avoid bias

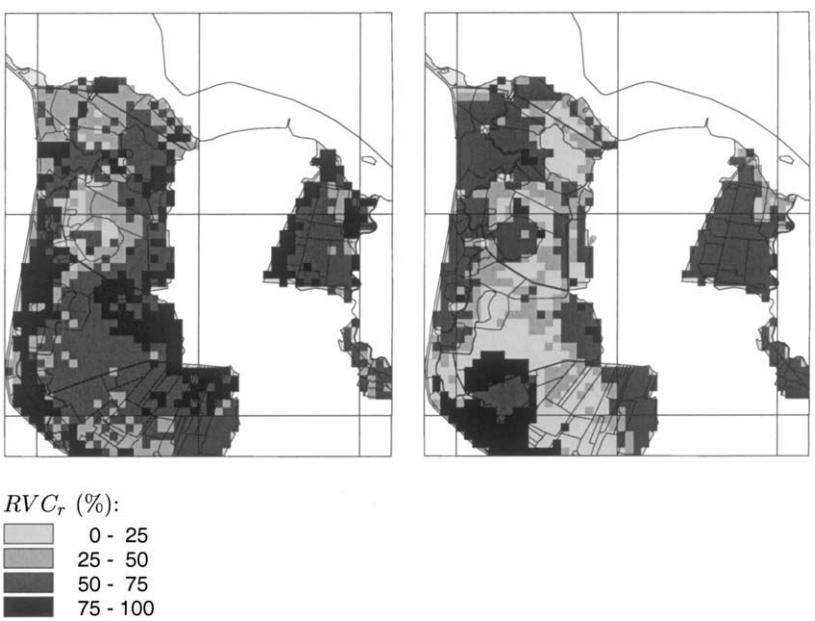

Fig. 3. Relative variance contribution of regression (RVC_r) for *Calamagrostis canescens* (left) and *Sphagnum palustris* (right).

resulting from back transforming block averages (King, 1991; Phillips and Marks, 1996; Heuvelink and Pebesma, 1998).

- The 95% prediction limits would have been calculated more accurately by using 97.5 and 2.5 sample percentiles and thus avoiding the assumption of normality of the error distribution. To do this, a larger Monte Carlo sample, and therefore a larger computational effort is needed.

The methodology presented here is not restricted to logistic regression modeling. More in general, when predictions can be expressed as some function of known predictor variables $x(s_0)$ and estimated regression coefficients $\hat{\beta}$,

$$\hat{p}(s_0) = f(x(s_0), \hat{\beta})$$

a Monte Carlo analysis can be carried out to quantify the effect of uncertainty with respect to $x(s)$ on $\hat{p}$. Here, $f(\cdot)$ may be anything from a generalized additive model to a neural network (Hastie et al., 2001). Addressing estimation error of β for such cases is usually done by bootstrapping (Efron and Tibshirani, 1993).

The methods used here to predict spatial patterns of explanatory variables and propagate errors through the regression model have not often been used in ecological studies. Phillips and Marks (1996) used a similar approach to perform uncertainty analysis in a spatially distributed model for the spatial prediction of potential evapotranspiration, however, they did not use block kriging nor stochastic simulation of the input variables. The use of geostatistics in ecological studies is usually confined to description of local spatial variance patterns of soil nutrients with variograms (e.g. Robertson, 1987; Schlesinger et al., 1996) but rarely to spatial prediction of abundance of organisms (Villard and Maurer, 1996). A notable exception is Gotway and Stroup (1997), who address spatial correlations in residuals for prediction under the generalized linear model framework.

The results of this study raises the question about the value of spatial predictions of logistic regression models without an uncertainty analysis. It is unclear whether the large prediction intervals in this type of spatial modeling are due to inadequate modeling methods, e.g. the incapability of logistic regression modeling to handle (spatially) correlated data, or due to the quality and spatial coverage of the data.

References

Barendregt, A., Wassen, M.J., 1989. Het Hydro-Ecologisch Model ICHORS (versie 2.0 and 3.0). Department of Environmental Studies, Utrecht, The Netherlands. In Dutch, with English summary.

Barendregt, A., Nieuwenhuis, J.W., 1993. ICHORS, hydro-ecological relations by multi-dimensional modelling of observations. In: Hooghart, J.C., Posthumus, C.W.S. (Eds.), The Use of Hydro-Ecological Models in the Netherlands, TNO Committee on Hydrological Research, Proceedings and Information no 47, Delft, The Netherlands, pp. 11–30.

Burrough, P.A., McDonnell, R.A., 1998. Principles of Geographical Information Systems. Oxford University Press, Oxford, UK.

Chambers, J.M., Hastie, T.J., 1993. Statistical Models. Chapman and Hall, New York, USA.

Cressie, N.A.C., 1993. Statistics for Spatial Data, revised edition. Wiley, New York, p. 900.

Dunning, J.B. Jr, Stewart, D.J., 1995. Spatially explicit population models: current forms and future uses. Ecol. Model. 67, 81–102.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman and Hall, London, p. 436.

Elston, D.A., Buckland, S.T., 1993, Statistical modelling of regional GIS data: an overview. Ecological Modelling, 67, 81–102.

Gotway, C.A., Stroup, W.W., 1997. A generalized linear model approach to spatial data analysis and prediction. J. Agric. Biol. Environ. Stat. 2 (2), 157–178.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Verlag, New York, p. 536.

Heuvelink, G.B.M., Pebesma, E.J., 1998. Spatial aggregation and soil process modelling. Geoderma 89, 47–65.

Jongman, R.H.G., Ter Braak, C.J.F., Van Tongeren, O.F.R., 1995. Data Analysis in Community and Landscape Ecology, new edition with corrections. Cambridge University Press, Cambridge, UK, p. 454.

Journel, A.G., Huijbregts, C.J., 1978. Mining Geostatistics. Academic Press, London, p. 600.

King, A.W., 1991. Translating models across scales in the landscape. In: Turner, M.G., Gardner, R.H. (Eds.), Quantitative Methods in Landscape Ecology. Springer-Verlag, New York, pp. 497–517.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, second edition. Chapman and Hall, London, p. 511.

Myers, D.E., Begovich, C.L., Butz, T.R., Kane, V.E., 1982. Variogram models for regional groundwater geochemical data. Math. Geol. 14 (6), 629–644.

Pebesma, E.J., De Kwaadsteniet, J.W., 1997. Mapping groundwater quality in the Netherlands. J. Hydrol. 200, 364–386.

Pebesma, E.J., Wesseling, C.G., 1998. Gstat, a computer program for geostatistical modelling, prediction and simulation. Comput. Geosci. 24, 17–31.

Phillips, D.L., Marks, D.G., 1996. Spatial uncertainty analysis: propagation of interpolation errors in spatially distributed model. Ecol. Model. 91, 213–229.

Robertson, G.P., 1987. Geostatistics in ecology: interpolating with known variance. Ecology 68, 744–748.

Schlesinger, W.H., Raikes, J.A., Hartley, A.E., Cross, A.E., 1996. On the spatial pattern of soil nutrients I desert ecosystems. Ecology 77, 364–374.

Schot, P.P., 1989. Grondwatersystemen en grondwater kwaliteit in het Gooi en de randgebieden. Vakgroep Milieukunde, Universiteit Utrecht, Utrecht, The Netherlands In Dutch.

Schot, P.P., Molenaar, A., 1992. Regional changes in groundwater flow patterns and effects in groundwater composition. J. Hydrol. 130, 151–170.

Schot, P.P., Van Der Wal, J., 1992. Human impact on regional groundwater composition through intervention in natural flow patterns and changes in land use. J. Hydrol. 134, 297–313.

Turner, M.G., Arthoud, G.J., Engstrom, R.T., Heil, S.J., Liu, J., Loeb, S., McKelvey, K., 1995. Usefulness of spatially explicit population models in land management. Ecol. Appl. 5 (1), 12–16.

Villard, M., Maurer, B.A., 1996. Geostatistics as a tool for examining hypothesized declines in migratory songbirds. Ecology 77, 59–68.