Geoderma 316 (2018) 56–69

Contents lists available at ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma

Mapping the probability of ripened subsoils using Bayesian logistic regression with informative priors

Luc Steinbuch[a,b,c,*], Dick J. Brus[d], Gerard B.M. Heuvelink[b,c]

[a] Wageningen Environmental Research (Alterra), PO Box 47, 6700 AA Wageningen, The Netherlands
[b] Soil Geography and Landscape group, Wageningen University, PO Box 47, 6700 AA Wageningen, The Netherlands
[c] ISRIC - World Soil Information, PO Box 353, 6700 AJ Wageningen, The Netherlands
[d] Biometris, Wageningen University, PO Box 16, 6700 AA Wageningen, The Netherlands

ARTICLE INFO

Keywords:
Bayesian statistics
Binomial logistic regression
Soil ripening
Informative priors
Soil mapping
Soil mapping uncertainty

ABSTRACT

One of the first soil forming processes in marine and fluviatile clay soils is ripening, the irreversible change of physical and chemical soil properties, especially consistency, under influence of air. We used Bayesian binomial logistic regression (BBLR) to update the map showing unripened subsoils for a reclamation area in the west of The Netherlands. Similar to conventional binomial logistic regression (BLR), in BBLR the binary target variable (the subsoil is ripened or unripened) is modelled by a Bernoulli distribution. The logit transform of the `probability of success' parameter of the Bernoulli distribution was modelled as a linear combination of the covariates soil type, freeboard (the desired water level in the ditches, compared to surface level) and mean lowest groundwater table. To capture all available information, Bayesian statistics combines legacy data summarized in a 'prior' probability distribution for the regression coefficients with actual observations. Our research focused on quantifying the influence of priors with different information levels, in combination with different sample sizes, on the resulting parameters and maps. We combined subsamples of different size (ranging from 5% to 50% of the original dataset of 676 observations) with priors representing different levels of trust in legacy data and investigated the effect of sample size and prior distribution on map accuracy. The resulting posterior parameter distributions, calculated by Markov chain Monte Carlo simulation, vary in centrality as well as in dispersion, especially for the smaller datasets. More informative priors decreased dispersion and pushed posterior central values towards prior central values. Interestingly, the resulting probability maps were almost similar. However, the associated uncertainty maps were different: a more informative prior decreased prediction uncertainty. When using the 'overall accuracy' validation metric, we found an optimal value for the prior information level, indicating that the standard deviation of the legacy data regression parameters should be multiplied by 10. This effect is only detectable for smaller datasets. The Area Under Curve validation statistic did not provide a meaningful optimal multiplier for the standard deviation. Bayesian binomial logistic regression proved to be a flexible mapping tool but the accuracy gain compared to conventional logistic regression was marginal and may not outweigh the extra modelling and computing effort.

1. Introduction

One of the first soil forming processes in marine and fluviatile clay soils is ripening, which is the irreversible change of physical and chemical soil properties, such as consistency, under influence of air. The ripening stage is an important factor in determining land use suitability. Moreover, it is also an indicator for forecasting soil shrinkage (Pons and Zonneveld, 1965). In the central western part of The Netherlands, clay soils have been waterlogged almost since deposition, and part of these soils are thus still ripening. The ripening process is ongoing, and as a result the current maps, created between 1960 and 1995, are getting outdated. These maps must be updated to accurately represent the current situation.

Soil ripening is mapped as a binary property, i.e. on each location, the soil is considered either 'ripened' or 'unripened'. It is unripened if any part of the profile (0–80 cm) contains unripened clay. If point

Abbreviations: AUC, Area under curve; BBLR, Bayesian binomial logistic regression; BIC, Bayesian Information Criterion; BLR, Binomial logistic regression; FPR, False Positive Rate; GLM, Generalized Linear Model; ML, Maximum likelihood; MLE, Maximum likelihood estimator; MLR, Multinomial logistic regression; MLW, Mean lowest ground water table; ROC, Receiver Operating Characteristics; TPR, True Positive Rate; UMF, Uncertainty multiplication factor

* Corresponding author at: Wageningen Environmental Research (Alterra), PO Box 47, 6700 AA Wageningen, The Netherlands.
E-mail address: luc.steinbuch@wur.nl (L. Steinbuch).

https://doi.org/10.1016/j.geoderma.2017.12.010
Received 4 October 2017; Received in revised form 5 December 2017; Accepted 9 December 2017

observations of soil ripening and maps of covariates related to soil ripening are available, a map of the probability of a ripened subsoil can be obtained by binomial logistic regression (BLR). In BLR, the logit transform of the 'probability of success' parameter of the Bernoulli distribution which represents in our case the probability that the soil is ripened, is modelled as a linear combination of covariates. With more than two classes the data follow a multinomial distribution and a similar approach, multinomial logistic regression (MLR), can be applied to map class probabilities. Kempen et al. (2009) and Vasques et al. (2014) applied MLR to map probabilities of multiple soil classes. Collard et al. (2014) compared MLR with classification trees and random forests. They found that MLR performed remarkably well for predicting soil classes. In contrast, Heung et al. (2016) showed MLR to perform worse for predicting soil classes in a comparison of ten machine learning approaches (e.g. logistic model trees, artificial neural networks).

BLR and MLR only use the observations of the variable of interest at the sampling points and the maps of the covariates. Models might better reflect reality and give more accurate predictions if we were able to exploit all available information in the model calibration process, especially in situations with scarce data. In particular, we may think of 'prior' knowledge about the regression coefficients of the BLR (MLR) model, which is not used in BLR (MLR) calibration. Bayesian statistics is equipped to capture all available knowledge by combining multiple information streams, i.e. information summarized in a 'prior' probability distribution of the model parameters, and information contained in the actual observations. For instance, Stanaway et al. (2011) used knowledge of plant properties and observation accuracy in Bayesian mapping of the risk of invasive plant species in Australia. Frigessi and Stander (1994) used deterministic terrain data to support Bayesian classification of satellite spectral images. Truong et al. (2014) used expert guesses of point-support variogram parameters to support Bayesian area-to-point kriging for remotely sensed air temperature.

To our best knowledge, Bayesian logistic regression has not yet been used to create soil property maps. In this research, we extensively explain, and apply, Bayesian binomial logistic regression (BBLR) for mapping clay soil ripening probability. In particular, we assess the added value of incorporating prior information derived from case-related legacy data. We investigate the added value of prior information with different degrees of information level in combination with different sample sizes of recent soil ripening data. Furthermore, this work includes a brief explanation of Bayesian generalized regression, the Metropolis algorithm and the validation statistics 'overall accuracy' and 'area under curve', with the purpose to familiarize soil scientists with these concepts.

2. Theory

2.1. The binomial logistic regression (BLR) model

Binary responses on discrete or continuous covariates can be modelled with the binomial logistic regression (BLR) model, which is an instance of the Generalized Linear Models family.

Let y_i, $i = 1…n$ be observations of a binary target variable, where each y_i equals 1 or 0 and n is the number of sampling locations. In BLR, the data are modelled as independent draws from a Bernoulli probability distribution:

$$y_i \sim \text{Bernoulli}\,(\pi_i) \tag{1}$$

with π_i the 'probability of success' parameter at the i-th sampling location. The logit transform of π_i is modelled by a linear combination of covariates:

$$\text{logit}(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = d_i^T\beta \tag{2}$$

where d_i is an $(m + 1)$ vector, the first element of which equals 1 and the remaining elements of which contain the values of m covariates at the i-th sampling location, and β is a vector of regression coefficients, including an intercept term. The inverse logit is written as:

$$\pi_i = \text{logit}^{-1}(d_i^T\beta) = \frac{exp(d_i^T\beta)}{1 + exp(d_i^T\beta)} \tag{3}$$

For all locations together, Eq. (2) can be written as Eq. (4) with π a column vector of π_1, …, π_n and D the design matrix, which contains the m covariates at the n sampling locations, including a column of leading ones:

$$\text{logit}(\pi) = D\beta \tag{4}$$

Having described π as a function of a vector of regression parameters β, we obtain an estimate of β that fits the data best, and use this calibrated BLR model for estimating the probability of a ripened subsoil at new locations. Note that we assume that the regression residuals are independent. In other words, we assume that the spatial structure in parameter π is fully captured by the covariates.

2.1.1. Estimation of regression parameters using maximum likelihood

Likelihood is a central concept in statistical model calibration, selection and comparison. In the scope of this paper, the likelihood $\mathscr{L}(\beta)$ equals the probability of the observations y as a function of the regression coefficients vector β, given in Eq. (5) (Collet, 1991):

$$\mathscr{L}(\beta) = p(y|\beta) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i} \tag{5}$$

Note that parameter π_i is a function of β as given in Eq. (3). Note also that $p(y\,|\beta)$ is a proper probability distribution when considered a function of y, i.e. it integrates to one over all possible values for y, but it is a likelihood when considered a function of β.

We calibrate a given model structure, i.e. a model with a given combination of covariates, on the data by finding the estimate $\hat{\beta}$ for β that maximizes the likelihood. Analytical solutions are not always available and numerical, iterative search algorithms are used instead (Collet, 1991). The uncertainty in $\hat{\beta}$ is expressed by its variance-covariance matrix:

$$\text{var}(\hat{\beta}) = (D^T\hat{V}D)^{-1} \tag{6}$$

with $\hat{V} = diag\,(\hat{\sigma}_1^2, \hat{\sigma}_2^2, ..., \hat{\sigma}_n^2)$, where $\hat{\sigma}_i^2 = \hat{\pi}_i\,(1-\hat{\pi}_i)$ and $\hat{\pi}_i$ the estimate for π_i resulting from plugging in $\hat{\beta}$ in Eq. (4). The diagonal of $var(\hat{\beta})$ contains the squared standard errors, i.e. the modelling variance of $\hat{\beta}$.

2.1.2. Estimation of probability of ripened subsoil at new locations

Point estimates for the model parameter $\hat{\pi}$ at a new location can be obtained by substituting $\hat{\beta}$ for β and d_0 for D in Eq. (3), with d_0 the covariate values at the new location. The modelling uncertainty in $\hat{\pi}$ at a new location as result of uncertainty in $\hat{\beta}$ can be investigated by the Monte Carlo method: simulate a large number of independent vectors with regression coefficients (using $\hat{\beta}$, $var(\hat{\beta})$ and a pseudo-random number generator, while assuming $\hat{\beta}$ has a multivariate normal distribution and accounting for correlation of the different regression coefficients) and calculate the corresponding $\pi_0^{(j)}$ using Eq. (4), where (j) indicates the iteration number, $(j) = 1, 2…r$ of r iterations. The resulting empirical distribution at the new location can be visualised by a histogram of all simulated $\pi_0^{(j)}$.

2.1.3. Selecting model structure

The regression model structure, i.e. the combination of covariates, may be chosen by minimizing the Bayesian Information Criterion (BIC) (Neath and Cavanaugh, 2012). Model selection criteria such as BIC favour models that explain the data well – quantified with a high maximum likelihood – but penalizes for model complexity, expressed as the number of model parameters m, which equals the number of