Remote Sensing of Environment 192 (2017) 45–57

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

Bayesian principal component regression model with spatial effects for forest inventory variables under small field sample size

Virpi Junttila[a,*], Marko Laine[a,b]

[a] Lappeenranta University of Technology, School of Engineering Sciences, P.O. Box 20, Lappeenranta FI-53851, Finland
[b] Finnish Meteorological Institute, P.O. Box 503, Helsinki FI-00101, Finland

ARTICLE INFO

Article history:
Received 17 August 2016
Received in revised form 5 January 2017
Accepted 27 January 2017
Available online xxxx

Keywords:
Remote sensing
Multicollinearity
Spatial correlation
MCMC
Forest inventory
Laser scanning
PCA
Bayesian analysis
Geostatistics

ABSTRACT

Remote sensing observations are extensively used for analysis of environmental variables. These variables often exhibit spatial correlation, which has to be accounted for in the calibration models used in predictions, either by direct modelling of the dependencies or by allowing for spatially correlated stochastic effects. Another feature in many remote sensing instruments is that the derived predictor variables are highly correlated, which can lead to unnecessary model over-training and at worst, singularities in the estimates. Both of these affect the prediction accuracy, especially when the training set for model calibration is small. To overcome these modelling challenges, we present a general model calibration procedure for remotely sensed data and apply it to airborne laser scanning data for forest inventory. We use a linear regression model that accounts for multicollinearity in the predictors by principal components and Bayesian regularization. It has a spatial random effect component for the spatial correlations that are not explained by a simple linear model. An efficient Markov chain Monte Carlo sampling scheme is used to account for the uncertainty in all the model parameters. We tested the proposed model against several alternatives and it outperformed the other linear calibration models, especially when there were spatial effects, multicollinearity and the training set size was small.

1. Introduction

Remotely sensed data, e.g. from satellites, digital aerial images, or airborne laser scanning, are increasingly used for mapping ecological variables over large geographical areas. Typical examples of such use are habitat and biodiversity monitoring (McDermid et al., 2009; Nagendra et al., 2013), lake water quality (Matthews et al., 2010) and forest inventory (Masek et al., 2015). Remotely sensed observations provide only indirect information of the area and modelling is needed for the interpretation of the data in terms of the variables of interest. In forest inventory, these variables include forest characteristics such as average forest biomass, median tree height, per hectare stem number or average timber volume, see e.g. Næsset (1997), Means et al. (1999), Rooker Jensen et al. (2006), and Magnussen et al. (2010). In some cases, direct physical models might be available (e.g. based on emission and scattering of light), but generally a simpler black-box type model to translate the remotely sensed data to the ecological variables is needed, e.g. linear regression (see the references above) or other methods (see e.g. Powell et al., 2010; Gleason and Im, 2012; Belgiu and Drăguţ, 2016). In forest inventory, this is typically done by linear regression.

A benefit of remotely sensed observations is that they can cover the whole spatial area of interest and the geographically located ecological variable can be predicted over the whole area in a pixel or some other sub-area level. To calibrate the model for prediction, i.e., to estimate the model parameters, remotely sensed data need to be accompanied by a set of field measurements of the ecological variables at chosen test locations. For instance, in area based prediction of forest inventory variables, the field measurements can be given as per hectare values estimated in circular field sample plots with a given radius. In forest inventory models, the number of field sample plots needed for accurate calibration can be several hundreds for an area between 10,000 and 100,000 hectares (Maltamo et al., 2011). The design of the field measurement locations has to account for not only the obvious statistical properties, but also the landscape properties such as mountainous terrain or thick forest, and it may be laborious and costly to reach these locations for the measurement work. Thus, to decrease the cost of the predictions, it is preferable to keep the number of field measurements to a minimum. This ambition for a small training set causes additional challenges to the

* Corresponding author.
E-mail addresses: virpi.junttila@lut.fi (V. Junttila), marko.laine@fmi.fi (M. Laine).

http://dx.doi.org/10.1016/j.rse.2017.01.035
0034-4257/

model parameter estimation process since the regression problem may become under-determined and easily suffer from the effects of over-training (Junttila et al., 2013).

LiDAR (Light Detection and Ranging) is an active remote sensing system based on laser light. In airborne LiDAR, a sensor in an aeroplane or a helicopter sends laser pulses towards the ground and records the time lapses between the launch of the beams and the return of the signals. In area based models, the LiDAR predictor variables are usually some statistical aggregates of the actual LiDAR pulse measurements over the geographical sub-areas. These variables are typically highly correlated and this multicollinearity can cause singularities in the model. If the number of LiDAR predictor variables is large compared to number of field sample plots, the multicollinearity can also cause unnecessary model over-training because, in some sense, the highly correlated predictor variables contain the same information about the response and thus add no new information, only redundant variables. A general approach to overcome problems caused by multicollinearity is to use variable selection algorithms or principal component regression. In a recent study, Junttila et al. (2015) achieved good results with a small training set and highly correlated multidimensional data by utilizing singular value decomposition combined with Bayesian regularization.

Many ecological variables are spatially correlated, which means that experimental units geographically close to each other are likely to be more similar than those far away. If the model explains this variability well, the model based predictions follow the same correlation. However, any lack-of-fit, which is inevitable in most linear calibration models, may produce spatially correlated model residuals, i.e., residuals geographically close to each other are more similar than those far away by residual sign and amplitude. In such occasions, the model performance is improved if the predictions are corrected toward those field measurements close to the prediction location.

In this paper, we build a linear model for prediction of the ecological variable of interest with a small number of field measurements which is both efficient and robust against modelling assumptions. We use a Bayesian approach that allows us to implement effective estimation of complex, hierarchically structured parameter associations appropriate for accommodating the strong multicollinearity typical in remotely sensed predictors and spatial autocorrelation among the model residuals. The method is applicable for many problems where strongly correlated data are translated to spatial observations with a linear model. In the proposed model, the problems caused by the multicollinearity of the predictors and by the small number of field measurements are overcome by using predictor orthonormalization and regularization. We utilize Bayesian regularization to emphasize those linear combinations of principal component predictors that explain most of the variability of the original predictor variables and that have predictive information on the ecological variable of interest. A general spatial dependency is allowed for the residuals of the model by a spatial random effect. The hierarchical model is estimated and uncertainty in the spatial model parameters is carried through to the predictions by using an efficient adaptive Metropolis Markov chain Monte Carlo (MCMC) algorithm.

We validate model performance by using both synthetic data with different noise levels and spatial correlation, and with real-world observations for forest inventory. In both cases, we assume a given design for the field plot locations and show how the spatial correlation structure and a full Bayesian treatment of model parameter uncertainties improve the model based predictions.

This work is based on earlier studies by Junttila et al. (2013), who used a similar spatial model, but with plug-in estimates of the spatial parameters instead of MCMC, and studies by Junttila et al. (2015), who used combination of singular value decomposition and regularization for regression parameters, but solved them with maximum likelihood estimation instead of MCMC. The parametric uncertainties of the two earlier papers are now dealt with using a sampling based approach. We show, by comparing the predictive power, that the approach chosen here outperforms the earlier methods in the presence of spatial correlation in the model error.

The article is organized as follows: we first define the proposed model in Section 2; the used datasets, both synthetic and real, are described and the validation procedure is explained in Section 3; the results of the validation are given in Section 4; and finally conclusions based on the results are given in Section 5.

2. Statistical methods

2.1. The proposed model

In this study, we use a linear regression model with a spatial random effect and hierarchical shrinkage prior for the regression parameters. The model combines the spatial modelling and singular value decomposition regularization described in Junttila et al. (2013) and Junttila et al. (2015). Instead of maximum a posteriori (MAP) estimates, the model parameters are estimated using Markov chain Monte Carlo (MCMC) simulation. With MCMC we can implement a hierarchical Bayesian model that can handle complex structured parameter associations and fully account for the uncertainty in all the model parameters for the model based predictions of the ecological variables of interest outside the training set.

We write our model as

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\eta} \sim N(\mathbf{0}, \mathbf{C}), \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \tau^2\mathbf{I}), \tag{1}$$

where the response $\boldsymbol{y}$ is a vector of the ecological variable of interest containing n observations, $\mathbf{X}$ is an $n \times (p+1)$ matrix that includes an intercept column and p columns of principal components of the remotely sensed data based variables, as described in Section 2.3, $\boldsymbol{\beta}$ is a $p+1$ vector of regression parameters, $\boldsymbol{\eta}$ is an n vector for the spatial random effect, and $\boldsymbol{\epsilon}$ is an n vector for non-spatial error. Both the random terms, $\boldsymbol{\eta}$ and $\boldsymbol{\epsilon}$, are assumed Gaussian. A full covariance matrix $\mathbf{C}$ defines the spatial correlation structure of the model residuals by using distances between the field measurement locations. The errors in $\boldsymbol{\epsilon}$ are assumed independent and identically distributed with variance τ^2.

Data at n locations containing the field measurements of the ecological variable are referred as the training set, while the locations where the variable needs to be predicted, are referred as the validation set. The predictors and the geographical coordinates are assumed to be known in each training set and validation set location.

To estimate the model parameters, we use hierarchical formulation to define the priors. To obtain the predictor regularization effect using the priors, we follow the formulation of Tipping (2001). For the regression parameter $\boldsymbol{\beta}_i, i = 0, 1, 2, \ldots, p$, the prior is zero mean Gaussian with inverse of variance $\boldsymbol{\alpha}_i$. The variance parameters $\boldsymbol{\alpha}_i$ are assumed to be unknown and they are estimated too. The prior for $\boldsymbol{\alpha}_i$ is defined by using a scaled χ^2 distribution and we have

$$\boldsymbol{\beta}_i \sim N(0, \boldsymbol{\alpha}_i^{-1}), \quad i = 0, 1, \ldots, p, \tag{2}$$

$$\boldsymbol{\alpha}_i \sim \chi^2(\nu_i, a_i), \quad i = 0, 1, \ldots, p. \tag{3}$$

The scaled χ^2 distribution is common in Bayesian analyses and can be defined by the standard Gamma distribution as $\chi^2(\nu, a) = \Gamma(\nu/2, a\nu/2)$ (see, e.g. Gelman et al., 2003). The scaled χ^2 parameterization is convenient in applications. We can interpret it as if knowing the value of $\boldsymbol{\alpha}$ to be a from ν (virtual) previous observations. In addition, it is the conjugate distribution for the inverse variance, which allows for the Gibbs sampling approach outlined below.