Available online at www.sciencedirect.com

ScienceDirect

Procedia Environmental Sciences 26 (2015) 123 – 126

Procedia Environmental Sciences

Spatial Statistics 2015: Emerging Patterns

Data fusion of remote-sensing and in-lake chlorophyll$_a$ data using statistical downscaling

Craig J. Wilkie[a]*, E. Marian Scott[a], Claire Miller[a], Andrew N. Tyler[b], Peter D. Hunter[b], Evangelos Spyrakos[b]

[a]*School of Mathematics and Statistics, University of Glasgow, Glasgow, G12 8QQ, UK*
[b]*School of Natural Sciences, University of Stirling, Stirling, FK9 4LA, UK*

Abstract

Chlorophyll$_a$ is a green pigment, used as an indirect measure of lake water quality. Its strong absorption of blue and red light allows for quantification through satellite images, providing better spatial coverage than traditional in-lake samples. However, grid-cell scale imagery must be calibrated spatially using in-lake point samples, presenting a change-of-support problem. This paper presents a method of statistical downscaling, namely a Bayesian spatially-varying coefficient regression, which assimilates remotely-sensed and in-lake data, resulting in a fully calibrated spatial map of chlorophyll$_a$ with associated uncertainty measures. The model is applied to a case study dataset from Lake Balaton, Hungary.

Peer-review under responsibility of Spatial Statistics 2015: Emerging Patterns committee

Keywords: statistical downscaling; Bayesian spatially-varying coefficients; chlorophyll$_a$; spatial calibration; MCMC.

1. Background

Chlorophyll$_a$, found in most plants and phytoplankton, is a green pigment which can be used as an indirect measure of phytoplankton biomass and hence of lake water quality. Its strong absorption of light at blue and red wavelengths allows quantification from water reflectance measurements made from Earth-observing satellites. Satellite imagery provides much better spatial coverage than traditional in-lake water samples, but must first be

* Corresponding author. Tel.: +44 141 330 2474.
E-mail address: c.wilkie.2@research.gla.ac.uk

Peer-review under responsibility of Spatial Statistics 2015: Emerging Patterns Committee
doi:10.1016/j.proenv.2015.05.014

calibrated, introducing the change-of-support problem: dense, grid-based remotely-sensed data must be adjusted using sparse point-scale in-lake samples.

This work is linked to the GloboLakes project (www.globolakes.ac.uk), a five-year consortium project investigating the state of lakes and their response to environmental change on a global scale. Lakes are vital components of the global biosphere, but they are vulnerable to climate- and human-induced change[1]. This means that a greater understanding of spatial and temporal patterns of change in lakes must be obtained.

Statistical downscaling is a method of data fusion, where information from two different sources, here namely remote-sensing and in-situ chlorophyll$_a$ data, is combined. Originally developed to resolve Global Climate Model output onto a scale required for impact assessment, these models have recently been applied in the air quality literature[2]. In this paper, the aim is to spatially calibrate the remote-sensing data using "in-situ" data obtained from in-lake, laboratory-analysed samples, which are assumed to be accurate within measurement error.

2. The model

The statistical downscaling model given in equation (1) is a Bayesian spatially-varying coefficient regression:

$$Y(\mathbf{s}) = \alpha(\mathbf{s}) + \beta(\mathbf{s})x(\mathbf{B}) + \varepsilon(\mathbf{s}), \qquad \varepsilon(\mathbf{s}) \overset{ind}{\sim} N(0, \sigma^2). \tag{1}$$

The response, in-situ data Y(**s**) at spatial location **s**, is regressed on remotely-sensed data x(**B**), where **B** is the remote-sensing grid cell in which in-situ location **s** lies. Errors ε(**s**) are assumed to be independent and Normally distributed. Intercept and slope coefficients are given Multivariate Normal distributions:

$$\boldsymbol{\alpha}(\mathbf{s}) \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\alpha)$$

$$\boldsymbol{\beta}(\mathbf{s}) \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\beta)$$

These distributions are given exponential spatial covariance structures:

$$\boldsymbol{\Sigma}_\alpha = (1/\tau_\alpha)\exp(-(\varphi_\alpha \mathbf{D}))$$

$$\boldsymbol{\Sigma}_\beta = (1/\tau_\beta)\exp(-(\varphi_\beta \mathbf{D}))$$

Here $\sigma_\alpha^2 = 1/\tau_\alpha$ and $\sigma_\beta^2 = 1/\tau_\beta$ are the spatial variances and φ_α and φ_β are the spatial decay parameters of the exponential covariance functions. **D** is a matrix of distances between spatial locations used in the model. The spatial decay parameters control how fast spatial correlation decreases as distance between points increases. The intercept and slope coefficients α(**s**) and β(**s**) are assumed to vary smoothly over space, meaning that the model assumes smoothly changing relationships between the response and explanatory variables over space. The spatial decay parameters are given Uniform prior distributions, with endpoints chosen for a sensible range of plausible values, while spatial variances are given uninformative Inverse-Gamma prior distributions:

$$\varphi_\alpha \sim \text{Unif}(0.05, 20)$$

$$\varphi_\beta \sim \text{Unif}(0.05, 20)$$

$$\tau_\alpha \sim \text{Gamma}(0.001, 0.001)$$

$$\tau_\beta \sim \text{Gamma}(0.001, 0.001)$$

These priors are chosen to be uninformative, meaning that they have little influence on the posterior distributions, compared to the data. Model (1) is related to a model developed by Berrocal et al.[2], referred to here as model (2), which was used to calibrate modelled air quality data using data from ground pollution monitoring stations. Unlike

model (2), model (1) does not assume that slope and intercept parameters co-vary over space, since there is no reason to assume that a change in slopes will be related to a change in intercepts, for chlorophyll$_a$ data.

3. Case study

3.1. Data

The model was applied to a case study lake, Lake Balaton in Hungary. There are monthly-averaged chlorophyll$_a$ data available for 7607 remote-sensing pixels and also in-situ data available for 9 in-situ locations, all recorded in mg/m^3. The data for July 2008 are analysed here with in-situ data obtained from the GloboLakes project and remote-sensing data obtained from the Diversity II project (www.diversity2.info/products/inlandwaters/). Exploratory analysis suggested a log transformation was appropriate for both in-situ and remotely-sensed data, so log(chlorophyll$_a$) data are used in all analyses.

Lake Balaton, Europe's largest shallow lake[3], is of particular interest, as it is a popular leisure site, but is prone to eutrophication. Steps were taken in the 1980s to improve the situation[4], after the water quality had deteriorated dramatically in the 1960s and 70s[3], and it is of interest to investigate how the health of the lake is improving.

Throughout the analysis, it is assumed that the in-lake data are accurate within measurement error, as they are obtained directly from water samples taken from the lake and analysed in a laboratory. However, the remote-sensing data are converted from satellite Earth surface reflectance data via an algorithm, and hence must be calibrated using the in-situ data, before being used to identify patterns in chlorophyll$_a$.

3.2. Results

The model was fitted using Markov-Chain Monte Carlo, using JAGS[5] via R, to a dataset of 7607 remotely-sensed pixels and 9 in-lake sampling locations. A plot showing the original remotely-sensed data and in-situ data and also the resulting calibrated spatial surface is produced below (Fig. 1).

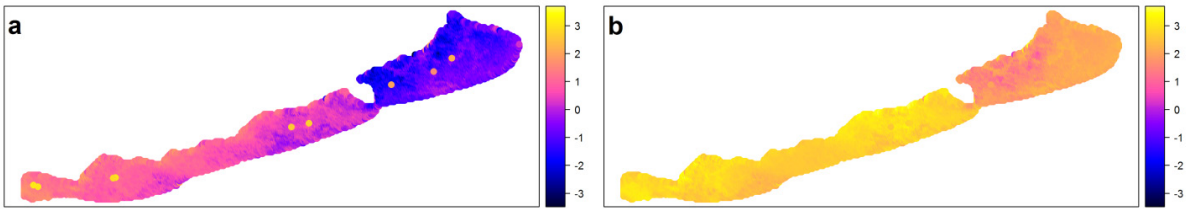

Fig. 1. (a) Log-transformed remotely-sensed data with in-situ data superimposed; (b) downscaled prediction surface.

The right plot in Fig. 1 shows the downscaled surface spatially calibrated, with different areas of the lake adjusted at different rates, based on the relationship with the in-situ data compared to the original remotely-sensed data (left plot). It can be seen that the resulting surface retains most features of the remotely-sensed dataset, which should allow the identification of local-scale events and changes in chlorophyll$_a$ levels.

Results from model (1) were compared to those from model (2) and also a Bayesian kriging model applied to the in-situ data, referred to as model (3). Predictions were made using a leave-one-out and predict method, where one location was removed at a time and the log(chlorophyll$_a$) was predicted. Root mean square errors (RMSEs) were then calculated. These were 0.38, 0.99 and 0.24 for models (1), (2) and (3), respectively. This analysis was repeated for several months with available data, giving similar results. The lower RMSE for model (1) than for model (2) suggests that the assumption of co-variation may not be appropriate for chlorophyll$_a$ data, at least for this dataset. It can be noted that the RMSE for Bayesian kriging is lower here than for the two downscaling models. However, the resulting smoothed surface does not take into account the remotely-sensed data and so produces an overly-smoothed surface based only on in-situ data, which would not allow the identification of local bloom events. Both downscaling models were able to produce a fully calibrated spatial map of log(chlorophyll$_a$), along with associated measures of

uncertainty. However, model (1) was found to have lower RMSEs and so may be preferred in this chlorophyll$_a$ context.

4. Discussion

Novel contributions in this paper include the direct modelling of $\alpha(\mathbf{s})$ and $\beta(\mathbf{s})$, rather than assuming they co-vary. The remotely-sensed data used here can be considered as being observed data, rather than being modelled as a smooth surface[2]. The method of statistical downscaling applied here has assimilated uncalibrated remote-sensing data and in-lake laboratory-analysed data from a very small number of spatial locations, to produce a fully calibrated spatial map of log(chlorophyll$_a$), with associated uncertainty measures. This should allow the assessment of spatial changes in chlorophyll$_a$ levels, which would not be possible solely from investigating in-lake data.

Future work will develop the model to incorporate spatio-temporal covariate information. Other developments include multivariate statistical downscaling, where several variables are downscaled simultaneously, to improve predictions by including information from variables related to chlorophyll$_a$. This will introduce a further change-of-support problem, as remote-sensing data are resolved on multiple grid sizes.

Acknowledgements

CM and EMS were partly funded for this work through the NERC GloboLakes project (NE/J022810/1). CJW acknowledges the support of ISD and also the School of Mathematics and Statistics, University of Glasgow. We acknowledge the ESA DUE DIVERSITY II project for providing ENVISAT data and derived indicator products.

References

1. Tranvik LJ, Downing JA, Cotner JB, Loiselle SA, Striegl RG, Ballatore TJ et al. Lakes and reservoirs as regulators of carbon cycling and climate. *Limnol Oceanogr* 2009;**54**: 2298-2314.
2. Berrocal VJ; Gelfand AE, Holland DM. A spatio-temporal downscaler for output from numerical models. *J Agr Biol Envir St* 2010;**15**:176-197
3. Padisák J, Reynolds CS. Selection of phytoplankton associations in Lake Balaton, Hungary, in response to eutrophication and restoration measures, with special reference to the cyanoprokaryotes. *Hydrobiologia* 1998;**384**:41-53.
4. Tátrai I, Mátyás K, Korponai J, Paulovits G, Pomogyi P. The role of the Kis-Balaton Water Protection System in the control of water quality of Lake Balaton. *Ecol Eng* 2000;**16**:73-78.
5. Plummer M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In: Hornik, K., Leisch, F., & Zeileis, A, editors. *Proceedings of the 3rd international workshop on distributed statistical computing* 2003;**124**.