Some issues in modelling biodiversity using spatially modelled covariates

Hideyasu Shimadzu¹, Scott D. Foster²

¹Marine and Coastal Environment Group, Geoscience Australia, ²CSIRO Wealth from Oceans Flagship and CSIRO Mathematical and Information Sciences.

Statistical models have enhanced the understanding of the relationship between biodiversity and the environment. Typically, some sort of regression analysis is performed where physical variables are covariates. It is frequently the situation that the covariates are not observed; they are spatial predictions. This study indicates that this process may bias the statistical distribution and the resulting parameter estimates if the variance of the predictions is ignored.

Great Barrier Reef data

Data comprises of 1189 sites, where biological and physical variables were measured (Pitcher et al. 2007, see Figure 1). The biological outcome variables were presence/absence of a particular species, and species richness (number of species) in a benthic sled sample. The physical variables used in this study were depth, %carbonate and %mud. We randomly selected 200 sites from this complete set to mimic performing a biological survey (without measuring physical data).

Model and approximate bias

Let \( Y_i \) be a biological response such as species presence/absence or richness, and let \( x_i \) be the vector of physical covariates at the survey site \( i \). A generalised linear model is often used with

\[
E[Y_i | x_i] = h(y_i) = h(\theta_i) = h(\theta_i^{g(x_i, \tau)})
\]

where \( h \) is an inverse link function and \( \tau \) is a \( p \times 1 \) vector of unknown parameters. However, the physical covariates are commonly not observed but are estimated by spatial predictions, \( \tilde{x}_i \), based on observations at other sites. Allowing for variance from the spatially predicted covariate, the approximate mean and covariance of the outcomes are

\[
E[E[Y_i | x_i] | x_i] = E[y_i | \tilde{x}_i] = \tilde{h}(\tilde{\theta}_i) = \tilde{h}(\tilde{\theta}_i^{g(\tilde{x}_i, \tau)})
\]

and

\[
Cov(Y_i, Y_j | x_i) = Cov(Cov(Y_i, Y_j | x_i)) = \frac{dh}{d\theta_i} \frac{dh}{d\theta_j} \Sigma_{ij} \tau_i \tau_j
\]

where \( \Sigma_{ij} \) is the cross-covariance of prediction for the \( i \)th and \( j \)th physical covariates and \( \{X_i\} \) are the observed physical covariates.

Simulation studies

To assess the size of relative bias a simulation study was performed. We simulated biological data given the observed data and analysed them using the spatially predicted covariates \( \tilde{x}_i \) and the observed covariates \( x_i \). A total of 1000 simulations were performed and are indexed by \( k \) in the equations below.

Simulation model 1

\[
E[Y_{i(k)}] = h(\tilde{x}_i^{g(\tilde{x}_i, \tau)})
\]

Simulation model 2

\[
E[Y_{i(k)}] = h(x_i^{g(x_i, \tau)})
\]

Results

<table>
<thead>
<tr>
<th>Relative size of bias</th>
<th>Min.</th>
<th>Median</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Presence/absence</td>
<td>-19.0%</td>
<td>100.1%</td>
<td>839400.0%</td>
</tr>
<tr>
<td>Richness</td>
<td>1.3%</td>
<td>3.5%</td>
<td>17.2%</td>
</tr>
<tr>
<td>Summary statistics of the 200 sites</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Conclusions

> Using spatially modelled covariates leads to bias in outcome distribution and parameter estimates

> Simulation shows this is a real problem not just a theoretical one

> Will need to account for uncertainty in future models

Acknowledgements

We would like to thank everyone who contributed to the collection of the Great Barrier Reef data. This study was supported by the Climate Research Facilities Program (CRF), administered through the Australian Government’s Department of the Environment, Water, Heritage and the Arts. Further information can be obtained from: www.marinehub.org

References


Further information

www.marinehub.org