Uncertainty in Spatially Predicted Covariates: Is It Ignorable?

Scott D. Foster†

CSIRO Mathematics, Informatics and Statistics, and CSIRO's Wealth from Oceans Flagship, Hobart, Australia.

Hideyasu Shimadzu

Geoscience Australia, Environmental Geoscience Division, Canberra, Australia.

Ross Darnell

CSIRO Mathematics, Informatics and Statistics, Brisbane, Australia.

Summary. In ecology, a common form of statistical analysis relates a biological variable to variables that delineate the physical environment, typically by fitting a regression model or one of its extensions. Unfortunately, the biological data and the physical data are frequently obtained from separate data sources. In such cases there is no guarantee that the biological and physical data are co-located and the regression model cannot be used. A common and pragmatic solution is to predict the physical variables at the locations of the biological variables and then use the predictions as if they were observations. In this article, we show that this procedure can cause potentially misleading inferences and use generalised linear models as an example. We propose a Berkson-error model which overcomes the limitations. The differences between using predicted covariates and the Berkson error model are illustrated using data from the marine environment, and a simulation study based on these data.

1. Introduction

One important goal of ecological research is to understand how ecological quantities are related to the physical environment in which they reside. Common variations on this theme include: species distribution modelling (e.g. Lehmann et al., 2002; Guisan et al., 2002, 2006, all are editorials to special editions devoted to the topic), realised-niche delineation (e.g. Chase and Leibold, 2003), community prediction/mapping (e.g. Leathwick et al., 2005; Ferrier and Guisan, 2006), and biodiversity modelling (e.g. Foster and Dunstan, 2010). The analytical methods used to perform these tasks range from simple to complex. The complexity depends on the ecological quantity under consideration, the assumed functional relationships, the assumed statistical model, and the number of variables used to delineate the environment. They all have certain aspects in common: they relate biological data to physical using a regression model or one of its extensions. Focus is typically on prediction of the ecological quantity at spatial locations that are not sampled. We will do the same in this paper.

Commonly, the physical variables used for delineation are not directly measured at the locations of the ecological data as they often come from separate data sources. To facilitate

†Address for correspondence: CSIRO Mathematics, Informatics and Statistics, GPO box 1538, Hobart 7001, Tasmania, Australia. E-mail: scott.foster@csiro.au. Telephone: +61 (3) 6232 5178. Facsimile: +61 (3) 6232 5000.

the analysis, the physical variables are often predicted at the locations of the biological data from the locations of the physical data using geostatistical methods. It is important to note that the predictions will not be as variable as the actual observations would be. This two-stage analysis is used in preference to forming a joint geostatistical model for the biological *and* physical variables as the raw physical data are rarely made available to the analysts of the biological data, due to inter-organisational data-ownership issues. However, data products, such as predictions, are commonly made available. Examples from the marine environment are the National Oceanographic Data Center's World Ocean Atlas, and CSIRO's Atlas of Regional Seas.

It is not immediately clear what effect ignoring this extra level of variation will have on the validity of the ecological models. The purpose of this paper is to demonstrate how the extra variability can be included in the statistical model by specifying a Berkson-error model (see Carrol et al., 2006) and to compare it to the commonly performed analysis. We use the popular class of generalised linear models (GLMs; McCullagh and Nelder, 1989) to demonstrate the effect of the often-ignored Berkson errors. GLMs are analytically tractable and often form the basis for more complicated models.

The approach taken here differs from previous work that considers measurement error in ecological models (Elston et al., 1997; Van Niel and Austin, 2007). Those studies all consider extra uncertainty in the physical observations arising from observations with imprecise measuring equipment – a classical measurement error model (see Carrol et al., 2006). In this paper we add to this discussion by considering a previously ignored source of uncertainty, that associated with not observing the covariate data directly. This source of uncertainty is extremely prevalent in ecological modelling and cannot be captured by a classical measurement error model. The ecological problem under consideration here is similar to the misalignment problem in environmental epidemiology studies, see Lopiano et al. (2011) for a recent review and comparison. However, in those studies the statistical model is linear, which simplifies the problem substantially. The ecological problem cannot be tackled directly by these methods and many of the results are not applicable to non-linear models.

In the remainder of this article we describe some example data from the marine environment (Section 2) and outline models to analyse such data (Section 3). Analytical approximations to the bias induced from using predicted covariates are presented in Section 4 and the sizes of the biases for the example data are given in Section 4.2. Section 5 describes a simulation study and the example data are analysed in Section 6. Section 7 provides a summary and discussion. The code and the synthetic data, described in Section 2.1, are available from the journal's website as an R-package called `SEIC`.

2. Great Barrier Reef Data

Data were collected during a survey of the Great Barrier Reef (GBR) lagoon off the north eastern coast of Australia (Pitcher et al., 2007). The purpose of data collection was to characterise biodiversity for conservation purposes. These data were chosen for use in this study as they are atypical in that they are thorough and extensive for both biological and physical data and in that the physical data were collected at the same locations as the biological data. This enables us to mimic situations where prediction of covariates is necessary through degradation of the covariate data.

There were 1189 sites sampled using benthic sleds (biological data) in conjunction with

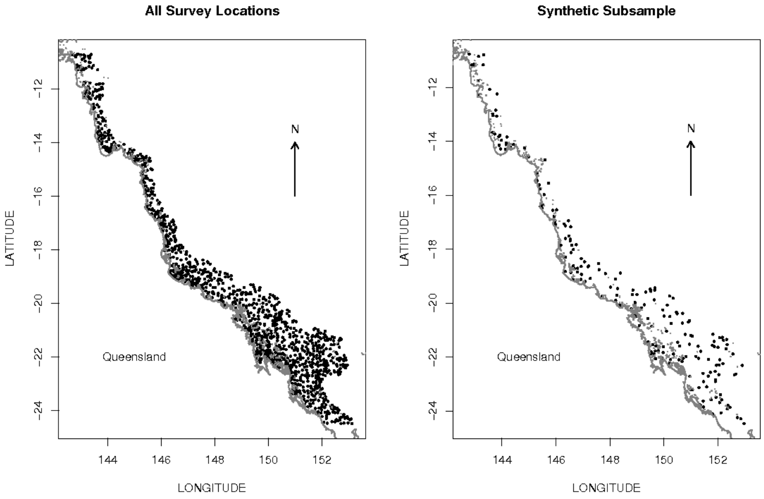

Fig. 1. Left panel: Survey locations within the Great Barrier Reef lagoon off Queensland, Australia. Right panel: A sub-sample of 200 locations taken completely at random.

sediment samples and sonar measurements (physical data); see Figure 1 for locations. The biological data considered here are the presence/absence of a Bryozoan species *Cheilostomata hippaliosina* and species-richness (number of species per square metre). The physical data considered in this paper were depth, %carbonates and %mud, and were chosen due to their high association with the two biological variables under consideration. The %mud and the %carbonates variables are not subject to any sum constraint. We transform the physical variables using the arcsine square-root transform (percentage variables) or the log transform (depth) but we continue to use the original variable names. The transformations are performed so that they vary over the entire real numbers, which simplifies the geostatistical modelling. For the purposes of this study the scale of the covariates is immaterial.

2.1. Degraded Data

We synthesise a survey by randomly sampling a set of 200 locations from the 1189 locations of the GBR data (see Figure 1 for locations). The synthetic survey is considered to be for biological data, while the locations in the original data (those that are not included in the synthetic survey) are considered to be the locations where physical data are measured. This mimics the data collection scenario under consideration. The physical variables at the locations of the synthetic samples are predicted based on the physical variables at the other locations in the GBR data. Prediction was performed using model-based kriging (Diggle and Ribeiro, 2007) with covariance structure defined by a normal convolution process (Higdon, 2002; Ver Hoef et al., 2004). Prediction covariances are also calculated. All cross-covariances are assumed to be zero, but we note that this is not a requirement of the methods used in this manuscript. Details of the prediction process are given in the Appendix.

3. Statistical Models for Macroecology and Biogeography

A common form for the statistical analysis that relates an ecological quantity to physical variables is that of a regression analysis and its many extensions. The regression model relates the expectation of the ecological quantity measured at n locations, $\boldsymbol{y}$ say, as a function of p physical covariates (arranged into the $n \times p$ matrix $\boldsymbol{X}$ with i^{th} row $\boldsymbol{x}_i^\top$). A completely general formulation for the mean model is

$$\mathrm{E}\left(y_i|\boldsymbol{x}_i\right) = h(\boldsymbol{x}_i), \tag{1}$$

where the function $h(\boldsymbol{x}_i)$ maps the i^{th} location's covariates, $\boldsymbol{x}_i$, to a scalar. It is common to assume that, conditional on the covariates, the observations $\boldsymbol{y}$ are independently distributed. In general, we believe this to be a reasonable assumption as the range of spatial dependence due to ecological sources is typically much smaller than the density of the samples. However, there are cases where this assumption will not hold, an example is animal movement data. Model (1) encompasses many that are used in the literature including generalised linear models (GLMs; see McCullagh and Nelder, 1989). The assumed probability density function (PDF) for model (1) is

$$f(\boldsymbol{y}|\boldsymbol{X}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{x}_i), \tag{2}$$

where $f(y_i|\boldsymbol{x}_i)$ is the PDF for location i.

Model (1) assumes that the p covariates are observed at the locations where the ecological variables are observed. Frequently, this assumption does not match reality and so the matrix $\boldsymbol{X}$ is replaced by predictions from a geostatistical model. This transforms (1) to

$$\mathrm{E}\left(y_i|\tilde{\boldsymbol{x}}_i\right) = h(\tilde{\boldsymbol{x}}_i), \tag{3}$$

where $\tilde{\boldsymbol{x}}_i$ is the vector of predicted covariates at location i.

An appropriate statistical model must account for the structure of the observations and, in particular, for the fact that the observed ecological and physical variables are not co-located. A natural way to do this is by considering the conditional distribution of the ecological variables given the observed physical variables. With all the observed physical variables given in the design matrix $\boldsymbol{X}_o$ the conditional PDF is

$$\begin{aligned} f(\boldsymbol{y}|\boldsymbol{X}_o) &= \int f(\boldsymbol{y}, \boldsymbol{X}|\boldsymbol{X}_o) d\boldsymbol{X} \\ &= \int f(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{X}_o) f(\boldsymbol{X}|\boldsymbol{X}_o) d\boldsymbol{X} \\ &= \int f(\boldsymbol{y}|\boldsymbol{X}) f(\boldsymbol{X}|\boldsymbol{X}_o) d\boldsymbol{X} \\ &= \int \left(\prod_{i=1}^{n} f(y_i|\boldsymbol{x}_i) \right) f(\boldsymbol{X}|\boldsymbol{X}_o) d\boldsymbol{X}, \end{aligned} \tag{4}$$

where $f(\boldsymbol{X}|\boldsymbol{X}_o)$ is the predictive distribution defined from a geostatistical model for the observations $\boldsymbol{X}_o$. This derivation assumes that an ecological observation depends only on the unobserved physical covariates at their locations. The PDF (4) is that for a Berkson-error model, a model whose covariate uncertainty stems from observing a smooth version of the actual variable (see Carrol et al., 2006, Chapter 1).

There are direct parallels between this approach and a Bayesian approach. In particular, there is now a set of random effects in the model (the Berkson errors), and these have to be integrated over to make inference about the GLM's parameters. The important difference between the model in (4) and a full Bayesian model is that the GLM's parameters are treated as fixed. Hence it is a direct extension of the usual GLM model, one that is more in keeping with the philosophy of the GLM than the Bayesian solution.

For likelihood inference the integral in (4) would have to be evaluated and subsequently maximised. This is only analytically possible here for normal GLMs, since the geostatistical model used assumes normality too (see the Appendix). However, in the next Section we show that approximations to the first two moments of this distribution are available for non-normal data. This enables assessment of the size of the bias terms that arise from using plug-in values for the unobserved physical variables.

3.1. Estimation Methods for Berkson-Error Models

There have been many methods proposed for finding the maximum-likelihood estimates from marginal PDFs, like that in (4). Special care needs to be used here as the dimension of the integrating variable is large with respect to the number of observations. The method that we employ, Laplace importance sampling (LIS: Kuk, 1999; Skaug and Fournier, 2006), is quite effective but it is by no means the only solution. A notable alternative is MCMC. However, given the dimensionality of the integral we do not expect that *any* integration method will provide a computationally inexpensive solution. Before settling on LIS we experimented using MC maximum likelihood (MCML; see Diggle and Ribeiro, 2007, Section 5.5.1) and Laplace's approximation. We found that the MC error in MCML was unacceptably large. Laplace's approximation generally performed well but it occasionally gave extreme estimates. In these cases the LIS behaved more predictably. We expect that MCMC, like LIS, would produce good results but it remains untested here.

The LIS method is an extension of the well-known Laplace approximation where the integrand in (4) is approximated by a multivariate normal with mean specified by the maximum of the integrand and variance by the Hessian of the integrand calculated at the location of the maximum. The Laplace approximation works well for many problems but not for all; the accuracy depends on the accuracy of the underlying normal approximation. The LIS method alters the Laplace approximation by using the integrand's normal approximation as a proposal distribution for integration by importance sampling (Kuk, 1999). Continuing from (4), the approximate log-likelihood for the GLM's mean parameters and its dispersion parameter, $\boldsymbol{\tau}$ and ϕ respectively, is

$$\ell^{[k]}(\boldsymbol{\tau}^{[k]}, \phi^{[k]}; \boldsymbol{y}) = \frac{1}{B} \sum_{b=1}^{B} f^{[k]}(\boldsymbol{y}|\boldsymbol{x}_b) \frac{f^{[k]}(\boldsymbol{x}_b)}{g^{[k]}(\boldsymbol{x}_b)}, \tag{5}$$

where $\boldsymbol{x}_b$ is a draw from the normal proposal distribution, $g^{[k]}(\boldsymbol{x}_b)$. The k-superscript on the densities in (5) is a consequence of iterative updating of the mean and dispersion parameters of the integrand.

The LIS approximation, like the Laplace approximation, requires two levels of optimisation. First, the maximum of the integrand must be obtained with respect to the random effects. The integrand has the same form as a penalised likelihood and we use Newton-Raphson for its optimisation. Once the normal approximation has been found, the marginal log-likelihood is obtained via importance sampling. For the optimisation of the marginal log-likelihood with respect to the model's parameters, we use derivative-free optimisation (Nelder-Mead simplex). All optimisation was done using the function `nlminb` in R. Iteration over these two optimisation steps is required. We consider the process converged when the difference in successive iterations' log-likelihood is less than $1e-5$.

Following Skaug and Fournier (2006) we choose a single set of standard normal random variates $\boldsymbol{x}_*$ that are used in all iterations. The importance samples for iteration b are given by the multivariate transformation

$$\boldsymbol{x}_b = \boldsymbol{Q}\boldsymbol{x}_* + \tilde{\boldsymbol{x}},$$

where $\tilde{\boldsymbol{x}}$ is the vector of predicted covariate values, $\boldsymbol{Q}$ is the symmetric square root of $\left(-\boldsymbol{H}^{[k]}\right)^{-1}$ where $\boldsymbol{H}^{[k]}$ is the Hessian of the joint PDF of observations and unobserved covariates (evaluated at the maximum). See Harville (1997, page 543) for a description of the symmetric square root.

We suggest choosing a large number of samples to try and reduce the Monte Carlo error. The number of samples is primarily dependent on the number of Berkson errors but will also be dependent on other attributes of the data. For the data analysis in Section 6 we use $B = 50,000$, which appears sufficient. It is advisable to repeat the estimation process, with a new set of random samples, to diagnose estimation abnormalities. Any differences in parameter estimates between estimation runs could be attributable to two factors. The first, which is likely to explain the majority of any differences, is due to varying the locations of the Laplace importance samples. The second is due to the numerical optimisation routine used for estimation and, hopefully, should not affect the estimates substantially.

An estimate of the variance-covariance matrix of $\hat{\boldsymbol{\tau}}$ and $\hat{\phi}$ is available through the negative of the inverse of the Hessian of the log-likelihood. We use a five-point finite-difference approximation to calculate the Hessian using the highly accurate method described in Fornberg and Sloan (1994, Table 1). The Hessian is not guaranteed to be positive definite and any singularities are likely to be due to optimisation problems rather than problems with the finite difference approximation. If singularities are encountered, our advice is to try increasing the number of Laplace importance samples, this provides a better estimate of the log-likelihood function.

4. Approximate Moments and Bias Terms

When specifying a statistical model, one of the critical choices is the form of stochastic variation. It needs to reflect accurately the unexplained variation of the observations around the modelled mean. If it is incorrectly specified then any inference and prediction from the model may be misleading as the score equations will be biased. With this as motivation, we now study the distribution of the observations, given that there are Berkson errors in the covariates. The results are compared to the case where the Berkson errors are ignored.

We assume that the model for the ecological variable, conditional on the unobserved physical variables, follows a GLM. This assumption is made for illustrative convenience but we note that the results carry over, with slight extension, to any model that is a smooth function of the covariates. The GLM assumption specifies

$$\mathrm{E}\left(y_i|\boldsymbol{x}_i\right) = h(\boldsymbol{x}_i^\top\boldsymbol{\tau}),$$

where $h(\cdot)$ is the inverse link function, $\boldsymbol{\tau}$ is a $p \times 1$ vector of unknown parameters, $\boldsymbol{x}_i$ is the i^{th} row of $\boldsymbol{X}$, and the observations, $\boldsymbol{y}$, are independently distributed. We assume that the elements of the first and second moments of the distribution of the unobserved covariate $\boldsymbol{z}_j$, conditional on its observed counterpart (j^{th} column of $\boldsymbol{X}_o$), are

$$\mathrm{E}\left(\boldsymbol{z}_j|\boldsymbol{X}_o\right) = \boldsymbol{\mu}_j \text{ and } \mathrm{var}\left(\boldsymbol{z}_j|\boldsymbol{X}_o\right) = \boldsymbol{\Sigma}_j.$$

In practice the quantities $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ are unknown and must be estimated. We perform estimation via univariate geostatistical modelling, described briefly in Section 2.1 and in more detail in the Appendix.

Let $\tilde{\boldsymbol{x}}_i$ and $\boldsymbol{V}_i$ be the conditional expectation and variance of $\boldsymbol{x}_i$. Note that $\tilde{\boldsymbol{x}}_i$ is a vector that consists of the i^{th} row of each of the individual covariates' expectations ($\boldsymbol{\mu}_j$), and the variances $\boldsymbol{\Sigma}_j$ and $\boldsymbol{V}_i$ will share corresponding elements. The conditional expectation of the biological outcome y_i is

$$\begin{aligned}
\mathrm{E}\left(y_i|\boldsymbol{X}_o\right) &= \mathrm{E}\Big(\mathrm{E}\left(y_i|\boldsymbol{X}\right)|\,\boldsymbol{X}_o\Big) \\
&\approx \mathrm{E}\left(h(\tilde{\boldsymbol{x}}_i^\top\boldsymbol{\tau})\middle|\boldsymbol{X}_o\right) + \mathrm{E}\left(\frac{1}{2}h''(\tilde{\boldsymbol{x}}_i^\top\boldsymbol{\tau})(\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i)^\top\left(\boldsymbol{\tau}\boldsymbol{\tau}^\top\right)(\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i)\middle|\boldsymbol{X}_o\right) \\
&= h(\tilde{\boldsymbol{x}}_i^\top\boldsymbol{\tau}) + \frac{1}{2}h''(\tilde{\boldsymbol{x}}_i^\top\boldsymbol{\tau})\boldsymbol{\tau}^\top\boldsymbol{V}_i\boldsymbol{\tau}, \qquad (6)
\end{aligned}$$

where $h'(s)$ and $h''(s)$ are the first and second derivatives of $h(s)$ with respect to s. Using the predicted covariates as plug-in values will give biased expected values for the ecological model unless one of two conditions are met:

- the inverse link function is linear, or
- the quadratic form ($\boldsymbol{\tau}^\top\boldsymbol{V}_i\boldsymbol{\tau} = \boldsymbol{0}$) is zero, which can occur when $\boldsymbol{\tau} = \boldsymbol{0}$ or when $\boldsymbol{V}_i = \boldsymbol{0}$. This occurs when the ecological variables are not related to the physical variable or when the physical variables are measured with certainty.

None of these conditions can be guaranteed as they imply: 1) a potentially inappropriate model (linear link); 2) no relationship of biological data with physical data; or 3) prediction of covariates with zero prediction variance. The last condition may hold approximately if the physical data's density is sufficiently high. The remaining question is not about the presence of bias, rather it is about the size of the bias.

Let $v_*\left(\boldsymbol{x}_i^\top\boldsymbol{\tau}\right) = v\left(h(\boldsymbol{x}_i^\top\boldsymbol{\tau}), \phi\right)$ be the GLM's variance function expressed as a function of the linear predictor, where the function $v(\cdot, \phi)$ is the standard GLM variance function (see McCullagh and Nelder, 1989). The conditional variances and covariances are

$$\begin{aligned}
\mathrm{var}\left(y_i|\boldsymbol{X}_o\right) &= \mathrm{E}\Big(\mathrm{var}\left(y_i|\boldsymbol{x}_i\right)|\boldsymbol{X}_o\Big) + \mathrm{var}\Big(\mathrm{E}\left(y_i|\boldsymbol{x}_i\right)|\boldsymbol{X}_o\Big) \\
&\approx \mathrm{E}\Big(v_*\left(\tilde{\boldsymbol{x}}_i^\top\boldsymbol{\tau}\right)|\boldsymbol{X}_o\Big) + \mathrm{var}\Big(h'(\tilde{\boldsymbol{x}}_i^\top\boldsymbol{\tau})(\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i)^\top\boldsymbol{\tau}|\boldsymbol{X}_o\Big) \\
&= v_*\left(\tilde{\boldsymbol{x}}_i^\top\boldsymbol{\tau}\right) + \left[h'(\tilde{\boldsymbol{x}}_i^\top\boldsymbol{\tau})\right]^2\boldsymbol{\tau}^\top\boldsymbol{V}_i\boldsymbol{\tau}, \qquad \text{and} \qquad (7) \\
\mathrm{cov}\left(y_i, y_{i'}|\boldsymbol{X}_o\right) &\approx 0 + h'(\tilde{\boldsymbol{x}}_i^\top\boldsymbol{\tau})h'(\tilde{\boldsymbol{x}}_{i'}^\top\boldsymbol{\tau})\boldsymbol{\tau}^\top\boldsymbol{V}_{ii'}\boldsymbol{\tau}, \qquad (8)
\end{aligned}$$

where $\boldsymbol{V}_{ii'}$ is the matrix of covariances and cross-covariances for locations i and i'. In both (7) and (8) the first term is that expected from the GLM using predicted covariates and the second term can be seen as a bias adjustment. Analogous to the bias terms for the expectation, these can be zero only when there is no relationship between ecological and physical variables, or when there is no variation in the prediction for the physical variable.

Failure to account for the variability of the predicted covariates can affect the assumed distribution of the biological outcomes in three ways. First, the assumed expectations do not represent the actual expectations and bias may be present. Second, the variance is inflated in comparison to the assumed variance giving a false level of confidence. Third, the covariances are non-zero implying an over-estimate of the effective degrees of freedom available. The second and third points together imply that any test of significance is likely to have unduly high power for hypothesis testing.

4.1. Accuracy of Approximations

The approximations (6), (7) and (8) arise from a low-order Taylor series expansion around the geostatistical prediction of the covariates. Taylor series approximations are commonly used as they provide a way to approximate non-linear functions by an easy to manipulate polynomial. Higher order terms will make the approximation more accurate but should have diminishing contribution, especially for well-behaved functions like those used for link functions in GLMs. The higher order terms, as a collective, should not have a positive or negative net effect on the bias in the expectation, variance and covariance. Even if there was a net effect, it should be less than the magnitude of the terms already considered.

The accuracy of the approximation will depend on how close the actual, unobserved values of the covariate are to their geostatistical predictions. If the predictions are close then the approximations should suffice. However, if there are large discrepancies then the approximation, and the resulting bias terms, may be questionable.

There were two main reasons for not considering higher order approximations. The first is that the simplicity of the low-order approximations provides straight-forward qualitative interpretation of the nature of the bias, even if the exact numerical values are inaccurate in certain situations. The second is that the higher order terms are very difficult to derive and compute in a multivariate function as they are based on expectations and variances of quadratic forms (and higher order products) of multivariate variables.

4.2. Size of Bias for GBR Data

We investigate the size of the bias terms by considering the synthetic survey described in Section 2. Two situations are considered: a Bernoulli GLM for the presence/absence of the species of the Bryozoan *Cheilostomata hippaliosina*, and a Poisson GLM for species-richness. The GLM's parameter values were obtained from an analysis of the observed data on the observed covariates. This analysis cannot typically be performed as the physical data are not directly measured. We use these estimates for simulation as they are likely to be a good representation of the true parameter values.

The relative size of the bias terms for expectation and variance are calculated for each location as the bias adjustment term divided by the expected moment if the GLM with predicted covariates is used (see equations (6) and (7)). The biases in covariances are summarised as correlations. A summary of the relative biases is given in Table 1.

Table 1. Summary of relative bias taken from (6) and (7), and correlations from (8) for Bernoulli and Poisson GLMs. Relative bias is defined to be the bias adjustment terms in (6) and (7) divided by that expected from a GLM using predicted covariates. The correlation measure is amongst all observations and is based on the covariance (8) and variances (7). The correlation is not relative but the simple GLM assumes all to be zero.

		Min	1^{st} Quant.	Median	3^{rd} Quant.	Max
Bernoulli	Expectation	–0.077	0.085	0.245	0.402	123.300
	Variance	0.000	0.017	0.079	0.195	2.858
	Correlation	–0.071	0.000	0.000	0.000	0.407
Poisson	Expectation	0.014	0.031	0.036	0.042	0.162
	Variance	0.812	2.569	3.816	6.223	14.990
	Correlation	–0.064	0.000	0.000	0.000	0.394

For the Bernoulli GLM the bias in the expectation is substantial and sometimes extreme. The bias in the variance for the Bernoulli GLM can be large, although it is generally moderate. For the Poisson GLM, the bias in the variance is large but not the bias in the expectation. In both models the correlation induced from ignoring the variance in covariate predictions is typically small, but occasionally it is quite large and positive.

5. Bias in Parameter Estimates – Simulation Study

The performance of the approach was investigated using two simulation studies, both based on the synthetic GBR lagoon data (Section 2). The first simulation study is based on the presence/absence of the Bryozoan *Cheilostomata hippaliosina*, and the second simulation is based on species-richness. These are modelled by a Bernoulli and a Poisson GLM respectively, with parameter values given in Tables 2 and 3. Each simulated data set is created by first simulating covariates as a realisation of the spatial process estimated from the geostatistical models and then simulating biological data conditional on the simulated physical covariates. We fitted three models to every simulated data set: 1) a GLM of the biological data on the predicted covariates, 2) a Berkson GLM using LIS, and 3) a GLM of biological data on the actual simulated covariates. We label these models as 'pred-GLM', 'Berk-GLM' and 'true-GLM' respectively. The third model cannot be fitted to real data as the observed covariates are not available. We include this model here to show the behaviour of the model that the analyst would like to fit, if he/she could.

The 1000 sets of estimated parameters and their theoretical standard errors are summarised in Tables 2 and 3 for the Bernoulli and Poisson models respectively. We inspect the mean of the estimates against the values used for simulating, and compare the empirical standard deviation of the estimates against the mean of the theoretical standard errors. These comparisons will indicate if there is bias in an estimate's mean and standard error attributable to ignoring prediction variance.

The Bernoulli simulation indicates that all three methods produce biased means. This is surprising as we would typically expect that the true-GLM would be unbiased. However, for any simulated data set, the realised physical covariates will be correlated through space, which will cause information loss and increase the chance of extreme estimates. The mean estimates from the pred-GLM were always lower than the means from the other two methods, which were quite close to each other for all parameters. The estimates' empirical standard deviation and the mean theoretical standard error matched reasonably well but were slightly larger for pred-GLM and true-GLM. Also the standard errors for the Berk-GLM were larger than those for the other models, reflecting the allowance for uncertainty in the covariates. These results follow from the approximate distribution derived in Section 4.2 (Table 1). There is substantial bias in expectation and negligible bias in the variance.

The Poisson simulation gave different, but complementary, results from the Bernoulli simulation. The parameter estimates from all methods were essentially unbiased for the mean, except for the intercept estimate from the pred-GLM, Table 3. This estimate is slightly inflated, which corresponds to a slight and relatively constant bias in the distribution of the observations (see Table 1). The empirical standard deviation and the theoretical standard error from the Berk-GLM and the true-GLM agreed well, indicating that all sources of variation were accounted for (see Table 3). However, these two statistics do not agree for the pred-GLM; the theoretical value is only about half the empirical value. This indicates that an analyst, using predicted covariates in a GLM, will obtain estimates that have unrealistically high confidence.

The reason for the differences in the results for the Bernoulli and Poisson simulations can be explained by considering the approximations in Section 4. The behaviour appears to be dictated by the link function of the different models. Consider the Poisson case first (log link). The bias in an observation's expectation is proportional to $\exp(\eta)$, where η is the linear predictor, and the bias in an observation's variance is proportional to $\exp(2\eta)$. The bias in variance is larger and it follows through to the estimates. Now consider the Bernoulli case with a logit link. The bias terms are proportional to $\pi(1-\pi)(1-2\pi)$ and $\pi^2(1-\pi)^2$ for the mean and variance bias respectively, with π the fitted value. Bias in expectation is zero if $\pi = 0.5$. So, if the set of fitted values has mean near zero then this term should average out. This is the situation in this simulation but it will not always be the case.

Combined, the two simulations suggest that bias in parameter estimates *and* their standard errors can be biased if predicted covariates are used as a direct substitute for the unobserved measurements. The presence of bias in the estimates depends on the conditions mentioned in Section 4. The size of bias appears to depend on the sign and size of the derivatives of the link function, the amount of variation in the covariates and the size of the parameter values.

Table 2. Summary of parameter estimates for the Bernoulli simulation study (1000 simulations). The second column contains the values used to generate the individual data sets. The three models types are: 1) a GLM using predicted covariates (pred-GLM), 2) a Berkson-error GLM (Berk-GLM), and 3) a GLM with each simulation's realised physical covariates (true-GLM).

Covariate	Value	Model	Mean($\pm$SE*)	Estimate SD†	Mean SE‡
Intercept	–2.550	pred-GLM	–2.342(0.011)	0.359	0.347
		Berk-GLM	–2.691(0.017)	0.538	0.539
		true-GLM	–2.669(0.013)	0.426	0.396
Depth	0.266	pred-GLM	0.244(0.011)	0.337	0.326
		Berk-GLM	0.278(0.012)	0.389	0.386
		true-GLM	0.283(0.010)	0.314	0.303
%Carbonates	0.704	pred-GLM	0.650(0.009)	0.300	0.287
		Berk-GLM	0.763(0.011)	0.360	0.352
		true-GLM	0.744(0.009)	0.278	0.272
%Mud	–1.841	pred-GLM	–1.691(0.013)	0.396	0.392
		Berk-GLM	–1.916(0.017)	0.549	0.570
		true-GLM	–1.923(0.012)	0.391	0.376

*Standard error of the mean of all simulated data sets' estimates

†Standard deviation of all simulated data sets' estimates

‡Mean asymptotic standard error (average of standard errors from each data set)

Table 3. Summary of parameter estimates for the Poisson simulation study (1000 simulations). The second column contains the values used to generate the individual data sets. The three models types are: 1) a GLM using predicted covariates (pred-GLM), 2) a Berkson error GLM (Berk-GLM), and 3) a GLM with each simulation's realised physical covariates (true-GLM).

Covariate	Value	Model	Mean($\pm$SE*)	Estimate SD†	Mean SE‡
Intercept	–1.742	pred-GLM	–1.706($\leq$0.001)	0.023	0.010
		Berk-GLM	–1.741($\leq$0.001)	0.022	0.023
		true-GLM	–1.742($\leq$0.001)	0.010	0.010
Depth	–0.071	pred-GLM	–0.070($\leq$0.001)	0.034	0.013
		Berk-GLM	–0.070($\leq$0.001)	0.030	0.030
		true-GLM	–0.070($\leq$0.001)	0.011	0.011
%Carbonates	0.323	pred-GLM	0.325($\leq$0.001)	0.036	0.013
		Berk-GLM	0.323($\leq$0.001)	0.033	0.031
		true-GLM	0.323($\leq$0.001)	0.010	0.011
%Mud	–0.487	pred-GLM	–0.486($\leq$0.001)	0.029	0.012
		Berk-GLM	–0.484($\leq$0.001)	0.023	0.024
		true-GLM	–0.487($\leq$0.001)	0.010	0.010

*Standard error of the mean of all simulated data sets' estimates

†Standard deviation of all simulated data sets' estimates

‡Mean asymptotic standard error (average of standard errors from each data set)

Table 4. GBR data: Parameter estimates and standard errors for a GLM using the predicted covariates (pred-GLM) and a GLM with Berkson errors (Berk-GLM). Two GLMs were considered, a Bernoulli GLM for presence/absence data and a Poisson GLM for species-richness data.

		Bernoulli			Poisson		
Covariate	Model	Estimate[*]	SE[*]	Esti-SD[†]	Estimate[*]	SE[*]	Esti-SD[†]
Intercept	pred-GLM	–2.536	0.376	–	–1.717	0.010	–
	Berk-GLM	–3.102	0.657	0.004	–1.971	0.029	0.002
Depth	pred-GLM	–0.230	0.332	–	–0.026	0.013	–
	Berk-GLM	–0.270	0.412	0.002	–0.409	0.049	0.011
%Carbonates	pred-GLM	1.053	0.326	–	0.301	0.013	–
	Berk-GLM	1.267	0.438	0.002	0.766	0.013	0.013
%Mud	pred-GLM	–1.769	0.432	–	–0.435	0.012	–
	Berk-GLM	–2.192	0.676	0.005	–0.926	0.049	0.007

[*] Taken from the first run of estimation algorithm

[†] Standard deviation of 20 runs of estimation algorithm

6. Analysis of the Great Barrier Reef Data

The results from the simulation study in Section 5 show that the GLM with Berkson-errors is a more realistic representation of the variation in the data. In this section we show how different the estimates from the two models can be for the synthetic data from the GBR lagoon (see Section 2). We fit the two models (predicted covariates GLM and Berkson error GLM) for both the presence/absence of the Bryozoan *Cheilostomata hippaliosina* and for species-richness. The resulting estimates are given in Table 4.

The two sets of estimates for the Bernoulli GLM are different but their standard errors are large with respect to the differences (Table 4). It is impossible to say from this single analysis if there is bias in the estimates. However, the estimates and their standard errors from the Berkson GLM are always larger than those from the GLM using predicted covariates. This is consistent with the simulation study in Section 5, even though the conclusions are not as compelling.

The two sets of estimates for the Poisson GLM are also quite different, as are their standard errors. The increase in standard error (except for %carbonates) is consistent with the simulation study, although the size of the increase is substantially larger for the real data. We note that the results presented in Table 3 are averages and there is no guarantee that each realisation will match, as is the case in Table 4. However, the differences cannot be totally attributable to this effect – they are too large with respect to the simulation study. The proportional differences in the standard errors in the simulation study were almost never observed to be as large as those observed in Table 4, except for %carbonates which was never as small. We can only speculate why this has occurred but it is almost certainly due to some unknown and unmodelled attribute of the data.

The variance in the parameter estimates was investigated by performing the estimation multiple (20) times. This helps diagnose the combined effect of Monte Carlo error in the LIS routine and estimation differences. The standard deviation between the 20 runs is given in Table 4. These values are small in comparison to the size of the estimate. For the purposes of this paper, this amount of estimation error is satisfactory.

Quantitative ecologists will typically use the estimated model to produce predictions on a dense grid of locations throughout the study region. These predictions are often displayed as a map. We now perform this prediction procedure for the predicted covariates and Berkson GLMs. In both cases we find the prediction and its standard error by simulation: we generate 1000 random draws from the asymptotic distribution of the estimates and use these to form a set of predictions at each location. We then compute the mean and standard deviation of the set of predictions for each location. For the predicted covariates GLM the predictions are obtained in the usual way by using the inverse link function on the randomly generated linear predictor. For the Berkson GLM the predictions are obtained using the expectation in (6). Different prediction methods are used as a reflection of the fact that the different models make different assumptions about the nature of the covariates. The predictions and their standard errors are compared in Figure 2.

The point predictions from the two models for the Bernoulli data are similar but the standard errors for the GLM using predicted covariates are smaller (Figure 2). That is, the point predictions from the predicted covariates GLM appear approximately correct but there is too much confidence placed in them. The point predictions from the Poisson model exhibit a substantial amount of bias, as do their standard errors (Figure 2). The predictions from the predicted covariates GLM can be substantially less than those from the Berkson GLM and the respective standard errors are nearly always substantially smaller. In both the Bernoulli and particularly the Poisson models there is potential for incorrect inference, which could lead to poor resource management decisions.

7. Summary and Discussion

In this paper we have identified and explored the effect of ignoring prediction variance when spatial predictions are used as covariates in a GLM. Our motivation was ecological modelling where the GLM relates biological data to spatially predicted physical data. We show that when the prediction variance is ignored the mean and variance of the sampling distribution of the biological data are biased. We expect that other, more complicated, types of models will produce similar results to those obtained for GLMs.

The bias in the sampling distribution transfers to bias in parameter estimates when the simple two-stage analysis is performed. Sometimes the bias manifests itself in the point estimates, sometimes in the estimates' standard errors, and sometimes both. We present a Berkson-error GLM for overcoming the bias in estimates and describe a method of estimation. A comparison of the estimates from the two models for the synthetic GBR

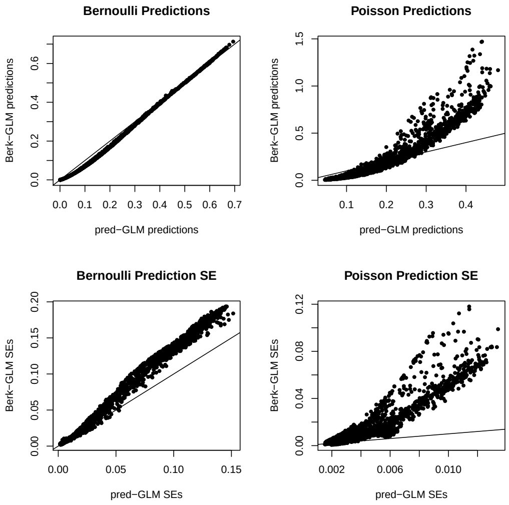

Fig. 2. A comparison of outcome predictions obtained from the predicted covariates GLM and the Berkson GLM. If the outcome predictions and their standard errors are the same then all points will lie on the 45° line. Left panels give predictions for the Bernoulli model and right panels are for the Poisson model. Upper panels are point predictions and lower panels are their standard errors.

data indicates that the bias is identifiable in real data, as well as simulated data.

The results presented in this paper can be used to provide an appropriate analysis of data that is not co-located. Further, they identify when a simple analysis could be performed, via a GLM with predicted covariates. The more complex analysis will add little rigour when the covariates are predicted with low variance, either through the observed covariates being located near the biological observations or from a spatially dense set of covariate observations. If there is substantial prediction variance in the covariates then the more complex analysis, described in Section 3.1, should be undertaken. Also, the results can guide planning of future surveys. In particular, analysis could be simplified by placing biological samples in areas of dense physical samples and/or close to the physical samples. Of course, ease of analysis is not the only consideration in survey design.

The dependence of the biological sampling distribution on the prediction mean and variance highlights that the geostatistical model is an important component of the modelling procedure. If the geostatistical models are inefficient then this will directly lead to inefficient models for the biological data too. If the predictions are provided by another researcher or organisation then a great amount of faith in those researchers and their practices is implicitly required.

Acknowledgements

Our sincerest thanks go to: Roland Pitcher, Zhi Huang, Kathy Haskard, Mark Palmer, Geoff Hosack, Bill Venables, Piers Dunstan, Jin Li, Camille Mellin, the Editor, the Associate editor and two referees. This work has been funded through the Commonwealth Environment Research Facilities (CERF) programme, an Australian Government initiative supporting world class, public good research. The CERF Marine Biodiversity Hub is a collaborative partnership between the University of Tasmania, CSIRO Wealth from Oceans Flagship, Geoscience Australia, Australian Institute of Marine Science and Museum Victoria. HS publishes in permission of the Chief Executive Officer of Geoscience Australia.

Appendix – Geostatistical Methods

We choose to formulate the task of predicting the physical variable in new locations as prediction of random site effects from a univariate geostatistical model (see Stein, 1999; Haskard, 2007; Diggle and Ribeiro, 2007). We use the usual mixed model process of 1) estimating the parameters of the geostatistical model (variances, and fixed effects), and 2) predicting the site random effects with plug-in values of the parameters.

Geostatistical prediction using mixed models requires the specification of a covariance structure (related to the classical variogram model); here we 'construct' a covariance structure rather than 'specifying' one (Higdon, 2002; Ver Hoef et al., 2004). Construction allows greater flexibility, and applicability, of an individual model through greater variability in the covariance structures permitted. The covariance structure is constructed using a moving average process over independent Gaussian effects specified on a predefined grid.

The mixed model for any one of the covariates (using simple kriging) is

$$\boldsymbol{z}_o = \boldsymbol{\alpha} + \boldsymbol{K}\boldsymbol{u} + \boldsymbol{e}$$

where $\boldsymbol{z}_o$ is the $n \times 1$ vector of observations for the covariate, $\boldsymbol{\alpha} = \alpha \mathbf{1}_n$ is the overall mean of the covariate, $\boldsymbol{K}$ is a $n \times k$ matrix for the spatial process, $\boldsymbol{u}$ is a $k \times 1$ vector of independent normal effects with zero mean and variance σ_o^2, and $\boldsymbol{e}$ is a vector of residuals with zero mean and variance σ^2. The matrix $\boldsymbol{K}$ is a moving average smoothing matrix that relates the locations of the observations to the locations of the spatial grid. The $(i,j)^{th}$ element of $\boldsymbol{K}$ is the kernel density of the i^{th} observation centred at the j^{th} grid point. In our implementation we choose a Gaussian kernel with both variances and covariance unknown. Hence, the matrix $\boldsymbol{K}$ is itself a function of unknown variance parameters which must be estimated along with the mean $\boldsymbol{\alpha}$, the spatial variance σ_o^2, and the residual variance σ^2.

We perform estimation using a profile (restricted-)likelihood approach. First, given the matrix $\boldsymbol{K}$, updates are made for the variance components and the mean. Second, given the variance components and the mean, the matrix $\boldsymbol{K}$ is updated. These two steps are iterated until convergence of parameters and log-restricted-likelihood.

The estimated variance components, fixed effects and $\boldsymbol{K}$ matrix are then used as plug-in values for formation of the spatial predictions. The predictions and prediction variances at a new set of m sites are given by

$$\begin{aligned} \hat{\mathrm{E}}\left(\boldsymbol{z}_m | \boldsymbol{z}_o\right) &= \hat{\mu}\mathbf{1}_m + \hat{\boldsymbol{\Sigma}}_{mo}\hat{\boldsymbol{\Sigma}}_{oo}^{-1}\left(\boldsymbol{x}_o - \hat{\mu}\mathbf{1}_n\right) \qquad \text{and} \\ \widehat{\mathrm{var}}\left(\boldsymbol{z}_m | \boldsymbol{z}_o\right) &= \hat{\boldsymbol{\Sigma}}_{mm} - \hat{\boldsymbol{\Sigma}}_{mo}\hat{\boldsymbol{\Sigma}}_{oo}^{-1}\hat{\boldsymbol{\Sigma}}_{om} \end{aligned}$$

where

$$\widehat{\mathrm{var}}\left(\begin{bmatrix} \boldsymbol{z}_m \\ \boldsymbol{z}_o \end{bmatrix}\right) = \begin{pmatrix} \hat{\boldsymbol{\Sigma}}_{mm} & \hat{\boldsymbol{\Sigma}}_{mo} \\ \hat{\boldsymbol{\Sigma}}_{om} & \hat{\boldsymbol{\Sigma}}_{oo} \end{pmatrix}.$$

The usual kriging expression can be obtained by substituting $\hat{\boldsymbol{\alpha}}$ with its generalised least squares estimator (see Haskard, 2007). The estimated joint-variance matrix is obtained from the specification of the linear mixed model. As an example consider the variance for the observed values of the covariate

$$\hat{\boldsymbol{\Sigma}}_{oo} = \hat{\sigma}_o^2 \hat{\boldsymbol{K}}\hat{\boldsymbol{K}}^\top + \hat{\sigma}^2 \boldsymbol{I}_n.$$

References

Carrol, R. J., D. Ruppert, L. A. Stefanski, and C. Crainiceanu (2006). *Measurement Error in Nonlinear Models: a Modern Perspective.* Chapman & Hall / CRC.

Chase, J. M. and M. A. Leibold (2003). *Ecological Niches: Linking Classical and Contemporary Aproaches.* Chicago: The University of Chicago Press.

Diggle, P. and P. Ribeiro (2007). *Model-based Geostatistics.* New York: Springer.

Elston, D., G. Jayasinghe, S. Buckland, D. MacMillan, and R. Aspinall (1997). Adapting regression equations to minimize mean squared error of predictions made using covariate data from a gis. *International Journal of Geographical Information Science 11*(3), 265–280.

Ferrier, S. and A. Guisan (2006). Spatial modelling of biodiversity at the community level. *Journal of Applied Ecology 43*, 393–404.

Fornberg, B. and D. Sloan (1994). A review of pseudospectral methods for solving partial differential equations. *Acta Numerica 3*, 203–267.

Foster, S. D. and P. K. Dunstan (2010). The analysis of biodiversity using rank abundance distributions. *Biometrics 66*, 186–195.

Guisan, A., T. C. Edwards Jr, and T. Hastie (2002). Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling 157*, 89–100.

Guisan, A., A. Lehmann, S. Ferrier, M. Austin, J. Overton, R. Aspinall, and T. Hastie (2006). Making better biogeographical predictions of species distributions. *Journal of Applied Ecology 43*, 386–392.

Harville, D. A. (1997). *Matrix Algebra From a Statistician's Perspective.* New York: Springer.

Haskard, K. A. (2007). *An anisotropic Matérn spatial covariance model: REML estimation and properties.* Ph. D. thesis, The University of Adelaide.

Higdon, D. (2002). *Quantitative Methods for Current Environmental Issues.* London: Springer-Verlag.

Kuk, A. (1999). Laplace importance sampling for generalized linear mixed models. *Journal of Statistical Computation and Simulation 63*(2), 143–158.

Leathwick, J. R., D. Rowe, J. Richardson, J. Elith, and T. Hastie (2005). Using multivariate adaptive regression splines to predict the distributions of new zealands freshwater diadromous fish. *Freshwater Biology 50*, 2034–2052.

Lehmann, A., J. Overton, and M. Austin (2002). Regression models for spatial prediction: their role for biodiversity and conservation. *International Journal of Biodiversity and Conservation 11*, 2085–2092.

Lopiano, K., L. Young, and C. Gotway (2011). A comparison of errors in variables methods for use in regression models with spatially misaligned data. *Statistical Methods in Medical Research 20*(1), 29–47.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (2nd ed.). Monographs on Statistics and Applied Probability. Boca Raton: CRC Press.

Pitcher, C. R., P. Doherty, P. Arnold, J. Hooper, N. Gribble, C. Bartlett, M. Browne, N. Campbell, T. Cannard, M. Cappo, G. Carini, S. Chalmers, S. Cheers, D. Chetwynd, A. Colefax, R. Coles, S. Cook, P. Davie, G. De'ath, D. Devereux, B. Done, T. Donovan, B. Ehrke, N. Ellis, G. Ericson, I. Fellegara, K. Forcey, M. Furey, D. Gledhill, N. Good, S. Gordon, M. Haywood, P. Hendriks, I. Jacobsen, J. Johnson, M. Jones, S. Kinninmoth, S. Kistle, P. Last, A. Leite, S. Marks, I. McLeod, S. Oczkowicz, M. Robinson, C. Rose, D. Seabright, J. Sheils, M. Sherlock, P. Skelton, D. Smith, G. Smith, P. Speare, M. Stowar, C. Strickland, C. Van der Geest, W. Venables, C. Walsh, T. Wassenberg, A. Welna, and G. Yearsley (2007). Seabed biodiversity on the continental shelf of the great barrier reef world heritage area. Technical report, AIMS/CSIRO/QM/QDPI CRC Reef Research Task Final Report. 320 pp.

Skaug, H. and D. Fournier (2006). Automatic approximation of the marginal likelihood in non-gaussian hierarchical models. *Journal of Computational Statistics and Data Analysis 51*, 699–709.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. New York: Springer-Verlag.

Van Niel, K. and M. Austin (2007). Predictive vegetation modelling for conservation: Impact of error propogation from digital elevation data. *International Journal of Geographical Information Science 17*(1), 266–280.

Ver Hoef, J. M., N. Cressie, and R. P. Barry (2004). Flexible spatial models for kriging and cokriging using moving averages and the fast fourier transform. *Journal of Computational and Graphical Statistics 13*(2), 265–282.