

A comparison of adaptive sampling designs and binary spatial models : a simulation study using a census of *Bromus inermis*

Kathryn M. Irvine^{a*}, Jamie Thornton^b, Vickie M. Backus^c,
Matthew G. Hohmann^d, Erik A. Lehnhoff^e, Bruce D. Maxwell^e,
Kurt Michels^f and Lisa J. Rew^e

Commonly in environmental and ecological studies, species distribution data are recorded as presence or absence throughout a spatial domain of interest. Field based studies typically collect observations by sampling a subset of the spatial domain. We consider the effects of six different adaptive and two non-adaptive sampling designs and choice of three binary models on both predictions to unsampled locations and parameter estimation of the regression coefficients (species–environment relationships). Our simulation study is unique compared to others to date in that we virtually sample a true known spatial distribution of a nonindigenous plant species, *Bromus inermis*. The census of *B. inermis* provides a good example of a species distribution that is both sparsely (1.9 % prevalence) and patchily distributed. We find that modeling the spatial correlation using a random effect with an intrinsic Gaussian conditionally autoregressive prior distribution was equivalent or superior to Bayesian autologistic regression in terms of predicting to un-sampled areas when strip adaptive cluster sampling was used to survey *B. inermis*. However, inferences about the relationships between *B. inermis* presence and environmental predictors differed between the two spatial binary models. The strip adaptive cluster designs we investigate provided a significant advantage in terms of Markov chain Monte Carlo chain convergence when trying to model a sparsely distributed species across a large area. In general, there was little difference in the choice of neighborhood, although the adaptive king was preferred when transects were randomly placed throughout the spatial domain. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: autologistic regression; Bayesian estimation; intrinsic conditionally autoregressive; low prevalence; strip adaptive cluster sampling

1. INTRODUCTION

A common goal in environmental and ecological studies is to model species distributions across a spatial domain of interest (e.g., Guisan and Thuiller, 2005). Typically, in these observational studies, it is of interest to both make predictions to unsampled areas and to understand species–environment relationships. Modeling nonindigenous plant species (NIS) across large landscapes, but at fine spatial resolution, presents additional challenges because they typically, depending on the stage of invasion, display sparse and patchy spatial distributions (Figure 1). Field based observations are the most reliable way to detect NIS populations. Consequently, data are collected by sampling a subset of the spatial domain of interest and, commonly, recorded as presence or absence due to logistical and cost constraints. For this type of data, two practical questions are (i) how should sampling be conducted to obtain optimal predictions and valid inferences concerning species–environment relationships; and (ii) what type of binary spatial model to use?

* Correspondence to: Kathryn M. Irvine, Northern Rocky Mountain Science Center, US Geological Survey, Bozeman, MT, U.S.A. E-mail: kirvine@usgs.gov

a Northern Rocky Mountain Science Center, US Geological Survey, Bozeman, MT, U.S.A.

b Thornton Statistical Consulting, Seattle, WA, U.S.A.

c Earth Sciences, Montana State University, Bozeman, MT, U.S.A.

d US Army Engineer Research and Development Center, Champaign, IL, U.S.A.

e Department of Land Resources and Environmental Sciences, Montana State University, Bozeman, MT, U.S.A.

f Department of Mathematical Sciences, University of Arizona, Tucson, AZ, U.S.A.



Figure 1. Map of *Bromus inermis* distribution at Little Bighorn Battlefield National Monument in Montana (with roads and trails represented) from survey conducted during Summer of 2010. Gray areas = *Bromus inermis* presences

For binary data that can be considered independent after accounting for environmental predictors, a common choice for analysis is logistic regression (LOGIT). Specifically, denote

$$Y_i = \begin{cases} 1 & \text{if species present in cell } i \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

the probability of species presence (π_i) at spatial location i is modeled as a linear function of environmental predictors via the logit link,

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = X\beta. \tag{2}$$

We consider two spatial extensions to the LOGIT model, autologistic regression (ALR) and an intrinsic Gaussian conditional autoregressive (ICAR) model. The ALR model extends (2) by including an autocovariate, an additional predictor variable, that is constructed as a function of the response values (Y) within a defined neighborhood around location i ,

$$\text{logit}(\pi_i) = X\beta + \delta \text{autocov}_i \tag{3}$$

where $\text{autocov}_i = \sum_{j \in \text{nbrs}(i)} Y_j$ where $\text{nbrs}(i)$ is the set of neighbors for cell i . Previous applications employing ALR for species distribution modeling have found that compared to LOGIT the coefficient estimates (β) in (3) can decrease toward zero, becoming both statistically and biologically non-significant (Gumpertz *et al.*, 1997; Wu and Huffer, 1997; Dormann *et al.*, 2007; Jewell *et al.*, 2007).

Autologistic regression was originally developed for modeling binary data collected on a lattice with all values observed. The issue with estimating the parameters using maximum likelihood is an intractable likelihood due to an unsolvable normalizing constant in the denominator, except for very small datasets (for further discussion, refer to Sherman *et al.*, 2006). Wu and Huffer (1997) and Huffer and Wu (1998) suggest a Markov chain Monte Carlo estimation of the MLE; however, they and Sherman *et al.* (2006) found in the presence of strong autocorrelation this estimation approach might perform poorly. An additional complication arises when the entire spatial lattice is not observed, and then the autocovariate is a function of observed and unobserved data. Missing (unobserved) data are easily accommodated in a Bayesian framework by initializing the Markov chains. Hoeting *et al.* (2000) originally developed a Gibbs sampling estimation method within a Bayesian framework to accommodate missing data. Another more accessible Bayesian estimation option, albeit slow, is to use the Gibbs sampling approach implemented in WinBUGS (Wintle and Bardos, 2006) as we do here.

Another common choice for modeling spatial binary data on a grid is the ICAR model that accounts for spatial autocorrelation using a random effect (Z_i) with a spatial prior (Banerjee *et al.*, 2004; Thomas *et al.*, 2004),

$$\text{logit}(\pi_i) = X\beta + Z_i. \tag{4}$$

The computational advantage with this model is the ability to parameterize the joint distribution of the random effects (\mathbf{Z}) as a series of conditional distributions, as follows:

$$Z_i | Z_{-i} \sim N \left(\bar{S}_i, \frac{\nu}{n_i} \right) \quad (5)$$

where $-i$ is all grid cells except i , $\bar{S}_i = \sum_{j \in nbs(i)} S_j / n_i$ with $j \in nbs(i)$ denoting the set of cells within the neighborhood of cell i , n_i is the number of neighbors for cell i , and ν is the parameter related to the variance of the conditional distributions. In this model, the mean of the conditional distribution is an average of the spatial effects of the neighbors, and the conditional variance is proportional to the number of neighbors. A joint specification of the random effects, such as when using a point-referenced spatial model, requires inversion of the variance-covariance matrix. This is typically an unreasonable computation burden for large datasets; however, Hooten *et al.* (2003) used orthonormal basis functions to decompose the joint distribution of (\mathbf{Z}) for prediction on a large spatial domain, similar in size to ours. For another application of modeling rare plant distributions using ICAR, refer to Ishihama *et al.* (2010); for a more statistical treatment and possible extensions, refer to Pettitt *et al.* (2002).

In this paper, we consider the effects of different sampling designs and choice of binary spatial model (ALR versus ICAR) on both predictions to unsampled locations and parameter estimation of the coefficients (β) and their standard errors using a simulation study. To date, simulation studies exploring prediction performance or parameter estimation for binary spatial models have been based on simulating virtual species distributions (ALR: Wu and Huffer, 1997; Hoeting *et al.*, 2000; Reese *et al.*, 2005; Sherman *et al.*, 2006; Wintle and Bardos, 2006; Dormann, 2007; Van Teefelen and Ovaskainen, 2007; Santika and Hutchinson, 2009; although see Augustin *et al.*, 1996, which subsampled an inventory of deer counts). Our study is unique in that we investigate these questions by virtually sampling a true known spatial distribution of a NIS.

Cluster type designs are preferred for both covariance parameter estimation and prediction of new observations using spatial models with Gaussian error distributions (Zimmerman, 2006; Irvine *et al.*, 2007). For Bayesian estimation of ALR, Wintle and Bardos (2006) found cluster sampling improved convergence properties, but predictive performance was not impacted by the choice of sampling design. These cluster designs consisted of randomly placed clusters or clumps of points or plots on a landscape. These designs do not necessarily maximize the number of presences in the data or increase detections as suggested by Reese *et al.* (2005) to improve Bayesian ALR predictions.

Another way to achieve a clustered pattern of plots on a landscape is by adaptive cluster sampling (ACS). ACS is an efficient and field-tested way to sample rare and sparsely distributed species spatially; thereby increasing the number of detections of a species (Thompson, 1990; Philippi, 2005; Samalens *et al.*, 2007; Morrison *et al.*, 2008). ACS consists of two stages, the first is a probabilistic selection of units from the population (initial set), and the second stage defines the inclusion of additional sampling units within some neighborhood of a selected unit that meets a prespecified condition, such as detection of a species within a plot (Thompson, 2002). Modifications of ACS have been introduced for easier field implementation, such as adaptively adding a larger diameter circular plot if a smaller circular plot meets the desired condition (Yang *et al.*, 2011). Adaptive sampling methods are commonly evaluated in terms of design-based estimates of population means, totals, or densities (refer to the review by Turk and Borkowski, 2005; Abrahamson *et al.*, 2011; Yu *et al.*, 2012). However, Reese *et al.* (2005) explored ACS for predicting species distributions using ALR and found this design resulted in fewer omission errors, to date no investigation of the ICAR and ACS has been performed.

We explore strip ACS designs, in which the initial set sampling unit is a transect or strip composed of 15 contiguous 10 by 10-m plots (Thompson, 2002). Transect designs are commonly employed for sampling NIS because of the decreased effort and time to sample walking along a transect compared to dispersed plots (Rew *et al.*, 2006). We considered two different strategies for selecting the initial sample set of transects: randomly on the landscape (hereafter simple random transects [SRT]), and targeted starting from roads and trails (hereafter targeted transects [TT]). Statistically, the SRT design allows for inferences to the entire landscape (if accessible); however, the logistical tradeoff with more effort to move and sample between transects (Rew *et al.*, 2006) may not overcome the added inferential scope. Populations of invasive plants commonly occur in patches close to roads and trails (Spellerberg, 1998; Parendes and Jones, 2000; Gelbard and Belnap, 2003; Watkins *et al.*, 2003; Pauchard and Alaback, 2004; Sharma and Raghubansi, 2009; Craig *et al.*, 2010; Seipel *et al.*, 2012). Also, any new invasions are likely to occur in such habitats rather than less disturbed sites (Hansen and Clevenger, 2005; Arevalo *et al.*, 2010; Haider *et al.*, 2010; Pollnac *et al.*, 2012). Therefore, TT designs are appealing from a field logistics and biological perspective. We explore whether the inferences concerning species–environment relationships are affected by using the targeted sampling approach compared to randomly placing transects throughout the spatial domain.

In our investigation, the second phase of sampling (adaptive phase) is triggered when an occurrence is detected along the transect. The second phase sampling unit is a 10 by 10-m plot. Various modifications to transect designs that include an adaptive phase were preliminarily assessed by Rew *et al.* (2006) and Maxwell *et al.* (2012) for optimal NIS patch detection, and Brummer *et al.* (2013) for predictions to unsampled areas using LOGIT. Here, we evaluate whether adding an adaptive phase leads to better model performance, particularly, for spatial binary models. We also consider different neighborhood definitions for determining how the adaptively added plot is selected based on the known dispersal characteristics of our censused species.

This article is organized as follows. We describe our simulation study based on virtually sampling a census of smooth brome (*Bromus inermis* Leyss.) from Little Bighorn Battlefield National Monument, Montana in 2010 (Section 2). We present our results exploring the effects of sampling designs and binary spatial models in terms of both prediction and determining species–environment relationships (Section 3). We conclude with a discussion of our results in the context of other simulation studies and provide guidance on preferred sampling designs and spatial model for sparse and clustered populations when spatial modeling is of interest (Section 4).

2. SIMULATION STUDY

Simulation studies investigating the effect of sampling design choice on spatial model performance typically generate multiple stochastic realizations (virtual landscapes) from a specified spatial model and then different sampling designs are overlaid. The benefit of such an approach is that the generating model and parameter values are known. In our case, we know the true spatial distribution across the landscape sufficient for considering predictive performance as the true presence/absence values are known over the entire spatial domain. However, as we are also interested in estimation of species–environment relationships, technically we do not know the true parameter values (β or δ) for the census data. Therefore, we begin with a brief description of the census data (Section 2.1), and then we use ALR, LOGIT, and ICAR to estimate the relationship between log-odds of *B. inermis* and environmental predictors using the complete census dataset (Section 2.2) for comparison to parameter estimation results based on sub-sampling the spatial domain.

2.1. Description of census data

Bromus inermis is a rhizomatous grass in a late stage of invasion at Little Bighorn Battlefield National Monument, and it is distributed in dispersed patches throughout the inventoried area. We chose *B. inermis* for the study because the species has not been managed, and it was one of the most abundant species inventoried by Lehnhoff and Lawrence (2010). Their census was conducted by following parallel 30-m wide transects distributed across their entire study area, and recording the length and width for each *B. inermis* patch encountered using a TrimbleGeoExplorer (XT or XH) global positioning system receiver. We used a 195 ha area of the censused landscape as the basis for the simulated sampling events (Figure 1). There was one main road traversing the study site, as well as smaller roads and interpretative trails. We used ArcGIS 9.3 (ESRI, 2009) and the census data to create the virtual landscape at a 10 m resolution by first using the length and width data to delineate *B. inermis* patches and then rasterizing these patches. The result was a 121 × 161 grid (19,481 cells total), where the presence or absence of *B. inermis* was known for each grid cell. Additional environmental predictors included slope, elevation, and aspect derived from a 10 m digital elevation model (US Geological Survey), and distances to roads and trails that were calculated from spatial data provided by the National Park Service.

2.2. Analysis of census data

We used logistic regression to model the probability of *B. inermis* occurrence as a function of the available environmental predictors. We used a backwards-forwards variable selection procedure using the stepAIC function in R version 2.12.2 (The R Development Core Team, 2011). The full set of predictor variables included: *X*-coordinate, *Y*-coordinate, elevation, slope (degrees), aspect, distance to roads, distance to trails, slope * aspect, elevation * distance to roads, and slope * distance to roads. This full set was determined based on exploratory data analysis; the full model had an AIC value of 3153.504. The final model with the lowest AIC (3150.138) was

$$\begin{aligned} \text{logit}(\pi_i) &= \beta_0 + \beta_1 X - \text{coordinate}_i + \beta_2 Y - \text{coordinate}_i + \beta_3 \text{Elevation}_i + \beta_4 \text{Slope}_i \\ &\quad + \beta_5 \text{Dist2rds}_i + \beta_6 \text{Dist2trails}_i + \beta_7 \text{Dist2rds} * \text{Elevation}_i + \beta_8 \text{Slope} * \text{Dist2rds}_i \\ &= X\beta \end{aligned} \tag{6}$$

where π_i is the probability that *B. inermis* is detected within grid cell *i* and all predictors were standardized, $(x_{ij} - \bar{x}_j)/SD(x_j)$ for all *j* predictors. The delta AIC was only 3 points suggesting marginal support for the top model, and only 14.9% of the deviance was explained by the model. The logistic regression model in (6) ignores the likely spatial autocorrelation in the data.

For ALR, we specified the same mean structure as for LOGIT, and the autocovariate was calculated by summing the eight neighboring cell values, unweighted. We explored the simplest type of autocovariate, but other more complicated functions of the neighboring cell values could be calculated based on the known biology of the species. We did explore whether the choice of neighborhood affected the AIC and there was no improvement with larger neighborhoods, following the suggested methods by Van Teeffelen and Ovaskainen (2007). Also, based on the fact *B. inermis* has relatively short dispersal distances, this definition seemed reasonable.

We estimated the LOGIT and ALR models using the glm function with (family = binomial) in the MASS package of R. Notice for the ALR model this produces maximum pseudolikelihood (MPL) estimates; this approach is pseudolikelihood because we are assuming conditional on the predictors the observations are still independent. We used Bayesian estimation for the parameters of the ICAR model. We assumed the neighbor set ($nbs(i)$) were the eight adjacent cells, as we did in the ALR model. We used the package geoBUGS and the car.normal prior on the random spatial effects to fit the ICAR model in WinBUGS (Lunn *et al.*, 2000; Thomas *et al.*, 2004). The ICAR model has an additional variance parameter (ν in (5)), we specified a gamma prior on $1/\nu$, specifically $\text{Gamma}(0.5, 0.0005)$, as suggested in Kelsall and Wakefield (1999). We used normal priors for the regression coefficients, $\beta \sim N(0, \tau)$ with τ , the precision parameter, equal to 0.01.

A comparison of ICAR, ALR (MPL estimation), and LOGIT models fit to the census data suggests that the LOGIT and ICAR point estimates are similar for the environmental predictors (Table 1). The standard errors were consistently larger for both spatial models compared to the LOGIT, as we would expect (Table 1). However, the standard error estimates were larger for the ICAR compared to the ALR model. This is not too surprising based on other studies that found the standard errors are under-estimated using MPL (e.g., Wu and Huffer, 1997; Sherman *et al.*, 2006).

2.3. Sampling designs

We explore two options for placing the initial sample set of transects (15 contiguous 10 m cells in a line) in a real population of *B. inermis*: SRT or a random starting location along roads or trails (TT, these represent perpendicular transects in Rew *et al.* 2006; Figure 2, first split

Table 1. Standardized coefficient estimates and standard errors of environmental variables for the complete census dataset, using logistic regression, the pseudomaximum likelihood estimation procedure of the autologistic regression model, and the intrinsic Gaussian conditional autoregressive (ICAR)

Predictor variable	Logistic	Autologistic	ICAR
Intercept	-5.46 (0.13)	-5.32 (0.13)	-7.84 (0.37)
X-coordinate	0.52 (0.10)	0.27 (0.12)	0.43 (0.20)
Y-coordinate	1.52 (0.10)	0.42 (0.12)	1.66 (0.18)
Slope	-0.30 (0.07)	-0.19 (0.08)	-0.32 (0.10)
Elevation	0.02 (0.16)	-0.43 (0.19)	-0.03 (0.28)
Distance to roads	-0.84 (0.16)	-0.70 (0.17)	-1.23 (0.29)
Distance to trails	1.04 (0.12)	0.29 (0.13)	1.08 (0.19)
Elevation* distance to roads	-0.93 (0.07)	-0.34 (0.09)	-1.10 (0.11)
Slope* distance to roads	-0.17 (0.06)	-0.04 (0.07)	-0.007 (0.12)
Autocovariate	NA	1.18 (0.05)	NA

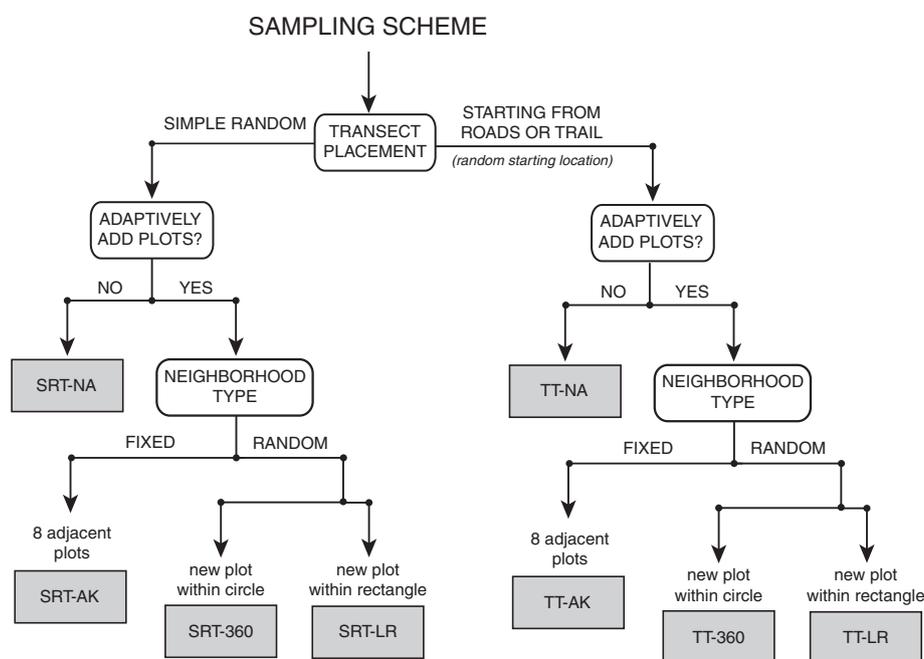


Figure 2. Decision tree describing eight different sampling designs explored in simulation study, abbreviations for designs are enclosed in boxes. The first split is how transects were placed within the spatial domain, the second whether or not plots were added if a detection was made along the transect, and the third was choice of neighborhood based on classic kings move (-AK) or modified based on dispersal distance (-360 or LR). All adaptive designs sampled eight adjacent cells if nonindigenous plant species was detected within the neighborhood. Sampling stopped once 50 additional plots were added.

on decision tree). Adaptive sampling is an alternative to conventional sampling, in which, for our case, the selection of additional grid cells (or 10 by 10-m plots) within a specified neighborhood was initiated if a NIS was observed in one of the plots within one of the initially selected transects (refer to Thompson, 2002; Figure 2 split two on decision tree). We explored three different neighborhood definitions (Figure 2, split three on decision tree). One neighborhood is fixed in the sense that the criterion of a presence is met and sampling proceeds in a prescriptive fashion, whereas the other two neighborhood types add an element of stochasticity for where the adaptively added plot should be sampled. For the fixed neighborhood, adaptive king design (AK), we defined the neighborhood as the eight adjacent plots surrounding an occupied plot (a king's move in the game of chess). If additional NIS were detected in any of the neighborhood plots, a new king's move neighborhood was added to the survey. Only plots not previously surveyed or added to the sample were included in the new neighborhood, sampling without replacement. Surveying continued until no neighboring plots contained NIS or the number of neighborhood plots surveyed reached 50 (Figure 3(a)).

We also investigated whether there was any benefit to defining the neighborhood based on the known dispersal of the species. For our species, *B. inermis*, reproduction is both vegetative and by seed, and dispersal distances are short. We modified the AK design where upon

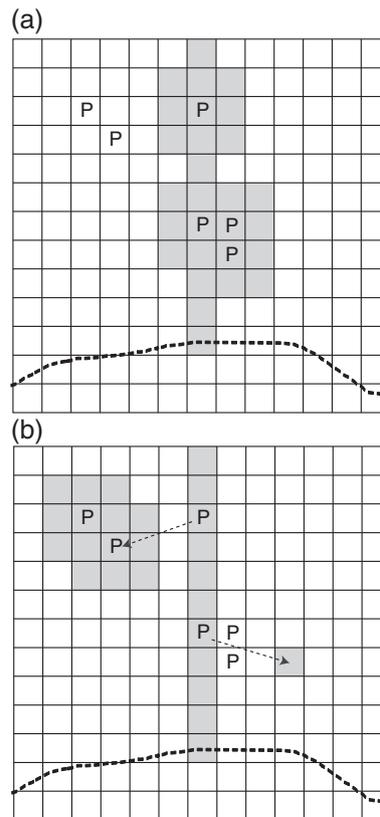


Figure 3. Explanatory depiction of a small portion of simulated sampling using adaptive cluster designs (a) adaptive king (AK) and (b) adaptive 360 (360), where white cells represent non-sampled, shaded cells represent sampled, and P represents cells with *Bromus inermis* present. The dash line indicates a road. For the 360 method (b), arrows indicate the cell selected randomly within a circle of radius 30 m and centered on the transect cell where *B. inermis* was detected. The adaptive LR design used a rectangle with the same width of 30m.

observing a NIS presence in the original transect, instead of adding the immediate neighboring plots to the transect, a randomly selected plot that was within a circle centered at the current occupied transect plot and at a radius equal to the 30 m was surveyed first. If NIS was detected within the random plot, additional neighbor plots were added to the survey and sampled as described for the AK design. We denote this design as 360 (Figure 3(b)). We also used a rectangle with length equal to the 30 m, such that if a detection was made along the original transect, then a random plot to the left or right of the plot with NIS on the transect was selected for sampling (abbreviated as LR).

A sampling event consisted of placing 64 transects (approximately 5% of the study area prior to adaptive sampling) based on either the SRT or TT design for the first phase. During the second phase of sampling, additional plots were sampled based on one of the three adaptive cluster designs (AK, 360 or LR). We simulated 10 independent sampling events for a total of 80 unique datasets using computer programs we developed using the open-source programming language Python (Python Software Foundation, 2007). Although we considered less intense sampling (1% and 2%), we encountered convergence issues for the spatial models, consistent with other simulation studies (Reese *et al.*, 2005; Wintle and Bardos, 2006) investigating Bayesian estimation of ALR.

2.4. Statistical models and estimation methods

For each of the 80 datasets, we fitted three distinct models, non-spatial LOGIT, ICAR, and ALR. Also, we explored two different estimation methods for the ALR. Bayesian estimation can be computationally time intensive for large datasets, so we also explored a commonly used short-cut estimation approach for the ALR model based on MPL extended to sampled data (Wu and Huffer, 1997). We used the same set of environmental predictors as in Table 1 for all models.

For Bayesian estimation of ALR in WinBUGS version 1.4.3 (Lunn *et al.*, 2000), we adopted the code from Wintle and Bardos (2006). Refer to Section 2.2 for our specification of the Bayesian ICAR model; the same was used for the sampled datasets. We used initial values based on the MPL estimates of the parameters. We used the same diffuse normal priors for all the environmental predictor regression coefficients and the autocovariate (if appropriate). We ran three chains of 200,000 iterations with a burn in period of 100,000. The remaining iterations were thinned by 200 to produce a total of 1500 posterior draws for each parameter. We calculated the Gelman–Rubin diagnostic to evaluate convergence for all parameters.

We considered a short-cut option for estimating the parameters of ALR using sampled data based on approaches described in the ecological literature (Augustin *et al.*, 1996; Dormann *et al.*, 2007; Van Teeffelen and Ovaskainen, 2007; Santika and Hutchinson, 2009). We created a prediction probability for each grid cell based on the observed sampled data. The predicted probability of occurrence is

$$\hat{\pi}_i = \exp(X_i \hat{\beta}) / [1 + \exp(X_i \hat{\beta})] \tag{7}$$

where $\hat{\beta}$ is estimated using LOGIT. The autocovariate was then calculated based on the predicted probability of occurrence values in the neighborhood of each grid cell as, $autocov_i = \sum_{j \in nbs(i)} \hat{\pi}_i$, where $nbs(i)$ is the set of neighbors for cell i . The parameters of the ALR were then estimated using the $glm()$ function; we denote this approach as MPL.

2.5. Predictive performance criteria

We calculated the area under the receiver operating characteristic (ROC) curve (AUC) as a measure of predictive performance. The ROC curve plots false-positive prediction rates against true-positive prediction rates. AUC scores range from 0 to 1.0 and a score of 0.5 indicates that the model discriminated among plots with or without *B. inermis* no better than random chance, whereas a score of 1.0 indicates that the model discriminated among plots with or without *B. inermis* perfectly. We used the ROCR package in R to estimate AUC (Sing *et al.*, 2009). For each of the 80 sampling events, we used the sample data to build the models (LOGIT, ICAR, ALR: Bayesian and MPL) and the remaining unsampled inventory values for the validation set (hold-out set).

We used paired *t*-tests to compare two model results for the same dataset, for example, ALR compared to ICAR for SRT-AK design, to assess whether the AUC values were significantly different. For a given model to determine whether there was any predictive benefit to adding an adaptive phase to the initial transect, we used a two-way ANOVA with all pairwise comparisons adjusted by Tukey correction factor. Within a given transect placement (SRT or TT), the simulation design is a randomized complete block design without replication. The sampling event was the blocking factor with 10 levels and the sampling design was the treatment factor with four levels (NA, AK, 360, LR).

2.6. Parameter estimation criteria

To compare the difference in parameter estimates of the environmental predictors, we calculated the median over the 10 different 95% posterior (confidence) intervals for each of the four model/estimation techniques (LOGIT, ICAR, ALR: Bayesian and MPL) for each sampling design. We were interested in assessing whether there was a substantial difference in both the uncertainty in the point estimates and the point

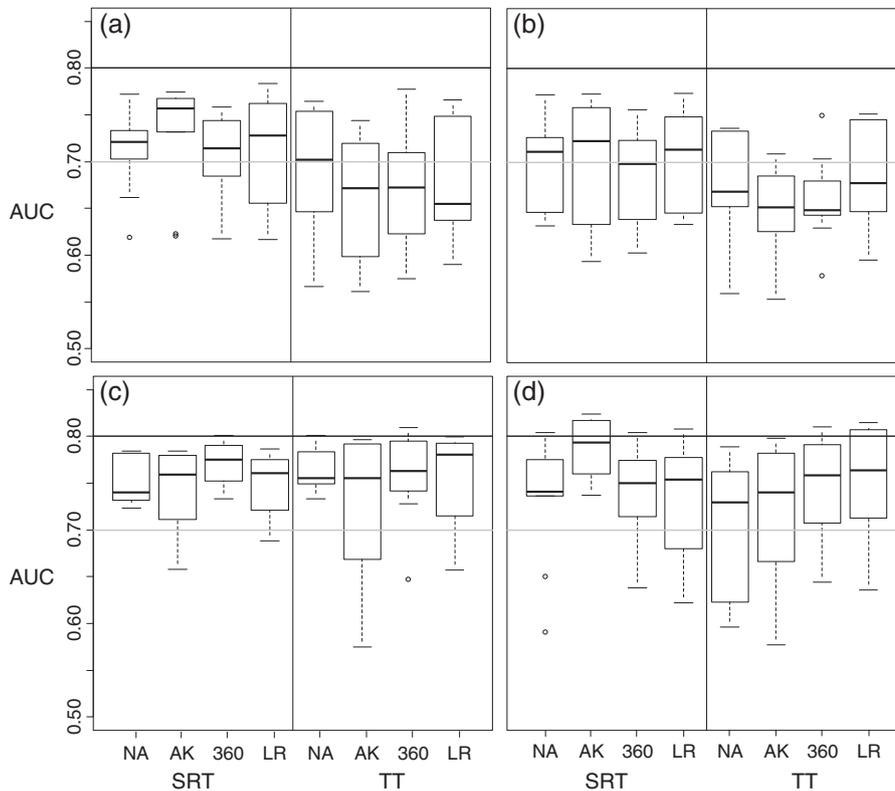


Figure 4. Box and whisker plots of area under the curve values (AUC) for (a) logistic regression with spatially explicit predictors, two different estimation methods for the autologistic regression model (b) MPL and (c) Bayesian; and (d) ICAR spatial model all used same environmental predictors. Each model was fit to the 10 replicate datasets for the eight different sampling designs (NA, non-adaptive transect method; AK, adaptive king; 360, adaptive 360; and LR, adaptive left or right methods) applied along simple random transects (SRT) or targeted transects (TT). Reference lines at AUC value = 0.70 and = 0.80 are included for easier comparison across panels.

estimates themselves. Our only basis for comparison of the effects of sampling design on inferences are the parameter point estimates based on the inventory data (Table 1), so we cannot compare model estimates to a true generating model as is common in most investigations to date.

3. RESULTS

In terms of predictions to unsampled areas, the Bayesian spatial models were superior for most sampling designs investigated (Figure 4(a) and (b) compared to (c) and (d)). The two non-Bayesian models, MPL estimation for ALR and LOGIT, were essentially equivalent in terms of predictive power for all designs (Figure 4(b)), on average different by only 0.01 to 0.02 AUC units (based on paired *t*-tests). For all designs, the Bayesian ALR produced higher median AUC compared to the MPL estimation of ALR (Figure 4(b) compared to (c)). Bayesian ALR and ICAR models were not statistically different except for the SRT-AK and TT-NA designs (Figure 4(c) and (d)). The Bayesian ALR model was better, on average, compared to the ICAR model for the TT-NA design, whereas the ICAR was superior for the SRT-AK design, on average.

There was not a consistently “best” design for prediction. For the TT designs, there was not a clear distinction among the different types of neighborhood in terms of prediction; further, there was not a substantial benefit to adding an adaptive phase for both Bayesian spatial models. In terms of the SRT, for Bayesian ALR, the SRT-360 was statistically better than the other neighborhood types (SRT-AK and SRT-LR), but not practically different, only 0.03 units (Figure 4(c)). However, the SRT-AK design was clearly superior when using the ICAR (Figure 4(d)).

The Bayesian ALR suffered convergence issues in all, or a subset of the regression parameters (Gelman–Rubin statistics > 1.2) for 20 of the 80 datasets. Interestingly, most occurred with the non-adaptive designs, when the adaptively added units for these sampling events were included in the analysis, convergence was attained and predictive performance improved dramatically. The ICAR did not show convergence issues based on the Gelman–Rubin statistics. However, some of the posterior intervals for the environmental predictors were unreasonably

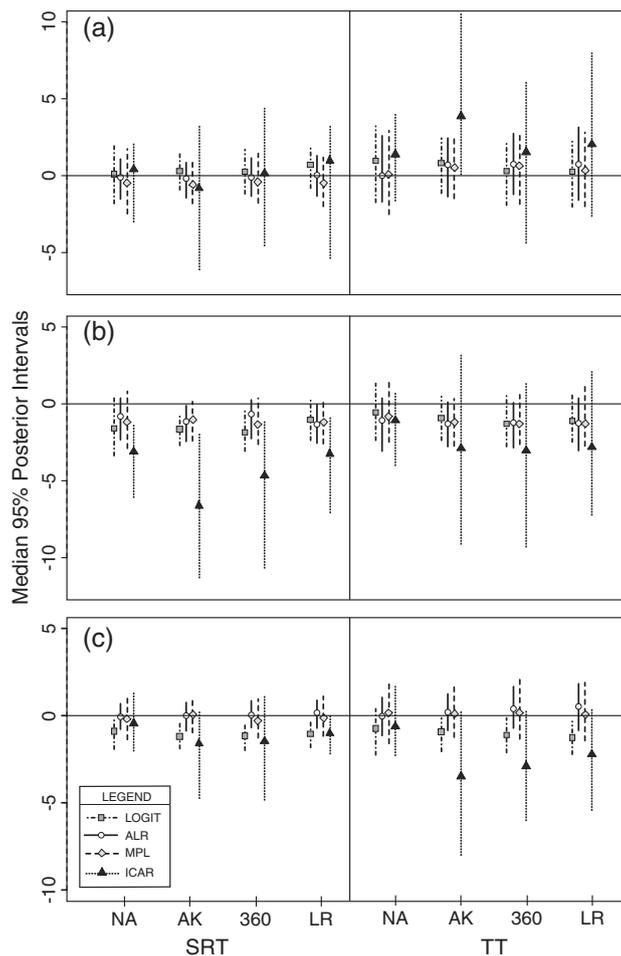


Figure 5. Median of the 95% interval estimates for (a) elevation, (b) distance to roads, and (c) elevation * distance to roads regression coefficients using Bayesian estimation for ALR and ICAR models, maximum pseudolikelihood estimation for ALR (MPL), and LOGIT for the eight different sampling designs (NA, non-adaptive; AK, adaptive king; 360, adaptive 360; and LR, adaptive left or right methods) applied along simple random transects (SRT) or targeted transects (TT). The black horizontal line is the zero reference line.

large. We adjusted the normal priors to be less diffuse $1.0E-2$ versus $1.0E-4$; in general, patterns were the same. We present the results for parameter estimation based on the less diffuse priors for the ICAR model (Figure 5).

In terms of evaluating the relationship between *B. inermis* and environmental predictors, the spatial models would produce similar conclusions across the different designs with most of the predictors' posterior intervals suggesting very weak or no association. For brevity, we show only estimation results for elevation (Figure 5(a)), distance to roads (Figure 5(b)), and elevation * distance to roads (Figure 5(c)). The most striking pattern was that for all the adaptive designs the ICAR posterior intervals were much wider compared to the other models. The ICAR point estimates were similar in sign and effect size to the LOGIT, although not always.

For elevation, the predictor with the smallest effect size (Figure 5(a)), the results for all models and designs would lead to the same general conclusions as using the fully censused dataset. For the predictor with the largest effect size (distance to roads Figure 5(b)), the ICAR estimates using the SRT designs would suggest a negative association with the log-odds of *B. inermis*, whereas the ALR would suggest no association (posterior intervals all contain zero). However, with the TT adaptive designs the models would all suggest no or a very weak negative association. For the interaction term (elevation * distance to roads, Figure 5(c)), LOGIT and ICAR produced negative point estimates compared to essentially zero for the ALR methods using the adaptive designs. These patterns were similar to model results based on the census dataset. Specifically, ICAR point estimates had larger uncertainties and effect sizes closer to LOGIT compared to ALR (Table 1).

A key practical question, based on the ubiquitous use of the MPL approach in ecology, is whether there is a trade-off for ecological inferences when using this short-cut estimation approach compared to the full Bayesian implementation. In terms of the autocovariate, the Bayesian ALR median 95% posterior intervals were substantially narrower compared to the MPL for all designs. The other predictors showed fewer differences between the two estimation methods for ALR (Figure 5). The choice of neighborhood for the adaptive phase (AK, LR, versus 360) appeared less important for either estimation method. Interestingly, both estimation methods had narrower median intervals for the adaptive designs compared to the non-adaptive designs.

4. DISCUSSION AND CONCLUSIONS

Our simulation study builds upon and deepens our understanding of the benefit of cluster sampling designs for spatial models by virtually sampling a complete inventory, as compared to simulating the spatial distribution (Hoeting *et al.*, 2000; Reese *et al.*, 2005; Wintle and Bardos, 2006; Irvine *et al.*, 2007). The strip adaptive cluster designs we explored provided a significant advantage when trying to model a sparsely distributed species across a large area in several aspects. A subtle benefit was improving convergence properties of the Bayesian ALR model. These issues were likely a result of the large spatial autocorrelation in our data, based on others' findings (Huffer and Wu, 1998; Sherman *et al.*, 2006). The spatial models had the same or better predictive performance using the adaptive designs compared to the non-adaptive design. The best predictive performance was achieved using ICAR and the SRT-AK design, randomly placed transects with the AK neighborhood to include additional plots. Adaptive designs produce clustered patterns of plots on a landscape based on attempting to maximize the number of observations with presences, so the fact they were better is consistent with other studies exploring sampling designs for spatial modeling (Hoeting *et al.*, 2000; Reese *et al.*, 2005; Zimmerman, 2006; Irvine *et al.*, 2007).

The use of TT designs, dependent on terrain and access, result in a large time savings; however, there can be a mismatch between the sampled and target population. This is an example of under-coverage as some areas have zero probability of being sampled, a source of non-sampling error (Thompson, 2002). We found that for *B. inermis*, there was little difference in predictive performance using the Bayesian spatial models between the TT and SRT designs. However, the LOGIT and MPL estimation of the ALR model tended to have lower AUC values using the TT designs. Similarly, the selection bias did not drastically affect the conclusions concerning environmental predictors as most were non-significant once spatial autocorrelation was modeled. However, for our landscape, the environmental gradients were not substantially different in those areas with zero probability of being selected using TT designs. The only exception was the predictor distance to roads; the SRT designs using ICAR suggested a negative association, whereas the TT designs suggested no association. This is likely because for this predictor the environmental gradient was sampled better using the SRT designs. This suggests perhaps the better option for sampling NIS is a combination of an SRT and TT design, easily accommodated in an unequal probability design.

In general, the choice of sampling strategy for NIS depends on both the species characteristics (particularly, stage of invasion) and the management goal. The species distribution data we used, *B. inermis*, is actually quite common relative to other NIS species. NIS presence on the landscape is typically quite low (e.g., <1% of the target area) particularly during the early stages of invasion when it is most critical to detect patches to prevent spread. An exploration of sampling designs for maximizing the number of patches detected for the purpose of early detection and rapid response management found optimal designs varied based on species characteristics. An optimal design for *B. inermis* was, essentially, TT-AK compared to a species (*Linaria dalmatica*) earlier in the invasion (<5 years) for which walking along road corridors was optimal for patch detection with the least amount of travel time (Maxwell *et al.*, 2012). Based on our results, for species considered at their maximum spread, no longer invading, a combination of the SRT and TT in an unequal probability design would be ideal for both detecting patches and modeling spatial distribution. Although, further investigation into optimal allocation with unequal probability surveys based on balancing travel time and sampling environmental gradients would be beneficial.

In terms of drawing inferences about the relationships between NIS probability of occurrence and environmental predictors, the posterior interval widths differed considerably between the ICAR and other models. In our simulation, the ICAR had larger uncertainty estimates and effect sizes closer to the LOGIT model for the adaptive designs. The dramatically larger uncertainty estimates for some situations was likely an artifact of an over-specified mean structure once we accounted for spatial autocorrelation. When we reduced the predictor set, the intervals were narrower albeit still wider than the independent LOGIT, as we would expect. Another approach would be to use a restricted spatial regression model (RSR), this is an emerging tool for spatial analysis to minimize the possible confounding of environmental predictors (fixed effects) and the spatial process (modeled via random effects) (Reich *et al.*, 2006; Hodges and Reich, 2010). In a comparison of ICAR and

RSR for binary data, Johnson *et al.* (2013) found dramatic improvements using the RSR in terms of reducing credible intervals. Future work exploring model selection for binary spatial models, including the emerging RSRs, would be insightful.

Bromus inermis has been widely introduced throughout the USA for forage and revegetation purposes. It has relatively short dispersal distances with both sexual and vegetative reproduction. Importantly for our study, the species has not been controlled in the survey area. Therefore, we assume the species has reached equilibrium with its environment. However, only 39.2% of the deviance in the inventory data was explained by the ALR model. This suggests that other relevant predictors *may* exist such as biophysical variables, likely leading to low predictive performance (AUC < 0.9) relative to other previous studies that used simulated species distributions (Hoeting *et al.*, 2000; Reese *et al.*, 2005; Wintle and Bardos, 2006). Our study is generalizable to other species distribution modeling efforts in which the species of interest has similar statistical properties; low prevalence and clusters of presences dispersed throughout the domain and located near corridors (in our case a road, but could be a stream or river corridor).

However, there are inherent drawbacks with using real data; there are a limited number of distinct sampling events possible depending on the size of the spatial domain and the true species–environment relationships are unknown. Our spatial domain was quite large (19,481 grid cells) compared to other previous simulation studies (Hoeting *et al.*, 2000; Reese *et al.*, 2005; Wintle and Bardos, 2006; Dormann *et al.*, 2007); however, we only had 10 unique sampling events for each design. The number of simulations was limited, primarily because of the high computational burden for using Bayesian estimation for ALR, and the fact that the size of the spatial domain resulted in redundancies in the simulated sampling events. Regardless, we feel the general patterns we describe would hold with NIS distributions with similar characteristics to *B. inermis*, that is, later in the invasion process with low prevalence (1.91%) and occurring in dispersed patches located close to roads and trails.

Acknowledgements

We would like to thank Patrick Lawrence and Dan Campbell for helping with field data collection, and Patrick for initial data collation and analysis; both John Long and William Barnett who assisted in the early work on the simulation study and statistical analysis of the inventory data; Cynthia Hollimon for LaTeX assistance; and Mike Ebinger and Tyler Brummer for improving figure quality. Funding from ERDC CERL grant W9132T-09-2-0012. Little Bighorn Battlefield National Monument for supply of research site and plant distribution data. Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the US Government.

REFERENCES

- Abrahamson IL, Nelson CR, Affleck DL. 2011. Assessing the performance of sampling designs for measuring the abundance of understory plants. *Ecological Applications* **21**: 452–464.
- Augustin NH, Muggleston MA, Buckland ST. 1996. An autologistic model for the spatial distribution of wildlife. *Journal of Applied Ecology* **33**: 339–347.
- Arevalo JR, Otto R, Escudero C, Fernandez-Lugo S, Arteaga M, Delgado JD, Fernandez-Palacios JM. 2010. Do anthropogenic corridors homogenize plant communities at a local scale? A case studied in Tenerife (Canary Islands). *Plant Ecology* **209**: 23–35.
- Banerjee S, Carlin BP, Gelfand AE. 2004. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall, CRC Press: Boca Raton, FL.
- Brummer TB, Maxwell BD, Higgs M, Rew LJ. 2013. Surveying non-native species occurrence and modeling realized distributions at local and landscape scales. *Diversity and Distributions* **19**: 919–932.
- Craig DJ, Craig JE, Abella SR, Vanier CH. 2010. Factors affecting exotic annual plant cover and richness along roadsides in the eastern Mojave Desert, USA. *Journal of Arid Environments* **74**: 702–707.
- Dormann CF. 2007. Assessing the validity of autologistic regression. *Ecological Modelling* **207**: 234–242.
- Dormann CF, McPherson JM, Araújo MB, Bivand R, Bollinger J, Carl G, Davis R, Hirzel A, Jetz W, Kissling WD, Kühn I, Ohlemüller R, Peres-Neto PR, Reineking B, Schröder B, Schurr FM, Wilson R. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography* **30**: 609–628.
- ESRI. 2009. ArcGIS 9 Desktop. Version 9.3.1, Environmental Systems Research Institute, Redlands, CA.
- Gelbard JL, Belnap J. 2003. Roads as conduits for exotic plant invasions in a semiarid landscape. *Conservation Biology* **17**: 420–432.
- Guisan A, Thuiller W. 2005. Predicting species distribution: offering more than simple habitat models. *Ecology Letters* **8**: 993–1009.
- Gumpertz ML, Graham JM, Ristaino JB. 1997. Autologistic model of spatial pattern of phytophthora epidemic in bell pepper: effects of soil variables on disease patterns. *Journal of Agricultural, Biological, and Environmental Statistics* **2**: 131–156.
- Haider S, Alexander J, Dietz H, Trepl L, Edwards PJ, Kueffer C. 2010. The role of bioclimatic origin, residence time and habitat context in shaping non-native plant distributions along an altitudinal gradient. *Biological Invasions* **12**: 4003–4018.
- Hansen MJ, Cleverger AP. 2005. The influence of disturbance and habitat on the presence of non-native plant species along transport corridors. *Biological Conservation* **125**: 249–259.
- Hodges JS, Reich BJ. 2010. Adding spatially correlated errors can mess up the fixed effect you love. *The American Statistician* **64**: 325–334.
- Hoeting JA, Leecaster M, Bowden D. 2000. An improved model for spatially correlated binary responses. *Journal of Agricultural, Biological, and Environmental Statistics* **5**: 102–114.
- Hooten MB, Larsen DR, Wikle CK. 2003. Predicting the spatial distribution of ground flora on large domains using a hierarchical Bayesian model. *Landscape Ecology* **18**: 487–502.
- Huffer FW, Wu H. 1998. Markov chain Monte Carlo for autologistic regression models with application to the distribution of plant species. *Biometrics* **54**: 509–524.
- Irvine KM, Gitelman AI, Hoeting JA. 2007. Spatial designs and properties of spatial correlation: effects on covariance estimation. *Journal of Agricultural, Biological, and Environmental Statistics* **12**: 1–20.
- Ishihama F, Takeda T, Oguma H, Takenaka A. 2010. Comparison of effects of spatial autocorrelation on distribution predictions of four rare plant species in the Watarase wetland. *Ecological Research* **25**: 1057–1069.
- Jewell KJ, Arcese P, Gergel SE. 2007. Robust predictions of species distribution: spatial habitat models for a brood parasite. *Biological Conservation* **140**: 259–272.
- Johnson DS, Conn PB, Hooten MB, Ray JC, Pond BA. 2013. Spatial occupancy models for large data sets. *Ecology* **94**: 801–808.

- Kelsall JE, Wakefield JC. 1999. Discussion of “Bayesian models for spatially correlated disease and exposure data” by Best *et al.* In *Bayesian Statistics 6*, Bernardo JM, Berger JO, Dawid AP, Smith AFM (eds). Oxford University Press: Oxford; 151.
- Lehnhoff E, Lawrence P. 2010. Revised Inventory of Non-Indigenous Plants at Little Bighorn Battlefield National Monument. Final Report on CESU task agreement J1380099006. http://www.cfc.umd.edu/CESU/NEWCESU/Assets/Individual%20Project%20Reports/NPS%20Projects/MSU/2009/09Lehnhoff_LIBI_weed%20inventory_Fnl_rpt.pdf.
- Lunn DJ, Thomas A, Best N, Spiegelhalter D. 2000. WinBUGS—a Bayesian modeling framework: concepts, structure, and extensibility. *Statistics and Computing* **10**: 325–337.
- Maxwell BD, Backus VM, Hohmann MG, Irvine KM, Lehnhoff E, Rew LJ. 2012. Comparison of transect based standard and adaptive sampling methods for invasive plant species. *Invasive Plant Science and Management* **5**: 178–193.
- Morrison LW, Smith DR, Young CC, Nichols DW. 2008. Evaluating sampling designs by computer simulation: a case study with the Missouri bladderpod. *Population Ecology* **50**: 417–425.
- Parendes LA, Jones JA. 2000. Role of light availability and dispersal in exotic plant invasion along roads and streams in the H. J. Andrews Experimental Forest, Oregon. *Conservation Biology* **14**: 64–75.
- Pauchard A, Alaback PB. 2004. Influence of elevation, land use, and landscape context on patterns of alien plant invasions along roadsides in protected areas of south-central Chile. *Conservation Biology* **18**: 238–248.
- Pettitt AN, Weir IS, Hart AG. 2002. A conditional autoregressive Gaussian process for irregularly spaced multivariate data with application to modelling large sets of binary data. *Statistics and Computing* **12**: 353–367.
- Philippi T. 2005. Adaptive cluster sampling for estimation of abundances within local populations of low-abundance plants. *Ecology* **86**: 1091–1100.
- Pollnac FW, Seipel T, Repath C, Rew LJ. 2012. Plant invasion at landscape and local scales along roadways in the mountainous region of the Greater Yellowstone Ecosystem. *Biological Invasions* **14**: 1753–1763.
- Python Software Foundation. 2007. Version 2.5. Available at <http://www.python.org>.
- R Development Core Team. 2011. R: A language and environment for statistical computing, R foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org>.
- Reese GC, Wilson KR, Hoeting JA, Flather CH. 2005. Factors affecting species distribution predictions: a simulation modeling experiment. *Ecological Applications* **15**: 554–564.
- Reich BJ, Hodges JS, Zadnik V. 2006. Effects of residual smoothing on posterior of the fixed effects in disease-mapping models. *Biometrics* **62**: 1197–1206.
- Rew LJ, Maxwell BD, Dougher FL, Aspinall R. 2006. Searching for a needle in a haystack: evaluating survey methods for non-indigenous plant species. *Biological Invasions* **8**: 523–539.
- Samalens JC, Rossi JP, Guyon D, Halder IV, Menassieu P, Piou D, Jactel H. 2007. Adaptive roadside sampling for bark beetle damage assessment. *Forest Ecology and Management* **253**: 177–187.
- Santika T, Hutchinson MF. 2009. The effect of species response form on species distribution model prediction and inference. *Ecological Modelling* **220**: 2365–2379.
- Seipel T, Kueffer C, Rew LJ, Daehler CC, Pauchard A, Naylor BJ, Alexander JM, Edwards PJ, Parks CG, Arevalo J, Cavieres LA, Dietz H, Jakobs G, McDougall K, Otto R, Walsh N. 2012. Processes at multiple scales affect richness and similarity of non-native plant species in mountains around the world. *Global Ecology and Biogeography* **21**: 236–246.
- Sharma GP, Raghubanshi AS. 2009. Plant invasions along roads: a case study from central highlands, India. *Environmental Monitoring and Assessment* **157**: 191–198.
- Sherman M, Apanasovich TV, Carroll RJ. 2006. On estimation in binary autologistic spatial models. *Journal of Statistical Computation and Simulation* **76**: 167–179.
- Sing T, Sander O, Beerenwinkel N, Lengauer T. 2009. ROCR: Visualizing the performance of scoring classifiers. <http://CRAN.R-project.org/package=ROCR>. R package version 1.0-3.
- Spellerberg I. 1998. Ecological effects of roads and traffic a review. *Global Ecology and Biogeography* **7**: 317–333.
- Thomas A, Best N, Lunn D, Arnold R, Spiegelhalter D. 2004. GeoBUGS User Manual Version 1.2.
- Thompson SK. 1990. Adaptive cluster sampling. *Journal of the American Statistical Association* **85**: 1050–1059.
- Thompson SK. 2002. *Sampling*, 2nd edn. John Wiley and Sons, Inc.: New York.
- Turk P, Borkowski JJ. 2005. A review of adaptive cluster sampling: 1990–2003. *Environmental and Ecological Statistics* **12**: 55–94.
- Van Teeffelen AJA, Ovaskainen O. 2007. Can the cause of aggregation be inferred from species distributions? *Oikos* **116**: 4–16.
- Watkins RZ, Chen J, Pickens J, Brosfoske KD. 2003. Effects of forest roads on understory plants in a managed hardwood landscape. *Conservation Biology* **17**: 411–419.
- Wintle BA, Bardos DC. 2006. Modeling species-habitat relationships with spatially autocorrelated observation data. *Ecological Applications* **16**: 1945–1958.
- Wu H, Huffer FW. 1997. Modelling the distribution of plant species using autologistic regression model. *Environmental and Ecological Statistics* **4**: 49–64.
- Yang H, Kleinn C, Fehrmann L, Tang S, Magnussen S. 2011. A new design for sampling with adaptive sample plots. *Environmental and Ecological Statistics* **18**: 223–237.
- Yu H, Jiao Y, Su Z, Reid K. 2012. Performance comparison of traditional sampling designs and adaptive sampling designs for fishery-independent surveys: a simulation study. *Fisheries Research* **113**: 173–181.
- Zimmerman DL. 2006. Optimal network design for spatial prediction, covariance parameter estimation and empirical prediction. *Environmetrics* **17**: 635–652.