Chapter 5

Special Topics in Point Pattern Analysis

Here we consider two useful modeling problems for spatial point patterns. The first is concerned with species distributions in ecology and occupies Sections 5.1 and 5.2. In Section 5.1 we consider the role of spatial point patterns in the analysis of presence-only species data. In Section 5.2 we consider preferential sampling, a concept that is attracting attention these days. The idea of preferential sampling is to assess whether there is stochastic dependence between the set of locations where observations are observed and the observations at those locations. First, we use preferential sampling in order to extend customary presence/absence modeling. Then, we use it to address fusion of presence-only data with presence/absence data. In Section 5.3 we consider multivariate spatial point pattern modeling. With M point patterns to model, we work in a marked point process setting. Customarily, as we developed in Section 2.6, we consider locations assigned, perhaps dependently, within a mark but independently across marks. Here, we consider the case where there is dependence between the locations for one mark and those for another. We focus on two flexible classes of models for multivariate point processes - multivariate Gibbs processes and multivariate log Gaussian Cox processes.

We need to add some more words with regard to Sections 5.1 and 5.2. Learning about species distributions is, arguably, a preoccupation in the ecology community. The literature separates two types of data collection to learn about species distributions: presence/absence and presence-only. The former imagines some sort of designed sampling where plots (grid cells, transects, etc.) are sampled and presence/absence or abundance of a species is observed for the sampled plots. Presence-only data is imagined in terms of randomly encountering a species within a region and is typically collected in the form of museum or citizen science data, In fact, the distinction between the two types of data collection can be murky since, if data collection is viewed through gridding of cells, then, conceptually, the observations associated with the cells can be imagined as capturing presence/absence as well as presence-only, as we elaborate below. In any event, the literature on modeling presence/absence data is enormous by now and, more recently, there has been a consequential growth in the literature addressing modeling for the presence-only setting.

The contribution here is to address some fundamental and occasionally contentious threads in the literature with regard to the foregoing data collection. It is asserted that a common modeling framework can be used for both data types, that presence/absence data modeling can be induced under a presence-only framework, and, moreover, that presence-only data can be used to infer about presence/absence [57, 96, 175]. A further implication is that fusion of general presence/absence and presence-only data sources can be implemented within what is essentially the presence-only framework [155].

We step into the fray first with discussion to define what "presence at a location" means, and argue that modeling for the two data types is distinct and incompatible.

Next, we introduce preferential sampling to clarify how potential bias in sampling locations can affect inference with regard to presence/absence, the so-called preferential sampling problem and the associated "shared process" perspective. Then, we turn to the fusion problem, again arguing that current versions of such fusion in the literature have fundamental flaws. We propose to employ the shared process perspective for implementing the fusion, extending preferential sampling ideas. This enables the two data sources to be probabilistically independent or dependent. Altogether, this perspective enables us to consider a collection of models and allow us to take the presence/absence modeling to a much richer explanatory level.

In order to examine these issues, we need to spend some time with the presence-only literature, describing suitable modeling. This is the contribution of Section 5.1. We also need to offer the same with regard to the presence/absence literature. This is addressed in Section 5.2. Evidently, we need to elaborate what preferential sampling is in order to reveal its utility for these issues. In the interest of keeping the explication at a concise and comfortable level, we only consider individual species models. However, extension to joint species distribution modeling can be developed.

5.1. Spatial modeling of presence-only data

5.1.1. A few initial words on presence/absence data

Again, learning about species distributions is a long-standing issue in ecology with an enormous literature. Useful review papers that organize and compare model approaches include [60, 216] and references therein. Following the overarching objective of this monograph, our focus here is on model-based approaches to study this problem. A substantial proportion of the model-based work focuses on modeling presence/absence where the data are available as a presence (1) or absence (0) at a collection of sampling locations. The goal is to explain the probability of presence at a location given the environmental conditions that are present. The natural approach is to build a binary regression model, with, say, logistic link, where the covariates can be introduced linearly (see below) or as smoothly varying functions.

The latter choice results in generalized additive models (GAMs) which tend to fit data well since they employ additional parameters that enable nonlinear and multimodal relationships with the data [60, 93]. They can also provide a qualitative picture of how species respond to environmental variables. The price we pay for using GAMs is a loss of simplicity in interpretation and the risk of overfitting with poor out-of-sample prediction. We don't consider GAMs further here, preferring the use of random effects specified through Gaussian processes. Random effects models are extremely flexible, offer direct interpretation, and good spatial prediction (kriging).

Much of this presence/absence work is *non-spatial* in the sense that, though it includes spatial covariate information, it does not model anticipated spatial dependence in presence/absence probabilities. Accounting for the latter seems critical since causal ecological explanations such as localized dispersal as well as omitted (unobserved) explanatory variables with spatial pattern such as local smoothness of soil or topographic features suggest that, at sufficiently high resolution, occurrence of a species at one location will be associated with its occurrence at neighboring locations [205]. In particular, such dependence structure, introduced through spatial random effects, facilitates learning about presence/absence for portions of a study region that have not been sampled, accommodating gaps in sampling and irregular

sampling effort. For point level categorical response, [98] use a Gaussian process (GP) prior for these spatial effects. For areal level count data, Markov random field (MRF) priors [16, 27] have been used in [7] and later incorporated into a hierarchical Bayesian model setting by [75, 76] and [39]. See also [118] in this regard. We return to presence/absence data, elaborating the foregoing modeling approaches in Section 5.2.

5.1.2. Review of presence-only data approaches

The focus of the work in this section is on the so-called *presence-only* setting. Analysis of presence-only data has seen growing popularity in recent years due to increased availability of such records from museum databases and other non-systematic surveys [see 90]. We note that presence-only data is not *inferior* to presence/absence data. In fact, it can be viewed as the converse; in principle, presence-only data offer a complete census while presence/absence data, since confined to a specified set of sampling sites, contains less information. However, in practice, a complete census of individuals is rarely achieved. The sampling effort required to achieve such censuses usually exceeds the available time and money resources.

One model-based strategy for presence-only data has attempted to implement a presence/absence approach. All of this work depends upon drawing so-called *background samples*, a random sample of locations in the region with known environmental features. Early work here characterized these samples as pseudo-absences [62, 67] and fitted a logistic regression to the observed presences and these pseudo-absences. Since presence/absence is unknown for these samples, work of [157, 211] shows how to adjust the resulting logistic regression to account for this. Additionally, all of this work is non-spatial in the above sense. More importantly, as we argue below, this approach attempts to condition in the wrong direction. The observed presences can be viewed as a point pattern, revealing its relevance for this monograph. See [38, 212] in this regard. With a point pattern of absences we could imagine a marked point pattern with a mark for presence and a mark for absence. However, pseudo-absences create an unobserved and artificial pattern of absences. In fact, even with a complete census of individuals, we will have a finite point pattern of presences with an *uncountable* set of absence locations.

Alternative algorithmic approaches include the genetic algorithm for rule-set prediction (GARP) approach [158] and the maximum entropy (Maxent) approach, [see, e.g., 159, 160]. GARP is based upon an artificial intelligence framework to produce a set of positive and negative rules that, together, give a binary prediction. Rules are favored according to their effectiveness (compared with random prediction) based upon a sample of background data and presence data. Maxent is a constrained optimization method which finds the optimal species density (closest to a uniform) subject to moment constraints. Maxent predictions have usually been found to have higher predictive accuracy on average than GARP [60]. Moreover, with the availability of an attractive software package[1], Maxent is now becoming the standard approach for presence-only data analysis. The point pattern analysis approach we present here provides an appealing alternative in that it is fully model-based, allowing full inference with associated uncertainty everywhere in the region.

Briefly, the Maxent approach produces a probability density surface which maximizes entropy given constraints imposed by the collection of vectors of environ-

[1] http://www.cs.princeton.edu/∼schapire/maxent

mental variable values at the sites at which the species has been observed. These constraints require that the average of each of the environmental covariates under this distribution *essentially* agrees with the empirical average for this covariate based upon samples over the region. The constrained optimization introduces regularization weights, one for each moment constraint. The optimization is solved only approximately, i.e., each constraint is satisfied within a specified precision to avoid overfitting. As an optimization strategy rather than a stochastic modeling approach, Maxent is unable to attach any uncertainty to resulting optimized estimates. The resultant surface is interpreted as providing the relative probability of observing a species at a given location compared to other locations in the region. However, Maxent is unable to provide an intensity, meaning we are unable to determine, for example, the expected number of individuals in a specified region.

Again, our approach is to model presence-only data as a point pattern with an associated intensity specified in terms of the available environments across the region. We do this through typical regression modeling, enabling natural interpretation for the coefficients. We employ a hierarchical model to introduce spatial structure for the intensity surface through spatial random effects, resulting in a log Gaussian Cox process following Section 2.3. We do not assume any background or pseudo-absence samples; rather, we assume that the covariates we employ are available as surfaces over the region in order to interpolate an intensity over the entire region. We acknowledge that the observed point pattern is biased through anthropogenic processes, e.g., human intervention to transform the landscape and non-uniform (in fact, often very irregular) sampling effort. Such bias in sampling is a common problem, see for example [128] and references therein. This requires adjusting the *potential* species intensity to a *realized* intensity which we treat as a *degradation* of the former.

Variation in site access is one of the factors that influences the likelihood of the site to be visited/sampled. For example, sites adjacent to roads or along paths, near urban areas, with public ownership, e.g., state or national parks, or with flat topography are likely to be over-sampled relative to more inaccessible sites. When bias implies that only a portion of the region is sampled, it is likely that only a portion of the overall point pattern is observed. In addition, there may be temporal bias in sampling. For example, as one learns more about the ecology of the species of interest, the nature of sampling effort may change [127]. One might build a regression model to attempt to explain sampling bias but no successful versions have appeared in the literature to date.

Land use, as a result of human intervention, affects *availability* of locations, hence, inference about the intensity. As a result of human intervention, some areas within a study region are not available for a species. Also, agricultural transformation and dense stands of alien invasive species preclude availability. Transformed areas are not sampled, and this information must be included in the modeling. Altogether, sampling tends to be sparse and irregular; we rarely collect a random sample of available environments.

Detection can affect inference regarding the intensity. That is, we may incorrectly identify a species as present which is actually absent (false presence) or fail to detect a species that is actually present (false absence) [166]. Evidently, the prevalence of these false records will affect the attempt of an explanatory model on response to environmental features [201]. Modeling for these errors can be attempted but requires information beyond the observed presence-only data and is not considered below.

Lastly, an attractive by-product of our proposed modeling is the opportunity to study species richness, that is, the expected number of distinct species in a specified

region. We can do this by thinking of the species as a mark/label associated with a point and that the entire point pattern arises as a superposition (Section 4.1). In particular, this enables us to obtain potential and observed richness surfaces.

5.1.3. Modeling details

We assume a log Gaussian Cox process (LGCP) model for the set of presence locations. We have to introduce degradation caused by sampling bias as well as by land transformation. As a result, we conceptualize a *potential* intensity, i.e., the intensity in the absence of degradation, as well as a *realized* (or effective) intensity that operates in the presence of degradation. Further, we imagine that the intensity is tiled to grid cells at the resolution of the available environmental covariate surface.

We imagine three surfaces over a region of interest, D. First, $\lambda(\mathbf{s})$ is the intensity in the absence of degradation. With $\int_D \lambda(\mathbf{s})d\mathbf{s} = \lambda(D)$, $f(\mathbf{s}) = \lambda(\mathbf{s})/\lambda(D)$ gives the potential density over D. Modeling for $\lambda(\mathbf{s})$ under a LGCP expects the environmental covariates, say $\mathbf{x}(\mathbf{s})$ to influence the intensity as a linear form in parameters. So, for any location $s \in D$, we have

$$\log\lambda(\mathbf{s}) = \mathbf{x}^T(\mathbf{s})\gamma + z(\mathbf{s}) \tag{5.1}$$

with $z(\mathbf{s})$, a zero-mean stationary, isotropic GP over D, to capture residual spatial association in the $\lambda(\mathbf{s})$ surface across grid cells. The Matérn family of covariance functions (Section 1.4) would provide a flexible class for isotropic dependence; a simple exponential covariance function, $\sigma^2 e^{-\phi||\mathbf{s}-\mathbf{s}'||}$ may be adequate. In the sequel, with regard to preferential sampling, we employ (5.1) as the model for an available presence-only dataset.

Next, we envision an availability surface, $U(\mathbf{s})$, a binary surface over D such that $U(\mathbf{s}) = 1$ or 0 according to whether location $\mathbf{s}$ is untransformed (hence, available) by land use or not. That is, assuming no sampling bias, $\lambda(\mathbf{s})U(\mathbf{s})$ can only be $\lambda(\mathbf{s})$ or 0 according whether $\mathbf{s}$ is available or not. Thirdly, we envision a sampling effort surface over D which we denote as $T(\mathbf{s})$. $T(\mathbf{s})$ is also a binary surface and $T(\mathbf{s})U(\mathbf{s}) = 1$ indicates that location $\mathbf{s}$ is both available and sampled. Altogether, $\lambda(\mathbf{s})U(\mathbf{s})T(\mathbf{s})$ becomes the degradation at location $\mathbf{s}$. This implies that in regions where no locations were sampled, the operating intensity for the species is 0.

Suppose we partition D into grid cells with $A_i, i = 1, 2, ...I$ denoting the geographical region corresponding to grid cell i. Typically the grid is at the resolution of the predictors used in explaining $\lambda(\mathbf{s})$. Then, if we average $U(\mathbf{s})$ over A_i, we obtain $u_i = \int_{A_i} U(\mathbf{s})d\mathbf{s}/|A_i|$ where $|A_i|$ is the area of cell i. Evidently, u_i is the proportion of cell i that is transformed. u_i can often be obtained through remote sensing for all grid cells. Now, bringing in the sampling effort surface, $T(\mathbf{s})$, we can set $q_i = \int_{A_i} T(\mathbf{s})U(\mathbf{s})d\mathbf{s}/|A_i|$ and interpret q_i as the probability that a randomly selected location in A_i was available and sampled. Thus, we can capture availability and sampling effort at areal unit scale.

Altogether, $\lambda(\mathbf{s})U(\mathbf{s})T(\mathbf{s})$ becomes the degradation at location $\mathbf{s}$. This implies that in regions where no locations were sampled, the operating intensity for the species is 0. Additionally, $\int_{A_i} T(\mathbf{s})d\mathbf{s}/|A_i|$ can be viewed as the sampling probability associated with cell i. Then, if $T(\mathbf{s})$ is viewed as random, the expectation of the integral would yield $\int_{A_i} p(\mathbf{s})d\mathbf{s}/|A_i|$ where, now, $p(\mathbf{s}) = P(T(\mathbf{s}) = 1) \in [0, 1]$. Clearly, $p(\mathbf{s})$ gives the local probabilities of sampling, not a probability density over D.

To go forward, we assume that $\lambda(\mathbf{s})$ is independent of $T(\mathbf{s})U(\mathbf{s})$. That is, the potential intensity for a species is independent of the degradation process. Then, omitting the details, we can write $\int_{A_i} \lambda(\mathbf{s})T(\mathbf{s})U(\mathbf{s})d\mathbf{s} = \lambda_i q_i$ where $\lambda_i = \int_{A_i} \lambda(\mathbf{s})d\mathbf{s}$ is the cumulative intensity associated with cell A_i and, again, $q_i = \int_{A_i} T(\mathbf{s})U(\mathbf{s})d\mathbf{s}/|A_i|$. However, it is not sensible to imagine that sampling effort is independent of land transformation. In fact, we might expect less sampling attention to be paid to more transformed areas [166, 204]. More directly, if $U(\mathbf{s}) = 0$ then $T(\mathbf{s}) = 0$. Hence, if we define $q_i = u_i p_i$, then $p_i = \frac{\int_{A_i} T(\mathbf{s})U(\mathbf{s})d\mathbf{s}}{\int_{A_i} U(\mathbf{s})d\mathbf{s}}$, i.e., p_i is the conditional probability that a randomly selected location in cell i is sampled given it is available. As an illustration, we might set p_i equal to 1 or 0 which we interpret as $T(\mathbf{s}) = U(\mathbf{s})\ \forall \mathbf{s} \in A_i$ or $T(\mathbf{s}) = 0\ \forall \mathbf{s} \in A_i$, respectively. That is, either all available sites in A_i were visited or no available sites in A_i were visited.

Again, to model the potential intensity surface $\lambda(\mathbf{s})$, we employ a log Gaussian Cox process model (LGCP). We expect the environmental covariates, say $\mathbf{x}(\mathbf{s})$ to influence the intensity and model the mean as a linear combination of them. Then for any location $s \in D$, we have the model given in (5.1) with $z(\mathbf{s})$ capturing residual spatial association in the $\lambda(\mathbf{s})$ surface across grid cells.

Recalling the grid above, suppose we have n_i presence locations $(\mathbf{s}_{i,1}, \mathbf{s}_{i,2}, \ldots, \mathbf{s}_{i,n_i})$ within grid cell i for $i = 1, 2, \ldots, I$. Following the discussion above, $U(\mathbf{s}_{i,j})T(\mathbf{s}_{i,j}) \equiv 1$, $0 \le j \le n_i, 1 \le i \le I$. Then the likelihood function becomes

$$L(\lambda_D; \{\mathbf{s}_{i,j}\}) \propto e^{-\int_D \lambda(\mathbf{s})U(\mathbf{s})T(\mathbf{s})\, d\mathbf{s})} \prod_{i=1}^{I} \prod_{j=1}^{n_i} \lambda(\mathbf{s}_{i,j}) \tag{5.2}$$

Although we have only finitely many presence locations, the integral term in L involves the uncountable random field $\lambda_D = \{\lambda(\mathbf{s}) : \mathbf{s} \in D\}$. Fortunately, we have a natural approximation to the stochastic integral at the scale of grid cells. That is, though we have geo-coded locations for the observed sites, with covariate information at grid cell level, we only attempt to explain the point pattern at grid cell level. In particular, for each cell $i = 1, 2, \ldots, I$, say we are given information on l covariates as $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{il})$. We will also have cell level information about land availability, u_i, across D. With many unsampled cells, many $n_i = 0$.

A computational advantage accrues to working at grid cell level; we can employ a product Poisson likelihood approximation rather than the point pattern likelihood in (5.2). That is, for cell i with geographic region A_i, suppose $\mathbf{c}_i$ is the centroid. Then, given the set $\{\lambda(\mathbf{c}_i), i = 1, 2, ..., I\}$, the n_i are independent and $n_i \sim \text{Po}(|A_i|\lambda(\mathbf{c}_i)q_i)$. Approximation of the point pattern likelihood using such a *tiled* surface over a lattice embedding the region was discussed in [20]. There it is shown that the approximation can be justified in the sense that the resulting approximate posterior converges to the true posterior as the partition gets finer.

Notice that, for any cell with $q_i = 0$ (which can happen if either $p_i = 0$ or $u_i = 0$) there is no contribution from A_i in the product Poisson likelihood. Since, from (5.1), $\log\lambda(\mathbf{s})$ follows a GP, the posterior distribution takes the form

$$\pi(\lambda(\mathbf{s}_{1:m}), \boldsymbol{\gamma}, \boldsymbol{\theta}|\mathbf{n}, \mathbf{x}, \mathbf{u}, \mathbf{q}) \propto \exp\left(-\sum_{i=1}^{I} \lambda(\mathbf{s}_i)\Delta_i q_i\right) \prod_{i=1}^{I} \lambda^{n_i}(\mathbf{s}_i) \times \phi_I(\log\lambda(\mathbf{s}_{1:I})|\boldsymbol{\gamma}, \mathbf{x}, \boldsymbol{\theta})\pi(\boldsymbol{\gamma})\pi(\boldsymbol{\theta})$$

where ϕ_I denotes an I dimensional Gaussian density and $\boldsymbol{\theta}$ denotes the parameters in the covariance function of $z(\mathbf{s})$ in (5.1). Model fitting for (5.2), simplified to (5.3) has been discussed in Section 4.2

We define species richness for a specified region as follows. Relative to a given set of species, the observed richness is the number of distinct species found in that region. Here, we show how our modeling above provides a parametric function for expected richness. By comparison, an often-used approach with Maxent is to merely sum over the individual species densities [149]. The interpretation of such a sum as a richness when integrated over a subregion is possibly unsatisfying and, in any case, no uncertainty can be attached to estimates made using this sum.

Under the presence-only setting, we imagine data arrives in the form, $(\mathbf{s}_j, L(\mathbf{s}_j))$, $j = 1, 2, ..., n$, i.e., a random location and a species label associated with that location, a marked point pattern. Suppose we use the foregoing modeling to create a species intensity, $\lambda_l(\mathbf{s})$, for species $l = 1, 2, ..., L$. For a set A within the study region, we define the richness for A to be the expected number of distinct species in A. Under this definition, we expect more species as A grows larger and no species as the area of A goes to 0.

Let $N(A)$ be the total number of observations in A, i.e., the total number of locations in A where a "presence-only" observation of any species was recorded. Let $N_l(A)$ be the number of locations in A where species l was observed. Finally, let $r(A) = \sum_l 1(N_l(A) > 0)$, where $1(\cdot)$ is the indicator function. Then, $r(A)$ is the "realized" richness in A. Thus, the quantity we seek to infer about is $E(r(A))$. Note that $E(1(N_l(A) > 0)) = 1 - e^{-\lambda_l(A)}$ since $N_l(A) \sim Po(\lambda_l(A))$. Hence, $E(r(A)) = \sum_l (1 - e^{-\lambda_l(A)})$. Evidently, richness is not additive, i.e., $E(r(A_1) \bigcup r(A_2)) \neq E(r(A_1)) + E(r(A_2))$.

With model fitting for each $\lambda_l(\mathbf{s})$, we can obtain posterior samples of $E(r(A))$ for any A by obtaining posterior samples for each $\lambda_l(A)$. Such samples are obtained through appropriate integration (summation for the Poisson approximation version) of $\lambda_l(\mathbf{s})$ over A. If we work with the collection of grid cells A_i, we can supply a richness surface for D. Adjustment for transformation and sampling intensity can be introduced, as above, to distinguish a potential and a degraded surface.

In the context of MCMC , employing a GP on a large collection of locations is computationally demanding because of the necessary repeated inversion of the covariance matrix arising from the process. There are a number of approximation techniques in literature, such as process convolution [97], approximate likelihood [190], fixed rank kriging [44], etc. The *predictive process* method [17] uses dimension reduction to accommodate a high dimensional GP as follows. If $z(\mathbf{s})$ is the zero mean GP under consideration, and our data consist of locations $\mathcal{S} \equiv (\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_I)$ where I is large, then the method proceeds by first choosing r locations $\mathcal{S}^0 = (\mathbf{s}_1^0, \mathbf{s}_2^0, \ldots, \mathbf{s}_r^0)$ from the region, called *knots*. Then, we replace $z(\mathbf{s}_{1:I})$ in the model equation by $\tilde{z}(\mathbf{s}_{1:I}) = E(z(\mathbf{s}_{1:I})|z(\mathbf{s}_{1:r}^0)) = \mathbf{C}z(\mathbf{s}_{1:r}^0)$ where the matrix $\mathbf{C}$ is calculated from the dependence structure of $z(\mathbf{s})$. $\mathbf{C}$ depends on correlation parameters but not on the process variance. In our setting, we apply this approximation to the $\{\lambda(\mathbf{s}_j)\}$ in (5.3) through the $\{w(\mathbf{s}_j)\}$. Bias correction, as discussed in [68], can be introduced.

An alternative strategy to fit such models uses the nearest neighbor Gaussian process (NNGP) [48]. This is a sparse Gaussian process modeling approach rather than a dimension reduction approach. The process is defined through a reference set of locations, along with the remaining locations of interest. A low dimensional neighbor set is attached to each location. The process is specified through univariate conditional normal distributions using the neighbor sets. This enables direct calculation of the inverse of the covariance matrix. Formal inversion of a high dimensional covariance matrix, the usual stumbling block for MCMC implementations with spatial random effects, is by-passed. See [48] for full details.

With regard to inference, posterior samples of $\boldsymbol{\gamma}$ help us to infer whether a particular factor has a significant impact (positive or negative) on species intensity. The ϕ parameter indicates the strength of spatial association for the realization of the intensity surface over D. This association may arise because some potentially important covariates are not available or because the covariate impact is not well captured using a linear form. That is, since Gaussian processes can capture a wide range of dependencies, using them in a hierarchical setting enhances predictive performance for the model.

In terms of informative displays of intensity surfaces, the $\lambda_i p_i$ surface will capture the (lack of) sampling effort. The $\lambda_i u_i$ surface reveals the effect of transformation. Of course, the $\lambda(\mathbf{s})$ surface is most interesting since it offers insight into the expected pattern of presences over all of D. Posterior draws of $\lambda_{1:I}$ can be used to infer about the potential intensity, displaying, say, the posterior mean surface. We can also learn about the potential density $g(\mathbf{s})$ in this discretized setting as $g_i = \lambda_i / \sum_{k=1}^{I} \lambda_k$, and the corresponding density under transformation as $g_{u,i} = \lambda_i u_i / \sum_{k=1}^{I} \lambda_k u_k$. [38] present an example using data for plant species from the Cape Floristic Region in South Africa.

5.2. Spatial modeling for presence/absence data using preferential sampling

5.2.1. Some presence/absence modeling details

Presence/absence data views the observations as binary responses, presence (1) or absence (0) at a collection of sampling locations See, e.g., [60] and references therein for a review. The goal is to explain the probability of presence at a location given the environmental conditions that are present there. The natural approach is to build a binary regression model with say logit or probit link.

Specification of basic individual presence/absence models can consider presence/absence at an areal scale. That is, for a given species, they score a 1 or 0 according to whether or not an individual of the species was present within a specified areal unit. The unit might be say, a grid cell or a quadrat. These models can also be specified at point level, i.e., presence or not at a specified geo-coded location. As noted in Section 1.1, model specification depends upon the choice of scale in terms of defining what probability of presence means. We consider this issue in greater detail below. However, the important point here is that we consider the sampling locations as *fixed* and the associated binary observation at the location to be random.

Suppose $Y(\mathbf{s})$ denotes the presence/absence (1/0) of the species at sample location $\mathbf{s}$. If the study region D is partitioned into grid cells with geographic area A_i, say at the level of resolution of the environmental covariates, then, summing up $Y(\mathbf{s})$ over the number of sample sites, n_i in region A_i, yields grid cell level counts: $Y_{i+} = \sum_{\mathbf{s}\in A_i} Y(\mathbf{s})$. This is an elementary illustration of scaling up from points to areal units. If the sampling site is viewed as the grid cell then $n_i = 1$, a single Bernoulli trial for the cell.

If we assume independence for the trials, a binomial distribution results for Y_{i+}, i.e.,

$$Y_{i+} \sim Binomial(n_i, p_i). \tag{5.3}$$

Explicitly, the probability that the species occurs in cell i, p_i, is related functionally to the environmental variables with a logit link function and a linear (in coefficients) predictor $\mathbf{w}_i^T\boldsymbol{\beta}$:

$$\log\left(\frac{p_i}{1-p_i}\right) = \mathbf{w}_i^T\boldsymbol{\beta}. \tag{5.4}$$

Here $\mathbf{w}_i$ is a vector of explanatory environmental variables associated with cell i and $\boldsymbol{\beta}$ is a vector of the associated coefficients. Here, and in the sequel, we could equally well use a probit link function.

If we model probability of presence at the sample site level, $Y(\mathbf{s})$ would be taken as

$$Y(\mathbf{s}) \sim Bernoulli(p(\mathbf{s})), \tag{5.5}$$

analogously relating the probability that the species occurs in site $\mathbf{s}$, $p(\mathbf{s})$, to the set of environmental variables as $\log\left(\frac{p(\mathbf{s})}{1-p(\mathbf{s})}\right) = \mathbf{w}^T(\mathbf{s})\boldsymbol{\beta}$. Such modeling requires that we have covariate levels $\mathbf{w}(\mathbf{s})$ for each site. This model is referred to as a spatial regression in the sense that the regressors are spatially referenced. If we set $\mathbf{w}(\mathbf{s}) = \mathbf{w}_i$ when $\mathbf{s}$ is within grid i, we return to the same model as in (5.3). Extension of this binary regression model would allow the covariates to be introduced as smoothly varying functions leading to generalized additive models (GAMs), as discussed above.

Next, we extend (5.4) to a simple spatially explicit model, accepting that spatial structure or autocorrelation in ecological pattern and process is pervasive. In the context of species distribution patterns, we would anticipate that the presence/absence of a species at one location may be associated with presence/absence at neighboring locations. This can be achieved by adding spatial random effects to the model. At the grid cell level a spatial term ρ_i associated with grid i is added to (5.4):

$$\log\frac{p_i}{1-p_i} = \mathbf{w}_i^T\boldsymbol{\beta} + \rho_i. \tag{5.6}$$

In (5.6), grid cell i has an associated random effect ρ_i which adjusts the probability of presence of the modeled species up or down, depending on the values in a *spatial neighborhood* of cell i. To capture this behavior, we customarily employ a Gaussian intrinsic or conditional auto-regressive (CAR) model [27]. Such a model proposes that the effect for a particular grid cell should be roughly the average of the effects of its neighboring cells and results in a multivariate normal as the joint distribution over all the cells. There are many ways to specify neighbor structure; see [16] for full discussion.

Most relevant for us for the remainder of this section, we consider a point level spatial model, extending (5.5). For point-referenced data, spatial dependence can be modeled directly between the points based on their relative locations, using Gaussian processes, creating geostatistical models [16]. Recalling the point level model above, we would augment the explanation of $p(\mathbf{s})$ through the form

$$\log\frac{p(\mathbf{s})}{1-p(\mathbf{s})} = \mathbf{w}^T(\mathbf{s})\boldsymbol{\beta} + \omega(\mathbf{s}). \tag{5.7}$$

Here, $\omega(\mathbf{s})$ is the spatial random effect associated with point $\mathbf{s}$, arising as a realization of a Gaussian process. A suitable covariance function would be selected. With a

binary response, this model is referred to a spatial generalized linear model (GLM); see [56]. The first stage sampling mechanism is a Bernoulli trial with the surface of probability of presence as a second stage specification. Inference from (5.7) would be about this surface at any location in the study region, where these probabilities are explained through the spatially referenced predictors. Another surface of interest is the realized presence/absence surface, i.e. $\{Y(\mathbf{s}) : \mathbf{s} \in D\}$.

We note (and discuss in Section 5.2.2, that presence/absence is not observable at point level. A *point* is unitless while we observe a location up to the scale of accuracy of the associated geo-coding for the location. However, this does not preclude useful point level modeling. Indeed, this is the case with all geostatistical modeling [16], e.g., temperature is never observed at a unitless location but we routinely model temperature surfaces. Taking (5.7), with say a probit link, it specifies $P(Y(\mathbf{s}) = 1) \equiv p(\mathbf{s}) = \Phi(\mathbf{w}^T(\mathbf{s})\boldsymbol{\beta} + \omega(\mathbf{s}))$. That is, $P(Y(\mathbf{s}) = 1) = P(z(\mathbf{s}) > 0)$ where $z(\mathbf{s})$ is a Gaussian process with mean $\mathbf{w}^T(\mathbf{s})\boldsymbol{\beta}$, variance 1 (needed to identify the βs), and a suitable correlation function, typically an exponential or Matérn. Under this model, the $Y(\mathbf{s})$ are drawn as conditionally independent Bernoulli trials given $p(\mathbf{s})$. As a result, even if $p(\mathbf{s})$ is smooth, realizations of the presence/absence surface are everywhere discontinuous. Of course, the $Y(\mathbf{s})$'s will be marginally dependent and smoothness of $p(\mathbf{s})$ will encourage a gridded image of a realization to offer a *locally* constant (0 or 1) appearance[2].

An alternative presence/absence specification is a first stage or direct model which introduces a latent Gaussian process at the first modeling stage, setting $Y(\mathbf{s}) = 1(z(\mathbf{s}) > 0)$. Now, if $z(\mathbf{s})$ is, again, a realization of a Gaussian process which is smooth, then the realized $Y(\mathbf{s})$ surface will be locally constant. For instance, if $z(\mathbf{s}) = \mathbf{w}^T(\mathbf{s})\boldsymbol{\beta} + \omega(\mathbf{s})$, as above, with an almost everywhere smooth mean surface, we have this behavior. The first stage modeling approach can be attractive for joint species distribution modeling [43] since it allows direct modeling of dependence between species rather than deferring it to the second stage [153, 154].

Unfortunately, we encounter a technical problem arising in the model fitting. This concerns the difference between the probability of presence surface, $p(\mathbf{s})$, that is, $\Phi(\mathbf{w}^T(\mathbf{s})\boldsymbol{\beta}+\omega(\mathbf{s}))$ under the second stage model and the realized presence surface under the direct model, $1(\mathbf{w}^T(\mathbf{s})\boldsymbol{\beta}+\omega(\mathbf{s}) \geq 0)$. The realized presence surface has to "agree" with the observed presences and absences while the probability of presence surface does not. We can observe a presence that has small probability of occurring or an absence that has a small probability of occurring. As a result, the probability of presence surface does not have to work as hard to fit the data. With $\omega(\mathbf{s})$ in the modeling, under the direct model, the GP has to react strongly to observed presences and absences. Under second stage modeling, it can react less so. Therefore, when fitting the direct model, the $\omega(\mathbf{s})$ surface becomes spiky in the neighborhood of a presence in order to explain well the observed presence. The flexibility of the GP is attractive but, here, its flexibility produces a posterior which is too sensitive to the data. In fact, under the direct model, the probability of presence surface becomes $\Phi(\mathbf{w}^T(\mathbf{s})\boldsymbol{\beta})$, the GP doesn't appear. As a result, for a region over which the covariates are essentially constant, this surface is essentially constant, regardless of the data, making it not very useful.

Can we achieve a locally constant realized presence/absence surface and a smoothed probability of presence surface? A proposal is the following. Still, we let $Y(\mathbf{s}) = 1, 0$ according to $z(\mathbf{s}) \geq 0, < 0$. However, we introduce two GP's in specifying $z(\mathbf{s})$, i.e., $z(\mathbf{s}) = \mathbf{w}^T(\mathbf{s})\boldsymbol{\beta} + \omega(\mathbf{s}) + \gamma(\mathbf{s})$. Here, $\omega(\mathbf{s})$ has a larger range,

[2]See Section 5.2 in this regard.

a smaller decay parameter while $\gamma(\mathbf{s})$ has a smaller range with a larger decay parameter. (Here, we are capturing the frequently used interpretation of the "nugget" as microscale dependence [16]). Then, we define the probability of presence surface as $p(\mathbf{s}) = P(z(\mathbf{s}) \geq 0|\boldsymbol{\beta}, \mathbf{w}(\mathbf{s}), \omega(\mathbf{s})) = \Phi(\mathbf{w}^T(\mathbf{s})\boldsymbol{\beta} + \omega(\mathbf{s}))$ while we define the realized presence/absence surface again as $1(z(\mathbf{s}) \geq 0)$. Since $\gamma(\mathbf{s})$ is smooth, we will have locally constant behavior in this surface. The γ's will be spiky but the ω's will be smoother. Evidently, strong prior information will be needed to control the decay parameters in the GP's. In fact, we would impose an order restriction on the ranges or decays, demanding more rapid decay for the $\gamma(\mathbf{s})$ process. Additionally, we can impose ranges for both the ω's and γ's which are appropriate for the spatial scale of D along with the smallest inter-point distance among the presence/absence locations.

Lastly, if we let $\gamma(\mathbf{s})$ be a pure error process, then we would again obtain the problem of $z(\mathbf{s})$ being everywhere discontinuous so that the realized $Y(\mathbf{s})$ surface would be everywhere discontinuous. However, a pure error process with very small variance will provide results similar to that for a GP with very short range, very rapid decay. The pure error process model will also be easier to fit.

5.2.2. What does "probability of presence" mean?

The models in Sections 5.1.3 and 5.2.1 above emanate from very different probability specifications. In order to integrate them, we need to take a more careful look at what "probability of presence" means. Loose thinking in this regard leads to irreconcilable modeling.

The issue is whether presence/absence is viewed at point level or at areal level. Is it a Bernoulli trial at a location or is it the probability that the number of individuals of a species in set A is ≥ 1? If we model presence/absence at point level, we know what $Y(\mathbf{s}) = 1$ means, but what does $Y(A)$ mean? A coherent probabilistic definition arises as a block average, i.e., a realization of $Y(A)$ is $\int_A 1(Y(\mathbf{s}) = 1)d\mathbf{s}/|A|$, the proportion of the $Y(\mathbf{s})$ that equal 1 in A; it is not a Bernoulli trial and $P(Y(A) = 1) = 0$! We can calculate $E(Y(A)) = \int_A p(\mathbf{s})d\mathbf{s}/|A|$ with $p(\mathbf{s})$ as in (5.7). That is, $E(Y(A))$ becomes the average probability of presence over A. It is the probability that the species is present at a randomly selected location in A.

If $p(\mathbf{s})$ is constant over A then $E(Y(A))$ is this constant probability. This takes us back to the case of gridded regions where we defined p_i, the constant probability over A_i using logistic (or probit) regressions, as in (5.4) and (5.6). Importantly, that areal definition of p_i is interpreted at point level; it is the probability of presence at any site in A_i.

Now, suppose we consider the locations of all individuals in a study region as a random point pattern. Then, if $N(A)$ is the number of individuals in set A, then $P(N(A) \geq 1)$ is probability of presence in A. Here, assuming a nonhomogeneous Poisson process or, more generally a log Gaussian Cox process, $N(A) \sim Po(\lambda(A))$ where $\lambda(A) = \int_A \lambda(\mathbf{s})ds$ for an intensity surface $\lambda(\mathbf{s})$. Then $P(Y(A) = 1) = P(N(A) \geq 1) = 1 - e^{-\lambda(A)}$. Since presence-only data alleges to sample the point pattern (although likely not fully, rather up to sampling effort over the region [38, 69]), it is compatible with this definition of presence/absence. However, the occurrence probability is only defined with regard to the size of A, a concern raised in [70]. Evidently, occurrence probability will vary with the size of A. As a result, it is unclear how to specify a meaningful probability of presence surface. Furthermore, the

definition of probability of presence as "one or more" observations of the species in A yields local distortion to any such surface; $N(A) = 1$ and $N(B) = 11$ are treated the same with regard to presence if $|A| = |B|$, [1]. Moreover, even if we ignore the size of A and return to a grid of cells over D, then it is clear that $p_i \equiv P(Y(A_i) = 1)$ has nothing to do with the p_i defined in the previous subsection.

The two foregoing definitions associated with the probability of presence in A, $P(N(A) \geq 1)$ and $P(Y(A_i) = 1)$, are incompatible and the fundamental difference between them has been ignored in the literature. The conceptualization for the first choice is that we go to fixed "point" locations and see what is there; we are not sampling a point pattern. There is a surface over a domain D which captures the probability of presence at every location in D. The conceptualization for the second is that we identify an area of interest D and we census it completely for all of the occurrences of the point pattern in it. It provides an intensity surface which can be scaled to a density surface. However, as with any probability density function, the density surface at a point is **not** the probability of presence at that point. The second version can not scale down to point level since then $\lambda(A) \to 0$.

Furthermore, if presented with a collection of plots and observed presence/absence for those plots, would one ever model the data as a point pattern? The answer seems clear; no point pattern was observed and there is no way to model an intensity. We would use one of the foregoing presence/absence models. Moreover, if we briefly consider a data fusion problem, suppose one obtains an additional set of presence-only data for the region. Why is it now appropriate to model the same presence/absence data using a point pattern model associated with the presence-only data?

We have articulated the issue with being too informal with regard to the notion of presence as well as the data fusion challenge. Of course, one can disregard the scaling issue, create an arbitrary discretization of the space, and calculate probabilities over the discretization, as in recent work of [155]. In summary, modeling presence/absence at the point level seems the preferable specification. However, in the literature to date, ignoring the incompatibility is the way that presence-only data has been used to provide presence/absence probabilities and the way presence-only data has been fused with presence/absence data.

We briefly digress to a related contentious issue in the recent literature. Can one use presence-only data to infer about presence/absence? [175], imagining an areal unit definition of presence, argue that "occurrence probability can be estimated from presence-only data." In particular, they assume that an environmental covariate, X, is a priori, uniformly distributed over the study region. Then, with $P(Y(A) = 1|X(A)) = \psi(\beta_0 + \beta_1 X(A))$ for link function ψ and a discrete uniform density for $X(A)$, using Bayes' Theorem,

$$f(z(A_i)|Y(A_i) = 1; \boldsymbol{\beta}) = \frac{\psi(\beta_0 + \beta_1 X(A_i))}{\sum_i \psi(\beta_0 + \beta_1 X(A_i))} \tag{5.8}$$

Equation (5.8) suggests that, by modeling environment/habitat given presence, we can learn about $P(Y(A_i) = 1|X(A_i)$. [96] point out that this model is flawed in the sense that the unconditional probabilities, $P(Y(A_i) = 1)$ are not identified; only relative probabilities are identifiable. We would add two further comments here. First, the likelihood in (5.8), in fact is $\Pi_i \psi(\beta_0 + \beta_1 X(A_i))/c(\beta_0, \beta_1)$. This is a different function of the β's than the likelihood for a binary regression with $P(Y(A_i) = 1|\beta_0, \beta_1, X(A_i))$. The parameters do not mean the same thing in the two models and would not provide the same estimates if we could fit the latter. Second, from above, it is unclear what the event $Y(A_i) = 1$ means and, regardless,

the occurrence probabilities being considered here suffer the same issues as above with regard to the size of the A_i.

5.2.3. Preferential sampling

We propose preferential sampling as a tool for both improving presence/absence prediction as well as for fusing presence-only data with presence/absence data. To begin we need to formally develop the concept of preferential sampling.

What is preferential sampling all about?

The notion of preferential sampling was introduced into the literature in the seminal paper of [52]. Subsequently, there has been considerable follow up research. Two useful papers in this regard are [156] and [37]. A standard illustration arises in geostatistical modeling [see e.g. 16, 46]. Consider the objective of inferring about environmental exposures. If environmental monitors are only placed in locations where environmental levels tend to be high, then interpolation based upon observations from these stations will necessarily produce only high predictions. The obvious remedy lies in spatial design of the locations, e.g., a random or space-filling design [151] for locations over the region of interest is expected to preclude such bias. However, sampling for presence/absence may not be designed in this fashion; ecologists may tend to sample where they expect to find individuals, introducing *bias* into the collection of sampling locations. Recognizing the possibility of such bias, can we revise presence/absence prediction to adjust for it? This is the intention of preferential sampling modeling. (The intention is not to attempt to remove preferential sampling.)

We proceed as follows. While the set of sampling locations may not have been developed randomly, we study it as if it were a realization of a spatial point process. That is, it may be designed in some fashion and be deterministic but not with the intention of being roughly uniformly distributed over D. Then, the question becomes a stochastic one: is the realization of the locations independent of the realization of the responses? If no, then we have what is called preferential sampling. Importantly, the dependence here is stochastic dependence. Notationally/functionally, the responses are associated with the locations. We will make this more clear below.

In our context, the presence/absence data has an associated probability of presence surface, as we develop below. This surface plays the role of the "exposure" surface, with the finite set of binary responses, $\mathcal{Y}$, informing about it. Suppose we view the set of sampling locations as a realization of a random point pattern, $\mathcal{S}$. The question we ask is whether $\mathcal{Y}$ is independent of $\mathcal{S}$, again in a stochastic sense? Below, we develop several models, using the idea of a *shared* process, that enable us to address this question and, furthermore, whether $\mathcal{S}$ enables us to improve our inference regarding the presence/absence surface, our prediction of presence.

Preferential sampling models for presence/absence data

To develop the stochastic specifications that formalize preferential sampling for a region D, we imagine two cases for the intensity associated with the point pattern of sampling locations, $\mathcal{S}$:

(i) $\log\lambda(\mathbf{s}) = \mathbf{x}^T(\mathbf{s})\boldsymbol{\gamma}$, i.e., a nonhomogeneous Poisson process (NHPP) and

(ii) $\log\lambda(\mathbf{s}) = \mathbf{x}^T(\mathbf{s})\boldsymbol{\gamma} + \eta(\mathbf{s})$, a log Gaussian Cox process (LGCP).

Here, $\mathbf{x}(\mathbf{s})$ is a vector of predictors with associated regression coefficients $\boldsymbol{\gamma}$ and $\eta(\mathbf{s})$ is a mean 0 GP with a suitable covariance function. Here, we consider only model (ii) to explain the set of sampling locations for the presence/absence data.

Consider modeling for $\mathcal{Y}$. Since we model $Y(\mathbf{s})$ directly through a latent Gaussian process, $z(\mathbf{s})$, i.e., $Y(\mathbf{s}) = 1(z(\mathbf{s}) > 0)$, as at the end of Section 5.2.1, we only need to propose models for $z(\mathbf{s})$.[3] We start with a simple spatial regression,

(a) $z(\mathbf{s}) = \mathbf{w}^T(\mathbf{s})\boldsymbol{\beta} + \epsilon(\mathbf{s})$,

where the predictors in $\mathbf{w}(\mathbf{s})$ and those in $\mathbf{x}(\mathbf{s})$ need not be identical. Extension to a customary geostatistical model for $z(\mathbf{s})$ becomes

(b) $z(\mathbf{s}) = \mathbf{w}^T(\mathbf{s})\boldsymbol{\beta} + \omega(\mathbf{s}) + \epsilon(\mathbf{s})$,

adding $\omega(\mathbf{s})$ as a mean 0 GP, independent of $\eta(\mathbf{s})$ above.

To illuminate the model structure, denote the point pattern over D by $\mathcal{S}$, the realization of ω over D as $\boldsymbol{\omega}_D$, and the realization of η over D as $\boldsymbol{\eta}_D$. Suppose we consider the joint distribution $[\mathcal{S}, \mathcal{Y}, \boldsymbol{\omega}_D]$. We have the natural factorization as $[\boldsymbol{\omega}_D][\mathcal{S}|\boldsymbol{\omega}_D][\mathcal{Y}|\mathcal{S}, \boldsymbol{\omega}_D]$ (suppressing $\boldsymbol{\eta}_D$ if case (ii)). Then, we say that there is no preferential sampling if $[\mathcal{S}|\boldsymbol{\omega}_D] = [\mathcal{S}]$. This is clearly the case with model (a) or (b) inducing $\mathcal{Y}$ and (i) or (ii) for $\mathcal{S}$. Only $\boldsymbol{\omega}_{\mathcal{Y}} = \{\omega(\mathbf{s}_i) : \mathbf{s}_i \in \mathcal{S}\}$ is needed to fit (b).

Now, we can extend model (i) to

(iii) $\log\lambda(\mathbf{s}) = \mathbf{x}^T(\mathbf{s})\boldsymbol{\gamma} + \psi\omega(\mathbf{s})$.

In this notation, with model (b) for $\mathcal{Y}$, $\omega(\mathbf{s})$ is a shared process for both $\mathcal{Y}$ and $\mathcal{S}$ so $\mathcal{Y}$ and $\mathcal{S}$ are not independent. Working with (b) and (iii), if $\psi = 0$, then, following [52], we have non-preferential sampling while if $\psi \neq 0$, we have *strong* preferential sampling.

[156] extended this idea so that $\mathcal{Y}$ follows the geostatistical model (b) while $\mathcal{S}$ follows model (ii). Then, they attempt to interpret $\eta(\mathbf{s})$ as a regressor to add to the geostatistical model for $\mathcal{Y}$. That is, now we have model

(c) $z(\mathbf{s}) = \mathbf{w}^T(\mathbf{s})\boldsymbol{\beta} + \delta\eta(\mathbf{s}) + \omega(\mathbf{s}) + \epsilon(\mathbf{s})$.

Here, the coefficient δ plays a preferential sampling role. For example, if the design $\mathcal{S}$ over-samples locations in D where we have presences, where $Y(\mathbf{s})$ tends to be 1, i.e., $z(\mathbf{s})$ tends to be high, then $\eta(\mathbf{s})$ will tend to be high around those locations. Therefore, $\eta(\mathbf{s})$ can be a significant predictor for $z(\mathbf{s})$ (hence for $Y(\mathbf{s})$) with $\delta > 0$. With (ii) and (c), $\eta(\mathbf{s})$ is the shared process. Only $\boldsymbol{\eta}_{\mathcal{Y}} = \{\eta(\mathbf{s}_i) : \mathbf{s}_i \in \mathcal{S}\}$ is needed to fit (c).

A further shared process model for $\mathcal{Y}$ that can be explored in this regard extends (a) to

[3] With a second stage model we would still introduce a latent Gaussian process, $z(\mathbf{s})$.

(d) $z(\mathbf{s}) = \mathbf{w}^T(\mathbf{s})\boldsymbol{\beta} + \delta\eta(\mathbf{s}) + \epsilon(\mathbf{s})$.

Here, interest is in comparing (d) and (ii) with (a) and (ii); is $\delta \neq 0$?

[52] focus on comparing (b) and (i) with (b) and (iii). [156] focus on comparing (ii) and (b) with (ii) and (c). [37] add another GP to the intensity for $\mathcal{S}$, i.e.,

(iv) $\log\lambda(\mathbf{s}) = \mathbf{x}^T(\mathbf{s})\boldsymbol{\gamma} + \eta(\mathbf{s}) + \xi(\mathbf{s})$.

That is, using model (iv) with model (c), there is a shared GP for $\mathcal{Y}$ and $\mathcal{S}$ as well as individual GP's for each, a total of three independent GP's altogether. [37] acknowledge identifiability problems with the three latent Gaussian fields. We offer Table 5.1 which provides a summary of the modeling choices for $\mathcal{S}$ and $\mathcal{Y}$.

TABLE 5.1
Summary of four modeling choices for $\mathcal{S}$ and $\mathcal{Y}$.

$\mathcal{S}$ models	$\mathcal{Y}$ models
(i) $\log\lambda(\mathbf{s}) = \mathbf{x}^T(\mathbf{s})\boldsymbol{\gamma}$	(a) $z(\mathbf{s}) = \mathbf{w}^T(\mathbf{s})\boldsymbol{\beta} + \epsilon(\mathbf{s})$
(ii) $\log\lambda(\mathbf{s}) = \mathbf{x}^T(\mathbf{s})\boldsymbol{\gamma} + \eta(\mathbf{s})$	(b) $z(\mathbf{s}) = \mathbf{w}^T(\mathbf{s})\boldsymbol{\beta} + \omega(\mathbf{s}) + \epsilon(\mathbf{s})$
(iii) $\log\lambda(\mathbf{s}) = \mathbf{x}^T(\mathbf{s})\boldsymbol{\gamma} + \psi\omega(\mathbf{s})$	(c) $z(\mathbf{s}) = \mathbf{w}^T(\mathbf{s})\boldsymbol{\beta} + \delta\eta(\mathbf{s}) + \omega(\mathbf{s}) + \epsilon(\mathbf{s})$
(iv) $\log\lambda(\mathbf{s}) = \mathbf{x}^T(\mathbf{s})\boldsymbol{\gamma} + \eta(\mathbf{s}) + \xi(\mathbf{s})$	(d) $z(\mathbf{s}) = \mathbf{w}^T(\mathbf{s})\boldsymbol{\beta} + \delta\eta(\mathbf{s}) + \epsilon(\mathbf{s})$

As a last comment here, we return to the question of independence of the various GPs. For example, is it appropriate to assume that $\eta(\mathbf{s}$ and $\omega(\mathbf{s})$ are independent? In fact, if we look say, at model (c) with model (ii), the mean surface $E(Y(\mathbf{s})) = \mathbf{w}^T(\mathbf{s})\boldsymbol{\beta} + \delta\eta(\mathbf{s}) + \omega(\mathbf{s})$ and $\lambda(\mathbf{s}) = \mathbf{x}^T(\mathbf{s})\boldsymbol{\gamma} + \eta(\mathbf{s})$ are dependent through coregionalization [16, chapter 8]. The coregionalization matrix is $\mathbf{A} = \begin{pmatrix} \delta\sigma_\eta^2 & \sigma_\omega^2 \\ \sigma_\eta^2 & 0 \end{pmatrix}$

In application, we might focus on a subset of model comparison. For instance, we might compare (a) and (ii) with (d) and (ii). That is, $[\mathcal{Y}|\mathcal{S}, \boldsymbol{\beta}][\mathcal{S}|\boldsymbol{\gamma}, \boldsymbol{\eta}_D]$ vs. $[\mathcal{Y}|\mathcal{S}, \boldsymbol{\beta}, \boldsymbol{\eta}_\mathcal{Y}, \delta][\mathcal{S}|\boldsymbol{\gamma}, \boldsymbol{\eta}_D]$. We might compare (b) and (ii) with (c) and (ii). That is, $[\mathcal{Y}|\mathcal{S}, \boldsymbol{\beta}, \boldsymbol{\omega}_\mathcal{Y}][\mathcal{S}|\boldsymbol{\gamma}, \boldsymbol{\eta}_D]$ vs.
$[\mathcal{Y}|\mathcal{S}, \boldsymbol{\beta}, \boldsymbol{\omega}_\mathcal{Y}, \boldsymbol{\eta}_\mathcal{Y}, \delta][\mathcal{S}|\boldsymbol{\gamma}, \boldsymbol{\eta}_D]$.
Since the intent is to improve the predictive performance of the model for $\mathcal{Y}$, model comparison criteria should focus on out-of-sample prediction for $Y(\mathbf{s})$.

5.2.4. Fusing presence/absence and presence-only data

We complete this section by turning to the data fusion question. Data fusion (also assimilation) is a widely employed objective when multiple data sources are available to inform about the same response of interest [150, 214]. A canonical example is the goal of modeling exposure to an environmental contaminant when we might have station data available, computer model output available, and perhaps satellite data. The conceptual modeling strategy is to imagine a latent *true* exposure surface and then build a model for each data source conditioned upon the true model. The joint modeling enables each of the sources to inform about the true exposure surface (ref), to enable improved prediction of the exposure surface. Other examples in the literature include application to weather data, sea surface temperature, and animal behavior patterns [177, 178, 215].

Our data fusion setting is different from customary settings. The usual data fusion setting envisions multiple data sources informing about a common response, e.g., ozone level. We have two different types of data. While both inform about

species distribution, we have argued above that presence/absence data is, stochastically, not the same as presence-only data. The fusion approaches in the literature [e.g., 57, 69, 83, 155] ignore this and assume a latent point pattern model for the presence-only data and that the presence/absence data is induced under this model as we described above. Since we argue that a point pattern specification is inappropriate for presence/absence data, we claim that a different type of fusion is required. We have a point pattern model for the presence-only data and a binary map model for the presence/absence data. So, we again turn to preferential sampling ideas [52] in order to explore fusion.

The extra information available to make a data fusion story is $\mathcal{S}_{PO}$, the set of observed presence-only locations. Formally, what information does $\mathcal{S}_{PO}$ bring with regard to learning about the probability of presence surface? Suppose we assume that $\mathcal{S}_{PO}$ is a complete census in D. Associated with $\mathcal{S}_{PO} = \{\mathbf{s}_1^*, \mathbf{s}_2^*, \ldots, \mathbf{s}_m^*\}$, we can imagine a $\lambda_{PO}(\mathbf{s})$ with a similar set of models to (i)-(iv). We expect $\lambda_{PO}(\mathbf{s})$ to be elevated near these observations. For example, analogous to (ii), let $\lambda_{PO}(\mathbf{s}) = \mathbf{x}^T(\mathbf{s})\boldsymbol{\gamma}_{PO} + \eta_{PO}(\mathbf{s})$, using the same predictors as with the presence/absence modeling. Because the mechanisms that created $\mathcal{S}_{PO}$ and $\mathcal{S}_{PA}$ are different, it doesn't make sense that $\mathcal{S}_{PO} \sim \mathcal{S}_{PA}$. So, in order to capture the influence of $\mathcal{S}_{PO}$ on the $p(\mathbf{s})$ surface associated with $\mathcal{Y}_{PA}$, it seems we should add $\delta_{PO}\eta_{PO}(\mathbf{s})$ to the mean for $z(\mathbf{s})$ in (c), i.e. a $\delta_{PA}\eta_{PA}(\mathbf{s})$ term and a $\delta_{PO}\eta_{PO}(\mathbf{s})$ term.

So, we have two sources for possible preferential sampling, one for each dataset. However, we might insist that $\delta_{PO} > 0$. Then, from the presence-only data, the probability of presence will be increased around the $\mathbf{s}_j^*$'s and decreased away from them. Indeed, the locations in $\mathcal{S}_{PO}$ are severely biased; they are locations where we see only 1's. We are severely over-sampling presences with $\mathcal{S}_{PO}$ and we should increase probability of presence where we do.

In summary, we now have four potential models for $\lambda_{PO}(\mathbf{s})$, parallel to those for $\lambda_{PA}(\mathbf{s})$ to combine with the model for $Y(\mathbf{s})$. Many of these models will be difficult to identify. We might focus our effort on a model for $\mathcal{S}_{PO}$ analogous to model (ii) for $\mathcal{S}_{PA}$. Then, we can add a $\delta_{PO}\eta_{PO}(\mathbf{s})$ term to the mean of $z(\mathbf{s})$ under (b), (c), or (d). In other words, the full model takes the form

$$[\mathcal{Y}|\mathcal{S}_{PA}, \boldsymbol{\beta}, \boldsymbol{\eta}_{PA,Y}, \delta_{PA}, \boldsymbol{\eta}_{PO,Y}, \delta_{PO}][\mathcal{S}_{PA}|\boldsymbol{\gamma}_{PA}, \boldsymbol{\eta}_{PA,D}][\mathcal{S}_{PO}|\boldsymbol{\gamma}_{PO}, \boldsymbol{\eta}_{PO,D}]. \tag{5.9}$$

As a last remark, in practice, with a partial realization of the presence-only point pattern, we need to degrade $\lambda_{PO}(\mathbf{s})$ in the model fitting. Section 5.1 shows how to adjust and fit (5.9) in the presence of a partially observed presence-only point pattern.

5.3. Multivariate point patterns

5.3.1. Introduction

Multivariate spatial point processes are stochastic processes generated in two dimensional space. Each generated point arises from one of two or more qualitatively distinguishable types [53]. So, multivariate point patterns in this context are often referred to as multitype point patterns. The spatial point pattern associated with each type corresponds to an associated sub-point pattern in the multivariate point pattern; the aggregation of say, M spatial point patterns is the multivariate point

pattern. For example, the common bivariate spatial point process consists of the collection of spatial points where each point is of one of two types.

Multivariate point process models are particularly important models in ecology for describing spatial patterns of a set of species and can be used to identify phylogenetic and functional diversity. For example, there is often interest in analyzing a multivariate point pattern consisting of competing species to assess interaction within and between between species. [203] offer a comprehensive review of spatial point pattern analyses in ecology, from data types and summary statistics, to methods of inference, model fit, and statistical tests.

Multivariate spatial point processes can arise as a marked point pattern where the marks identify "type." However, models for multivariate spatial point processes can differ from marked point patterns in some important ways. Recall Section 2.6 on marked point patterns. There, we considered marked point patterns where the spatial locations were assigned (possibly dependently) within a mark but the spatial locations were independent across marks. In addition, marked point processes can also be modeled with one process and a distribution for marks given locations $[m(\mathcal{S})|\mathcal{S}]$ where $\mathcal{S}$ is the entire point pattern.

For a dependent process, e.g., Gibbs process, these models assume the interactions between all spatial points, both points of the same type and points of different type, is the same. That is, the second order structure capturing the dependence or inhibition between points is associated with the realization $\mathcal{S}$ and is not mark-specific.

In the multivariate spatial point process models considered here, we are extending the model space to include the possibility of dependence between spatial locations of varying strength both within and between types. That is, we offer models to capture complex interactions between the spatial locations for two or more marks. In what follows, "mark" and "type" are used interchangeably.

We begin with summary statistics for multivariate spatial point patterns to detect spatial clustering and inhibition, as well as spatial segregation between types. We discuss the use of Monte Carlo randomization tests to assess significant pairwise interactions within multivariate spatial point patterns. Then, we introduce multivariate point process models that are commonly used to capture dependencies within and between two or mores spatial point patterns. We focus on two flexible classes of models for multivariate point process – Gibbs processes and log Gaussian Cox processes - and highlight important differences between them. In addition, we showcase recent and interesting applications of each type of model. For multivariate log Gaussian Cox processes, we give a detailed discussion regarding the rich model inference available. We also provide a detailed investigation of such modeling through an application to multiple tree species in Duke Forest, North Carolina.

5.3.2. Summary statistics and statistical tests for multivariate point patterns

Multivariate spatial point patterns can be assessed using extensions of the univariate summary measures introduced in Section 2.2. The second-order structure of a multivariate point pattern can be estimated through the K function (Ripley 1981, Lotwick 1982) and provides useful measures of multivariate point patterns to test hypotheses of spatial interaction. The bivariate K function $K_{m,m'}(d)$ for type m, $m' \in \mathcal{M}$ is $K_{m,m'}(d) = (\lambda_{m'})^{-1}\mathrm{E}(\#\text{number of points of type } m \text{ within distance } d \text{ of an arbitrary point of type } m')$. Analogous to the univariate statistic, under the null hypothesis of independent point processes of types m and m',

$K_{m,m'}(d) = \pi d^2$ for all distance d. For spatial point patterns $\mathcal{S}_m$ and $\mathcal{S}_{m'}$, which denote the collection of points with marks m and m', estimates of $K_{m,m'}(d) > \pi d^2$ identify positive dependence or clustering between the two types at distance d, whereas values $K_{m,m'}(d) < \pi d^2$ indicate negative dependence or repulsion.

Nearest neighbor methods using G functions can also be used to assess interactions between multivariate processes. Recall the univariate nearest neighbor distribution $G_m(d) = P(\text{nearest point is distance} \leq d)$ for spatial point process $\mathcal{S}_m$. For a bivariate point process with types m and m', $G_{m,m'}(d) = P(\text{nearest point} \leq d)$ for all points in $\mathcal{S}_m \cup \mathcal{S}_{m'}$. For spatial point patterns $\mathcal{S}_m$ and $\mathcal{S}_{m'}$ that are independent, $G_{m,m'}(d) = G_m(d)G_{m'}(d)$. The statistic $T_{m,m'}(d) = log(G_{m,m'}(d)) - log(G_m(d)) - log(G_{m'}(d))$ can be used to assess interactions between the point patterns [129]. Large values of $T_{m,m'}(d)$ denote attraction between the points of each type whereas negative values indicate repulsion or inhibition. Multivariate spatial point processes can be assessed analogously by computing the nearest neighbor distribution using the superposition of point patterns, $\cup_{m \in \mathcal{M}} \mathcal{S}_m$.

Hypotheses about multivariate spatial point processes can be assessed based on these statistics using Monte Carlo randomization tests [29, 168, 169]. Within an ecological framework, [5] discusses a series of Monte Carlo randomization tests for bivariate point patterns. The tests can assess independence between species, where alternative hypotheses could suggest clustering or inhibition at various distances.

A Monte Carlo test of spatial variation was developed by [108] and generalized to multivariate point patterns by [53]. In their approach, a multivariate inhomogeneous Poisson point process is assumed and nonparametric estimation of ratios of component-wise intensities is used to detect spatial segregation. Kernel regression estimators are defined for the conditional probability surface, $p_k(\mathbf{s})$, which is the conditional probability that an event at location $\mathbf{s}$ is of type k. Under the null model, the relative risk ratios of the component-wise intensities, which can be written as the ratio of conditional probabilities $p_k(\mathbf{s})/p_j(\mathbf{s})$, would be constant across the domain. Regions of increased or decreased relative risk estimates would indicate spatial segregation of one more type. Monte Carlo sampling is used to test the null hypothesis of no spatial variation in relative risk surfaces. Randomizations for the test are obtained by keeping the spatial locations fixed and randomly assigning marks, thus preserving the number of each type. Diggle et al (2005) apply this method to types of bovine tuberculosis in cattle herds in Michigan and find strong spatial segregation among types.

In an ecological context, measures of multivariate point patterns can also be used to assess species diversity. Using the formulation of $K(d)$ and $G(d)$ functions, [183] introduced functions $\alpha(d)$ and $\beta(d)$ to test for spatial variation of species diversity. The $\beta(d)$ function captures the conditional probability that two points belong to different types given that they are distance d apart. $\alpha(d)$ on the other hand, defines the probability that two points at distance less than or equal to d are of different types. These functions, which are aptly named to quantify α- and β-diversity in plant species, are extensions of the Simpson index, and are referred to as distance-dependent Simpson indices.

5.3.3. Multivariate Gibbs processes

Multivariate Gibbs processes, like their univariate counterparts, can be used to capture interactions, such as clustering or inhibition, within spatial point patterns. Multivariate Gibbs processes directly model the interaction between the points

such that, for an inhibition process, the observed location of an event decreases the likelihood of another event from occurring nearby. Importantly, the strength of these interactions can be specified within the multivariate Gibbs processes to vary within or between type (e.g., within species or between species).

Gibbs models assume symmetric interactions such that the effect of point $\mathbf{s}$ on the location of point $\mathbf{s}'$ is equal to the effect of point $\mathbf{s}'$ on the location of point $\mathbf{s}$. Gibbs processes with pairwise interactions are a common model choice for interacting point patterns. Let $(\mathbf{s}, m)$ denote an observed spatial location with mark m, where $\mathbf{s} \in D$ and $m \in \mathcal{M}$. Further, let $\mathcal{S}_m$ denote the spatial point pattern with mark m.

The density of a Gibbs process with pairwise interactions can be written as

$$f(\mathbf{s}, m) = \alpha \prod_{i=1}^{N} g_{m_i}(\mathbf{s}_i) \prod_{i<j} h_{m_i,m_j}(\mathbf{s}_i, \mathbf{s}_j)$$

where $g_m(\mathbf{s})$ for $m \in \mathcal{M}$ are functions capturing the first order trend of spatial points for mark m, and $h_{m,m'}(\mathbf{s}, \mathbf{s}')$ for $m, m' \in \mathcal{M}$ are the functions capturing the interactions between the pair of points for marks m and m'. Additionally, α is a scaling such that $f(\mathbf{s}, m)$ is a density. The interaction functions, $h_{m,m'}(\mathbf{s}, \mathbf{s}')$ must be symmetric such that $h_{m,m'}(\mathbf{s}, \mathbf{s}') = h_{m,m'}(\mathbf{s}', \mathbf{s})$. Various forms for $h_{m,m'}(\mathbf{s}, \mathbf{s}')$ have been proposed in the literature to capture the interaction between pairs of points. For example, multivariate hard-core processes might assume

$$h_{m,m'}(\mathbf{s}, \mathbf{s}') = \begin{cases} 1 & ||\mathbf{s} - \mathbf{s}'|| > d_{m,m'} \\ 0 & ||\mathbf{s} - \mathbf{s}'|| \leq d_{m,m'} \end{cases}$$

where $d_{m,m'}$ is the minimum allowable distance between points with mark m and m'. A more relaxed inhibition process between points with mark m and m' follows the multivariate Strauss process where

$$h_{m,m'}(\mathbf{s}, \mathbf{s}') = \begin{cases} 1 & ||\mathbf{s} - \mathbf{s}'|| > d_{m,m'} \\ \beta_{m,m'} & ||\mathbf{s} - \mathbf{s}'|| \leq d_{m,m'}. \end{cases}$$

Here, $d_{m,m'}$ are interaction radii and $\beta_{m,m'} \geq 0$ are interaction parameters such that smaller values of $\beta_{m,m'}$ exhibit stronger inhibition between components m and m'. Note that a bivariate Strauss process with $\beta_{1,2} = 1$ results in two independent Strauss processes. [55] show that even when all $\beta_{m,m'} < 1$, the marginal behavior of each component process could still exhibit spatial aggregation.

[99] propose a modification of the multivariate Gibbs process by, instead, applying univariate Gibbs models to build the multivariate process hierarchically. The benefit of the hierarchical approach is that it enables the assessment of the asymmetric strength and range of interaction within and between each type (e.g., species). Here, the hierarchical model is built based on scientific reasoning. For example, each type would be modeled univariately with a nonstationary process and adding a hierarchical structure to these processes would induce asymmetric dependence. That is, higher levels are driving heterogeneity in the lower levels but the lower levels do not affect the higher levels.

The hierarchical approach for building a multivariate Gibbs process begins by defining the density of the highest level Gibbs process. Let m_1 denote the mark for the highest level process. Then,

$$f(\mathbf{s}, m_1) = \alpha_{m_1} \prod_{\mathbf{s_i} \in \mathcal{S}_1} g_{m_1}(\mathbf{s}_i) \prod_{\substack{\mathbf{s}_i, \mathbf{s}_j \in \mathcal{S}_1 \\ i<j}} h_{m_1}(\mathbf{s}_i, \mathbf{s}_j)$$

where $g_{m_1}(\mathbf{s}_i)$ and $h_{m_1}(\mathbf{s}_i, \mathbf{s}_j)$ are analogous to above and α_{m_1} is the scaling factor for the univariate Gibbs process. Now, given the point pattern, $\mathcal{S}_1$, at the top of the hierarchy, the density for the second level point pattern can be written

$$f((\mathbf{s}, m_2)|(\mathcal{S}_1, m_1)) = \alpha_{m_1|m_2} \prod_{\mathbf{s_i} \in \mathcal{S}_2} g_{m_2}(\mathbf{s}_i) \prod_{\substack{\mathbf{s}_i, \mathbf{s}_j \in \mathcal{S}_2 \\ i<j}} h_{m_2}(\mathbf{s}_i, \mathbf{s}_j) \prod_{\substack{\mathbf{s}_i \in \mathcal{S}_2 \\ \mathbf{s}_j \in \mathcal{S}_1}} h_{m_2|m_1}(\mathbf{s}_i, \mathbf{s}_j).$$

Now, $h_{m_2}(\mathbf{s}, \mathbf{s}')$ controls the interaction between points of with mark m_2 and $h_{m_2|m_1}(\mathbf{s}, \mathbf{s}')$ captures the directed interaction of point at location $\mathbf{s}'$ with mark m_1 on a point at location $\mathbf{s}$ with mark m_2. Conditional intensities can continually be added in this fashion for general M marks. By computing the joint distribution, $f((\mathcal{S}_1, m_1), (\mathcal{S}_2, m_2)) = f(\mathcal{S}_1, m_1) f((\mathbf{s}, m_2)|(\mathcal{S}_1, m_1))$, we can easily see the difference between the hierarchical approach at the multivariate approach above. The densities are equivalent up to the scaling factor, which is now $\alpha_{m_1}\alpha_{m_2|m_1}$ where $\alpha_{m_2|m_1}$ depends on the observed locations, $\mathcal{S}_1$. It is important to note that the hierarchical model specification will vary with the ordering of the marks and should therefore be scientifically founded.

Estimation of both forms of multivariate Gibbs process models can be done using maximum likelihood methods. Traditional maximum likelihood requires computationally challenging approximations of the scaling factors. Maximum pseudolikelihood methods for Gibbs processes [88, 105] avoid the difficult approximation since the scaling factors cancel out of the ratio of density functions. The maximum pseudolikelihood method has been found to overestimate the interaction process, however, giving preference to the more cumbersome maximum likelihood approach [99].

[89] consider the asymmetric relationship for different tree species where the size of the tree determines the hierarchical level; largest trees are not influenced by smaller trees but small trees are influence by larger trees. Using multivariate Gibbs point processes with hierarchical interactions, [89] quantify the strength of competition between trees of different size classes. They find that the influence of large trees on small neighboring trees is stronger than large trees on other large trees at the same distance.

[103] employ a hierarchical multivariate spatial point process model to capture varying spatial inhomogeneity, clustering, and inhibition patterns of plant species. In particular, they specify the multivariate point pattern of "seeder" plant species conditionally given the observed point pattern of a collection of "resprouter" plant species. This is a simplification of the hierarchical interaction model defined by [99] where here, the resprouter plant species point patterns are treated as fixed and known and only the seeder plant species point patterns are modeled stochastically. The log-intensity function $\lambda_j(\mathbf{s})$ for each seeder species j is specified using a linear combination of K smooth interaction functions, each based on the distance between $\mathbf{s}$ and the observed spatial point pattern of resprouter plant species k, denoted $\mathcal{S}_k$. Each spatial point pattern of seeder plant species is assumed conditionally independent given the observed point patterns of the K resprouter plant species.

The hierarchical Bayesian framework allows for ecological information regarding species interaction radii distance to be included in the model through the interaction functions in order to assess the interaction strengths between each seeder and resprouter plant pair. The model is assessed using residual plots based on quadrat counts as well as estimated L functions. Their results indicate varying dependence structures between the seeder and resprouter plant species along with possible clustering within and between seeder species.

5.3.4. Multivariate Log Gaussian Cox Process

The multivariate log Gaussian Cox process [101, 142] provides an alternative approach for modeling multivariate spatial point patterns. Under the multivariate Gibbs process, the interactions between points are captured by the functions $h_{m,m'}(\mathbf{s}, \mathbf{s}')$. Under the multivariate LGCP, dependence is built between the intensity functions $\lambda_m(\mathbf{s})$ and $\lambda_{m'}(\mathbf{s})$. Like the univariate LGCP, realizations of the multivariate LGCP are conditionally independent given the multivariate intensity $\boldsymbol{\lambda}(\mathbf{s}) = (\lambda_1(\mathbf{s}), \ldots, \lambda_M(\mathbf{s}))^T$.

The multivariate LGCP was first introduced by [141] where processes $\mathcal{S}_1, 2, \ldots,$ $\mathcal{S}_M$, are modeled with intensity $\lambda_m(\mathbf{s}) = \exp(z_m(\mathbf{s}))$ for $m = 1, 2, \ldots, M$ where $\mathbf{z}(\mathbf{s}) = (z_1(\mathbf{s}), \ldots, z_M(\mathbf{s}))^T$ for $\mathbf{s} \in D$ is a multivariate Gaussian process with mean $\boldsymbol{\mu} = (\mu_1, 2, \ldots, \mu_M)'$ and covariance functions $c_{m,m'}(\mathbf{s}, \mathbf{s}') = cov(z_m(\mathbf{s}), z_{m'}(\mathbf{s}'))$. The covariance functions $c_{m,m'}(\mathbf{s}, \mathbf{s}')$ of the multivariate Gaussian process must specify a valid cross-covariance matrix. Conditional on $\mathbf{z}$, the processes, $\mathcal{S}_m, m = 1, 2, ..., M$ are independent Poisson processes with intensities $\lambda_m(\mathbf{s}), m = 1, 2, ..., M$.

[141] suggest affine transformations as an easy alternative way to build dependence between $z_m(\mathbf{s})$ and $z_{m'}(\mathbf{s})$. Here, letting $V_k(\mathbf{s})$ for $k = 1, 2, \ldots, K$, denote K independent univariate Gaussian processes with mean 0, variance 1, and valid correlation function $r_m(\mathbf{s}, \mathbf{s}')$, $z_m(\mathbf{s})$ is then defined as a linear combination of the $V_k(\mathbf{s})$. That is, $z_m(\mathbf{s}) = \sum_{k=1}^{K} A_{mk} V_k(\mathbf{s}) + \mu_m$ where μ_m is a type-specific adjustment, which could vary spatially using local regressors. Dependence between $z_m(\mathbf{s})$ and $z_{m'}(\mathbf{s})$, which implies dependence between $\lambda_m(\mathbf{s})$ and $\lambda_{m'}(\mathbf{s})$, is captured by the shared Gaussian processes, the $V_k(\mathbf{s})$'s. The covariance between $z_m(\mathbf{s})$ and $z_{m'}(\mathbf{s}')$ is

$$cov(z_m(\mathbf{s}), z_{m'}(\mathbf{s}')) = \sum_{k=1}^{K} A_{mk} A_{m'k} r_k(\mathbf{s}, \mathbf{s}')$$

for $m, m' \in \mathcal{M}$ and $\mathbf{s}, \mathbf{s}' \in D$. This offers a very flexible class of models for capturing dependence between intensity functions of multivariate spatial point processes. [141] show simulation results under different dependence structures for bivariate spatio-temporal processes with $K = 1$.

Coregionalization provides a convenient modification of this approach for building dependence through linear combinations of independent Gaussian processes [74]. Most applications of coregionalization are geostatistical, as the approach was intended to model measurements that co-vary jointly over a region. For multivariate spatial point patterns, we can use coregionalization at the process level in building the multivariate intensity surface. Letting $V_k(\mathbf{s})$ for $k = 1, 2, \ldots, M$ again denote independent Gaussian processes, we define the full rank $M \times M$ matrix $\mathbf{A}$, which we can assume to be lower triangular. Then, $\mathbf{z}(\mathbf{s}) = \boldsymbol{\mu} + \mathbf{A}\mathbf{V}(\mathbf{s})$. Here, $\boldsymbol{\mu} = (\mu_1, 2, \ldots, \mu_M)'$ is the vector of type-specific means, and $\mathbf{V}(\mathbf{s}) = (V_1(\mathbf{s}), \ldots, V_M(\mathbf{s}))^T$ is the vector of independent GPs at location $\mathbf{s}$. The *local* covariance matrix for the multivariate process is $\Sigma = \mathbf{A}\mathbf{A}^T$. The diagonal elements $A_{kk} \geq 0$ whereas the lower off diagonal elements $A_{kk'} \in \mathbb{R}^1$ where $k > k'$. For a bivariate spatial point process, $\mathbf{A}$ is a 2×2 dimensional matrix where $A_{21} < 0$ implies negative dependence between the two processes and $A_{21} > 0$ denotes positive dependence. In higher dimensions, similar positive and negative dependence can be deduced from either linear combinations of elements of $\mathbf{A}$ or through isolating some pairwise relationships by setting other lower off-diagonal elements of $\mathbf{A}$ to 0. For example, with a 3×3 dimensional coregionalization matrix, setting $A_{32} = 0$ means the sign of A_{31} indicates the direction of dependence between $z_1(\mathbf{s})$ and $z_3(\mathbf{s})$.

[207] use multivariate log Gaussian Cox processes to model multispecies point patterns of tree locations. They decompose the latent multivariate Gaussian processes into shared components and species-specific components to identify groups of species with similar patterns of dependence on the latent process. [32] extend the multivariate LGCP models to space-time models to capture both spatial and temporal heterogeneity in a time series of spatial point patterns of different weed species. The temporal dependence is modeled using a spatial birth process indexed in time and defined conditionally on a spatial multivariate Gaussian process. Dependence between weed type and across space is captured by the multivariate Gaussian processes defined using the affine transformations of [141]. The birth processes are assumed conditionally independent given the Gaussian processes.

Multivariate log Gaussian Cox process models can be fitted in the Bayesian framework analogously to the univariate LGCPs using the methods discussed in Section 4.2. Let $\mathcal{S} = \cup_{m=1}^{M} \mathcal{S}_m$ denote the collection of observed spatial locations from all species in the domain, D. Given the multivariate intensity surface, $\boldsymbol{\lambda}(\mathbf{s})$, the likelihood function of the multivariate LGCP model is the product of K independent nonhomogeneous Poisson process likelihoods. That is, the likelihood function with instantaneous intensity $\boldsymbol{\lambda}(\mathbf{s})$ and observations $\mathcal{S}$ is given by

$$\prod_{m=1}^{M} \prod_{\mathbf{s}_i \in \mathcal{S}_m} \lambda_m(\mathbf{s}_i) \exp^{-\int_D \lambda_m(\mathbf{s}) d\mathbf{s}}.$$

When fitting the model, the integrals are stochastic, requiring realizations of the multivariate process at representative points and approximate numerical integration to evaluate $\int_D \lambda_m(\mathbf{s}) d\mathbf{s}$.

Markov chain Monte Carlo can be used to obtain posterior samples of the model parameters. Note that sampling from the multivariate Gaussian process at a collection of representative points can be computationally cumbersome as both the number of locations and number of processes increase. The elliptical slice sampler provides one approach for increase computational efficiency. In fitting the model using coregionalization, we work in the parameter space of the independent Gaussian processes, $V_k(\mathbf{s})$, given $\boldsymbol{\mu}$ and $\mathbf{A}$. Simulating multivariate point pattern realizations to carry out full Bayesian inference is also conducted in this parameter space. Specifically, to simulate a realization of the multivariate point pattern, a sample from the posterior distribution of $V_k(\mathbf{s})$, $k = 1, 2, \ldots, M$, $\boldsymbol{\mu}$ and $\mathbf{A}$ yields a posterior sample of $\boldsymbol{\lambda}(\mathbf{s})$, $\mathbf{s} \in D$. Conditional on the multivariate intensity $\boldsymbol{\lambda}(\mathbf{s})$, realizations of the point patterns for each mark are independent. That is, for each m, we can obtain a realization of the spatial point pattern with intensity $\lambda_m(\mathbf{s})$ using the univariate simulation approach outlined in Section 4.1.3. The superposition of the M spatial point patterns yields a realization of the multivariate spatial point pattern from the multivariate LGCP model. A collection of realizations of the multivariate spatial point patterns using samples from the posterior distribution of the parameters enables full Bayesian inference.

The multivariate LGCP model provides rich posterior inference, which can be both parametric and predictive. Parametric inference converts posterior samples of the multivariate intensity, $\lambda_m(\mathbf{s})$, to conditional probabilities. For example, for an event occurring at location $\mathbf{s}$, $p(m|\mathbf{s})$ is the probability that it is of type m. This conditional probability can be computed as

$$p(m|\mathbf{s}) = \frac{\lambda_m(\mathbf{s})}{\sum_{m=1}^{M} \lambda_m(\mathbf{s})}$$

for type m at location $\mathbf{s}$. Analogous conditioning can be done for two events. Given that two events occurred, one at location $\mathbf{s}$ and another at $\mathbf{s}'$, the probability that the event at $\mathbf{s}$ was type m and the event at $\mathbf{s}'$ was type m' is denoted by $p(m, m'|\mathbf{s}, \mathbf{s}')$. This conditional probability is computed as

$$p(m, m'|\mathbf{s}, \mathbf{s}') = \frac{\lambda_m(\mathbf{s})\lambda_{m'}(\mathbf{s}')}{\sum_{m=1}^{M} \lambda_m(\mathbf{s}) \sum_{m=1}^{M} \lambda_m(\mathbf{s}')}.$$

With posterior samples of each $\lambda_m(\mathbf{s})$, we can obtain full posterior distributions for these conditional probabilities.

Posterior prediction of the number of events of each type in a particular set $B \subset D$ is also of interest. We can obtain the samples from the posterior predictive distribution of the number of events of type m in B, denoted $N_m(B)$, using composition sampling [16, 33]. The posterior distribution of $N_m(B)$ given the data, $\mathcal{S}$ can be written

$$[N_m(B)|\mathcal{S}] = \int_{\boldsymbol{\theta}} [N_m(B)|\boldsymbol{\theta}][\boldsymbol{\theta}|\mathcal{S}]d\boldsymbol{\theta}$$

where $\boldsymbol{\theta}$ represents all model parameters and latent process variables. The distribution $[N_m(B)|\boldsymbol{\theta}]$ follows a Poisson distribution with mean $\lambda_m(B) = \int_B \lambda_m(\mathbf{s})d\mathbf{s}$ and $[\boldsymbol{\theta}|\mathcal{S}]$ denotes the posterior distribution of the model parameters and latent process variables given the data. We can approximate the integral $\int_B \lambda_m(\mathbf{s})d\mathbf{s}$ using numerical integration. Samples from the posterior distribution of the total number of events in B, written $N(B) = \sum_{m=1}^{M} N_m(B)$, can be obtained from each posterior sample of $N_m(B)$. In the multivariate setting, we can also look at the joint posterior predictive distribution of $[N_m(B), N_{m'}(B')]$. We can obtain the joint posterior predictive distribution of the number of individuals of species m in the set B and the number of individuals of species m' in the set B'. Here, B and B' can be any sets in D, e.g., $B = B'$ or $B \cap B' = \emptyset$. Using the posterior distribution of the parameters given the data, we obtain samples from

$$\begin{aligned}[N_m(B), N_{m'}(B')|\mathcal{S}] &= \int_{\boldsymbol{\theta}} [N_m(B), N_{m'}(B')|\boldsymbol{\theta}][\boldsymbol{\theta}|\mathcal{S}]d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}} [N_m(B)|\boldsymbol{\theta}][N_{m'}(B')|\boldsymbol{\theta}][\boldsymbol{\theta}|\mathcal{S}]d\boldsymbol{\theta}\end{aligned}$$

since $N_m(B)$ and $N_{m'}(B')$ are conditionally independent Poisson random variables given $\boldsymbol{\theta}$.

In practice, samples from the posterior distributions of $N(B)$, $N_m(B)$, and $[N_m(B), N_{m'}(B')]$ for arbitrary $B, B' \in D$ and $m, m' \in \mathcal{M}$ can be obtained by simulating posterior spatial point patterns for the multivariate point process under the model. For each posterior spatial point pattern realization, the posterior sample of $N_m(B)$ is the number of events type m in B, and the sample of $N(B)$ the sum of $N_m(B)$ over all m. Samples from the joint posterior predictive distribution of $[N_m(B), N_{m'}(B')]$ are obtained from each realization as the number of type m in B and number of type m' in B'. Posterior simulated realizations of the multivariate spatial point pattern can also be used to learn about conditional probabilities associated with particular events, e.g., $P(N_m(B) = 0|N_{m'}(B') > 0)$. The conditional probabilities can be calculated from the joint distribution over the marginal distribution based on the posterior samples of the entire point pattern. Such conditional probabilities inform about the probability of absence of one type of event in a particular set given presence of another type of event in the same or perhaps, adjacent set. In certain applications these posterior distributions can be useful with regard to co-occurrence of events of different types.

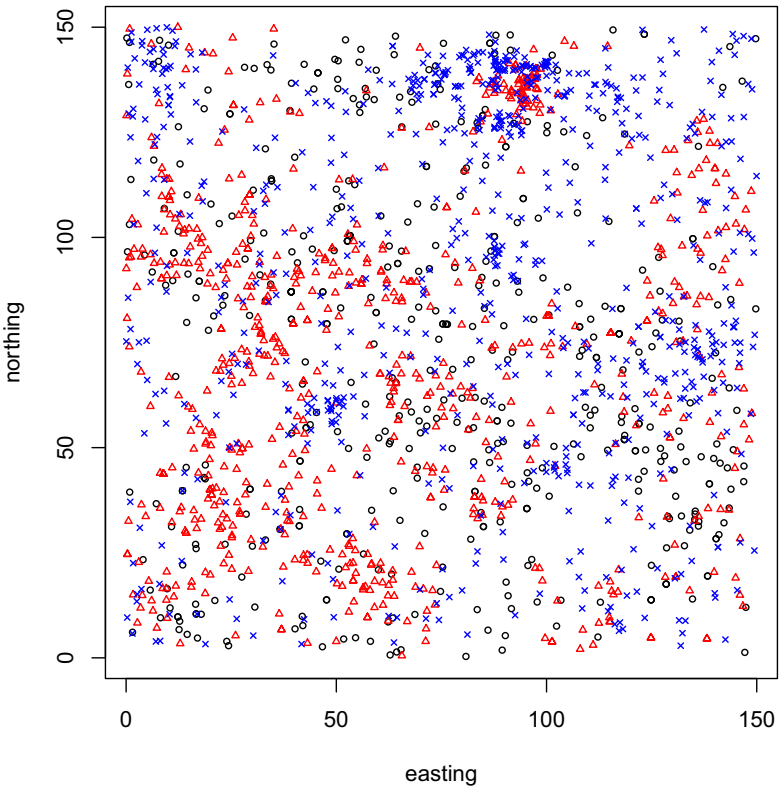

FIG 5.1. *Spatial locations of tree species* Acer rubrum *(red maple, ∘),* Fraxinus americana *(white ash, △), and* Liquidambar styraciflua *(American sweetgum ×)*

5.3.5. Application: Multivariate spatial point pattern of tree species in Duke Forest

We apply the multivariate LGCP model to tree data collected at Duke Forest in Durham, North Carolina. The Duke hardwood plot occupies mixed hardwood and pine stands [41]. Specifically, the data analyzed here include tree locations for three tree species, *Acer rubrum* (red maple), *Fraxinus americana* (white ash), and *Liquidambar styraciflua* (American sweetgum) in a 150m × 150m forest stand. The spatial locations of these trees are shown in Figure 5.1. In this region, there are 481 red maples, 717 ash, and 766 sweetgum trees. The elevation gradient for this forest plot is shown in Figure 5.2 highlighting generally lower elevation in the west and higher elevation in the east. Red maples tend to be spread out across the plot, whereas ash and sweetgum appear more clustered. Ash trees appear to be more prevalent in the west at lower elevations whereas sweetgum trees are more prevalent in the east at higher elevation.

We model the locations of the three species, red maple, ash, and sweetgum, using a multivariate log-Gaussian Cox process. The joint intensity is written $\boldsymbol{\lambda}(\mathbf{s}) = (\lambda_1(\mathbf{s}), \lambda_2(\mathbf{s}), \lambda_3(\mathbf{s}))^T$ where subscripts 1, 2, and 3 denote red maple, ash, and sweetgum, respectively. The joint intensity function is written

$$\log\boldsymbol{\lambda}(\mathbf{s}) = \boldsymbol{\gamma}\mathbf{x}(\mathbf{s}) + \mathbf{z}(\mathbf{s}) \tag{5.10}$$

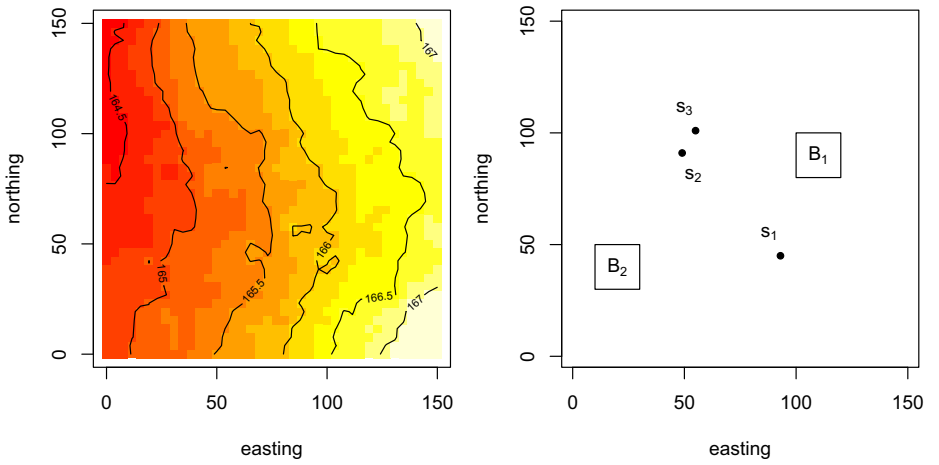

FIG 5.2. *(left) The elevation gradient for the Duke Forest plot, where the west is at lower elevation and east is at higher elevation. Elevation is reported in meters (m). (right) Three locations,* $\mathbf{s}_1$*,* $\mathbf{s}_2$*, and* $\mathbf{s}_3$*, as well as two subregions,* B_1*, and* B_2*, used for illustrative model inference.*

where $\boldsymbol{\gamma}$ is a matrix of coefficients, $\mathbf{x}(\mathbf{s})$ is a vector of location-specific covariates, and $\mathbf{z}(\mathbf{s})$ is a multivariate Gaussian process.

In fitting the model to the locations in the Duke Forest plot, $\mathbf{x}(\mathbf{s})$ includes elevation for location $\mathbf{s}$ as well as a intercept term. The coefficient $\boldsymbol{\gamma}$ contains species-specific intercept and slope coefficients to capture the varying effect of elevation across species. As a result $\boldsymbol{\gamma}$ is 3×2. The multivariate Gaussian process, $\mathbf{z}(\mathbf{s})$, models cross-species dependence as well as enabling local adjustment for remaining heterogeneity in the log intensity beyond what is explained by elevation.

We employ the linear model of coregionalization [74] to capture cross-species dependence. That is, $\mathbf{z}(\mathbf{s}) = \mathbf{A}\mathbf{V}(\mathbf{s})$, where $\mathbf{A}$ is a lower triangular 3×3 matrix and each $V_k(\mathbf{s})$, for $k = 1, 2, 3$, is a realization from an independent Gaussian process at a set of representative points. For the 150m $\times$ 150m plot, we employ 225 representative points on a grid of 10km resolution. The spatial covariance for $V_k(\mathbf{s})$ is defined by an exponential correlation function, $\text{cov}(V_k(\mathbf{s}), V_k(\mathbf{s}')) = \exp^{-||\mathbf{s}-\mathbf{s}'||/\phi_k}$ with range parameter ϕ_k.

Noninformative independent mean zero normal prior distributions were assigned to each of the coefficients in $\boldsymbol{\gamma}$. The diagonal elements of $\mathbf{A}$ were assigned diffuse, independent inverse-Gamma distributions with shape and scale equal 2. The lower off-diagonal elements of $\mathbf{A}$ were assigned mean zero normal distributions with variance 100. Lastly, the decay parameters, ϕ_k, were assigned Uniform distributions with lower and upper endpoints of 5 and 40. For the independent Gaussian processes, $V_k(\mathbf{s})$, this assumed an effective range, $3\phi_k$, to be greater than the resolution of the representative points and less than half the max distance of the domain.

Model inference was obtained in a Bayesian framework using MCMC and Metropolis-Hastings algorithms. Draws from the posterior distribution of the independent Gaussian processes at the representative points were obtained using elliptical slice sampling [144]. The model was fitted for 500,000 iterations, with the first 20% were disregarded as burn in. The chains were thinned to every 50th iteration to reduce dependence, and the remaining samples were used for posterior inference.

TABLE 5.2
Posterior mean and 90% CI for the multivariate LGCP model fitted to three tree species in Duke Forest. Here, γ_{m0} and γ_{m1} denote the intercept and elevation coefficient for species m and Σ denotes the local covariance matrix where Σ_{ij} is the (i,j)th element of the matrix $\Sigma = \mathbf{A}\mathbf{A}^T$.

	mean (90% CI)
γ_{10}	0.506 (0.091, 0.861)
γ_{11}	-0.254 (-0.639, 0.090)
γ_{20}	0.558 (-0.134, 1.092)
γ_{21}	-0.645 (-1.173 -0.146)
γ_{30}	0.929 (0.466, 1.398)
γ_{31}	0.333 (-0.107, 0.791)
Σ_{11}	0.468 (0.260, 0.748)
Σ_{22}	1.287 (0.798, 1.927)
Σ_{33}	0.886 (0.580, 1.302)
Σ_{21}	0.138 (-0.056, 0.355)
Σ_{31}	-0.047, (-0.232, 0.131)
Σ_{32}	0.006 (-0.204, 0.230)
ϕ_1	26.14 (13.26, 38.41)
ϕ_2	21.83 (11.74, 35.03)
ϕ_3	23.97 (14.03, 36.57)

Posterior means and 90% credible intervals are given for the model parameters in Table 5.2. The species specific-coefficients indicate that, in general, the intensity of red maple and ash decreases with elevation while it increases with elevation for sweetgum. The elements of the 3×3 matrix, $\mathbf{A}\mathbf{A}^T$ capture the *local* variance-covariance of the multivariate Gaussian processes. Specifically, the off-diagonal elements highlight dependence between the species. Red maple and ash have moderate positive dependence whereas sweetgum appear not correlated with either red maple or ash. The decay parameters are similar for each of the Gaussian processes. Under coregionalization, the posterior distribution of the effective range for each species can be computed from the posterior distributions of ϕ_1, ϕ_2, ϕ_3, and $\mathbf{A}$. See [16] for the functional forms for this computation.

The multivariate posterior mean intensity surface $\boldsymbol{\lambda}(\mathbf{s})$ is shown univariately in Figure 5.3 for each species, red maple, ash, and sweetgum. Red maples appear to have higher intensity in the central part of the region, whereas the intensity for ash and sweetgum favor the west and east part of the plot, respectively. Some similarity in patterns exists between red maple and ash in the central and western part of the plot. Ash and sweetgum both have high mean intensity in the north central region of the plot.

We also investigate the joint posterior probabilities of $p(m(\mathbf{s}), m'(\mathbf{s}')|\mathcal{S})$ for pairs of locations $\mathbf{s}$ and $\mathbf{s}'$ where m and m' take the values 1, 2, or 3. Three locations, $\mathbf{s}_1$, $\mathbf{s}_2$, and $\mathbf{s}_3$ shown in Figure 5.2 were randomly chosen such that $\mathbf{s}_1$ and $\mathbf{s}_2$ were far apart and $\mathbf{s}_2$ and $\mathbf{s}_3$ were close. The posterior mean probabilities for all pairs of species and pairs of locations $(\mathbf{s}_1, \mathbf{s}_2)$, and $(\mathbf{s}_2, \mathbf{s}_3)$ are given in Tables 5.3 and 5.4, respectively. For locations $(\mathbf{s}_1, \mathbf{s}_2)$, the highest probability is for a red maple at $\mathbf{s}_1$ and ash at $\mathbf{s}_2$. Marginally, red maple and sweetgum have very similarly large marginal probabilities at $\mathbf{s}_1$ while ash has the highest marginal probability at $\mathbf{s}_2$. When comparing nearby locations $\mathbf{s}_2$ and $\mathbf{s}_3$, the highest joint posterior mean probability is again ash and red maple, with ash at $\mathbf{s}_2$ and red maple at $\mathbf{s}_3$. Marginally, red maple has the highest probability at $\mathbf{s}_3$.

Additional model inference includes obtaining posterior distributions of $N_m(B)$, the number of species m in a subregion B. For the two regions, B_1, and B_2 shown in Figure 5.2, we compute the posterior predictive distributions for each species.

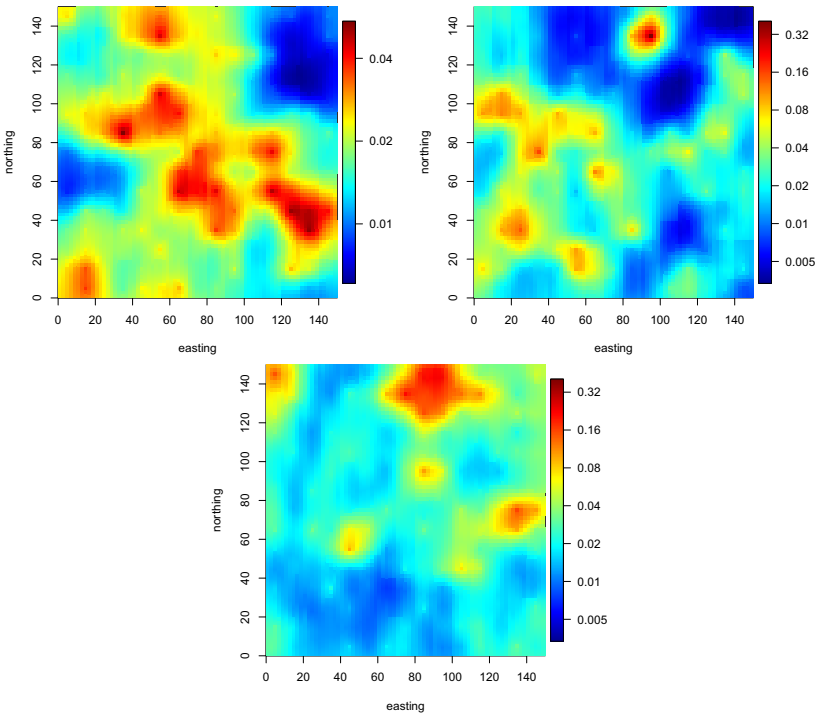

FIG 5.3.

TABLE 5.3
Posterior mean probabilities of the joint distribution $p(m(\mathbf{s}_1), m'(\mathbf{s}_2)|\mathcal{S})$ for the three species, red maple, ash, and sweetgum. The locations $\mathbf{s}_1$ and $\mathbf{s}_2$ are shown in Figure 5.2.

		$\mathbf{s}_2$		
		red maple	ash	sweetgum
$\mathbf{s}_1$	red maple	0.10	0.24	0.05
	ash	0.07	0.16	0.03
	sweetgum	0.09	0.22	0.04

TABLE 5.4
Posterior mean probabilities of the joint distribution $p(m(\mathbf{s}_2), m'(\mathbf{s}_3)|\mathcal{S})$ for the three species, red maple, ash, and sweetgum.

		$\mathbf{s}_3$		
		red maple	ash	sweetgum
$\mathbf{s}_2$	red maple	0.12	0.08	0.06
	ash	0.27	0.20	0.14
	sweetgum	0.05	0.04	0.03

TABLE 5.5
Posterior distributions of the number of each species in subregions B_1 and B_2.

	red maple	ash	sweetgum
B_1	7.3 (3.0, 12.0)	5.8 (2.0, 11.0)	8.0 (3.0, 14.0)
B_2	7.6 (3.0, 13.0)	37.8 (26.0, 49.0)	5.8 (2.0,11.0)

Each of the two regions are 20m $\times$ 20m, where B_1 is centered at approximately 2m higher in elevation than B_2. The posterior distribution of the number of red maples in each region are very similar, while the number of sweetgum is slightly higher in B_1 than B_2. The most notable difference between these two subregions is for ash trees. The expected number of ash trees is vastly greater in B_2 than B_1. This agrees with the posterior distribution of γ_{21}, which indicated a significant negative relationship with elevation for ash.