The Annals of Applied Statistics
2016, Vol. 10, No. 2, 1041–1062
DOI: 10.1214/16-AOAS929

LATENT SPATIAL MODELS AND SAMPLING DESIGN FOR LANDSCAPE GENETICS[1]

BY EPHRAIM M. HANKS[*], MEVIN B. HOOTEN[†,‡], STEVEN T. KNICK[§], SARA J. OYLER-MCCANCE[¶], JENNIFER A. FIKE[¶], TODD B. CROSS[‖,**] AND MICHAEL K. SCHWARTZ[‖]

Pennsylvania State University[*], U.S. Geological Survey, Colorado Cooperative Fish and Wildlife Research Unit[†], Colorado State University[‡], U.S. Geological Survey, Forest and Rangeland Ecosystem Science Center[§], U.S. Geological Survey, Fort Collins Science Center[¶], U.S. Forest Service, Rocky Mountain Research Station[‖] and University of Montana[**]

We propose a spatially-explicit approach for modeling genetic variation across space and illustrate how this approach can be used to optimize spatial prediction and sampling design for landscape genetic data. We propose a multinomial data model for categorical microsatellite allele data commonly used in landscape genetic studies, and introduce a latent spatial random effect to allow for spatial correlation between genetic observations. We illustrate how modern dimension reduction approaches to spatial statistics can allow for efficient computation in landscape genetic statistical models covering large spatial domains. We apply our approach to propose a retrospective spatial sampling design for greater sage-grouse (*Centrocercus urophasianus*) population genetics in the western United States.

1. Introduction. Landscape connectivity is "the degree to which the landscape facilitates or impedes movement among resource patches" [Taylor et al. (1993)]. Connectivity among subpopulations or habitat patches is an important factor in maintaining biodiversity, understanding the spread of infectious diseases and allocating resources for conservation. Many management efforts devote considerable resources to maintaining or restoring landscape connectivity [e.g., Crooks and Sanjayan (2006)]. Natural and anthropogenic disturbance can alter the landscape, affecting connectivity and increasing uncertainty about the functioning of current ecosystems. Data driven decision-making can lead to effective management of species and ecosystems.

Spatially-referenced genetic data are commonly used to study functional connectivity [e.g., Cushman et al. (2006), Durand et al. (2009), Guillot et al. (2005), McRae (2006), Slatkin (1987)]. Genetic variation in a metapopulation is a function of gene flow and mutation, and the spatial distribution of genetic variation is a function of migration within the metapopulation and the mating process which

Received December 2014; revised March 2016.
[1]Supported by the U.S. Geological Survey RWO 98.
Key words and phrases. Landscape genetics, sage grouse, optimal sampling.

mixes genes at a fine temporal scale. Genetic data consist of a snapshot in time of this spatio-temporal process, and are thus spatially-correlated observations. Spatial statistical models have been used in many forms in the field of landscape genetics [e.g., Guillot et al. (2009), Hanks and Hooten (2013), Selander (1970), Smouse and Peakall (1999), Sokal, Oden and Wilson (1991)]. Here we introduce a spatial statistical model with the specific aim to use the model to identify optimal spatial sampling designs for landscape genetic data.

In Section 2, we introduce our motivating landscape genetic study, which involves optimal retrospective sampling design for landscape genetics. In Section 3, we propose a hierarchical multinomial model with latent spatial autocorrelation for spatially-referenced genetic data. In Section 4, we describe a spatial covariance function that takes into account the lekking behavior of sage-grouse, and fit the hierarchical model from Section 3 to the sage-grouse genetic data. In Section 5, we describe approaches for retrospective spatial sampling design based on the hierarchical model from Section 3, and make recommendations on the allocation of the retrospective sampling effort for sage-grouse genetic data in the western United States. In Section 6, we discuss possible extensions to our latent variable approach to modeling landscape genetic data.

2. Greater sage-grouse in the western United States. The greater sage-grouse (*Centrocercus urophasianus*) is the largest species of grouse found in North America. This species depends on sagebrush (*Artemisia* spp.) for both food and cover, and the health of the sage-grouse population is a strong indicator of the health of sagebrush in the western U.S. [Connelly et al. (2000), Patterson (1952)]. Sage-grouse are a lekking species with leks often found in natural openings in sagebrush communities, surrounded by potential nesting habitat [e.g., Connelly, Hagen and Schroeder (2011)]. Leks are spatial locations where, during breeding seasons, male sage-grouse engage in competitive mating displays. Leks can be remarkably persistent in space and time, with some leks remaining active for up to 90 years [Dalke et al. (1963), Smith et al. (2005), Wiley (1973)].

Greater sage-grouse were historically widely distributed across 12 western states in the U.S. and three Canadian provinces, but have undergone both a population decline [Garton et al. (2011)] and a range contraction [Schroeder et al. (2004)]. Quantifying the distribution of sage-grouse genetic variability across the range of the species could be used to identify population structure and characteristics of gene flow that provide insight into regions where management actions could have the greatest impact.

2.1. *Sample collection.* From 2009 to 2012, sage-grouse feathers were collected from sage-grouse leks during the spring breeding season. An opportunistic sample of all the feathers observed on the lek was collected by collecting feathers observed while traversing the lek. Feathers were placed in paper envelopes with the spatial coordinates for the lek. Following collection, envelopes were stored in cool, dry facilities out of direct sunlight. All envelopes and feathers collected from

any given lek were pooled by lek, as it was impossible to determine before a lab analysis which feathers could have come from the same sage-grouse.

2.2. *Sample extraction and genotyping*. DNA was extracted from feathers using QIAGEN's DNeasy Blood and Tissue Kit and a user-developed protocol for purification of total DNA from nails, hair or feathers. Each sample was genotyped across a panel of 14 variable microsatellite loci, all of which have been redesigned specifically to optimize efficacy with low quality and quantity DNA acquired from noninvasively collected feather samples.

Each sample was amplified multiple times at each microsatellite locus to screen for genotyping error (e.g., allelic drop out, false alleles, scoring error, etc.). If there were any variation between the genotypes generated for each sample, samples were genotyped another two to six times to confirm genotype accuracy. As genetic material was coming from feathers, many of which were significantly weathered, some samples failed at some loci. We removed from the analysis any individual where samples failed at more than 1/3 of the loci. The program DROPOUT [McKelvey and Schwartz (2005)] was used to identify duplicate sample genotypes and screen for genotyping error.

By December 2012, genotype data had been obtained for 830 unique individuals at 243 distinct leks clustered in the southwestern half of the sage-grouse range (Figure 1). While feathers had been collected across the entire range, genetic data were not yet available for samples from many leks by December 2012.

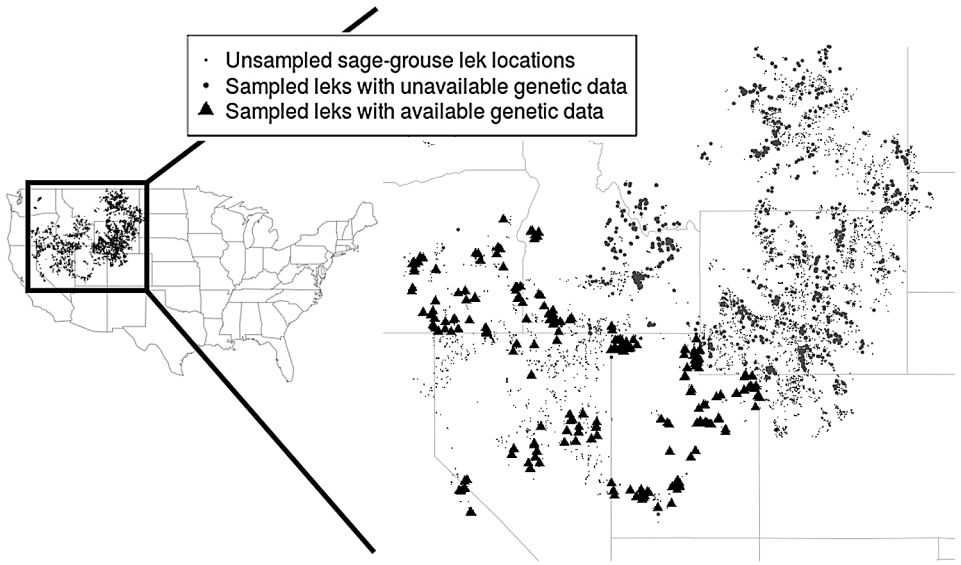

FIG. 1. *Sage-grouse lek locations in the western United States. Feather samples were collected from* 1177 *sage-grouse leks between* 2009 *and* 2012. *Microsatellite allele data from* 243 *of these leks were available for analysis in March*, 2013.

3. Latent spatial models for landscape genetics. A microsatellite is a portion of repetitive DNA in which certain short DNA sequences are repeated multiple times. In transcription, it is possible for the number of repeats of the DNA segment at a microsatellite to be either increased or decreased, as repeats are either transcribed multiple times or skipped. This leads microsatellites to often exhibit high mutation rates and corresponding high diversity in a population. An individual strand of DNA's microsatellite allele is characterized by the number of repeats of the short DNA sequence. Microsatellite alleles are inherited bi-parentally, and so are codominant, such that common alleles between two individuals shows shared ancestry, and a greater proportion of shared alleles indicates greater relatedness. While multiple feathers were collected at each lek, individual sage-grouse at each lek were identified by their unique genotype.

We propose a latent spatial model for spatially referenced microsatellite allele data where genetic relatedness is modeled using spatial correlation. In our specification, the categorical allele data are modeled using a multinomial probit data model with latent Gaussian spatial random effects.

Consider microsatellite allele data observed at L distinct loci for each spatially referenced individual in the study. At the ℓth locus, for $\ell = 1, 2, \ldots, L$, we denote the list of all distinct observed alleles from all individuals in the study as $\{a_{\ell 1}, a_{\ell 2}, \ldots, a_{\ell K_\ell}\}$. The actual numeric repeat length of the allele is not typically correlated with any important feature (e.g., fitness); rather, common alleles can indicate genetic relatedness. We will thus specify a model in which the alleles are considered as categorical observations. In particular, we model the two observed alleles for each (diploid) individual at each locus as arising from a multinomial distribution with spatially varying allele probabilities $\mathbf{p}_{s\ell} \equiv (p_{s\ell 1}, \ p_{s\ell 2}, \ \cdots, \ p_{s\ell K_\ell})'$, where $s \in \{1, 2, \ldots, S\}$ indexes the spatial location.

Let $\mathbf{y}_{sip\ell}$ be a vector coded such that the kth entry $y_{sip\ell k} = 1$ if the pth (indexing ploidy) observed allele at the ℓth locus is $a_{\ell k}$ for the ith individual sage-grouse at the sth spatial lek location, and $y_{sip\ell k} = 0$ otherwise. Then the multinomial probit model [e.g., Albert and Chib (1993)] for categorical data is specified in terms of latent variables, $\mathbf{z}$, as follows. Let

$$y_{sip\ell k} = \begin{cases} 1, & z_{sip\ell k} = \max\{z_{sip\ell a}, a = 1, \ldots, K_\ell\}, \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

where

$$z_{sip\ell k} \overset{\text{i.i.d.}}{\sim} N(\mu_{\ell k} + \eta_{s\ell k}, 1). \tag{2}$$

The allele $a_{\ell k}$ makes up a fraction $p_{s\ell k}$ of the genetic makeup of the sage-grouse at location s, where $p_{s\ell k} = \mathrm{P}(z_{sip\ell k} = \max\{z_{sip\ell a}, a = 1, \ldots, K_\ell\})$.

The mean of the latent variable $z_{sip\ell k}$ in (2) consists of the sum of two effects. The first is $\mu_{\ell k}$, an allele specific intercept which determines the relative frequency

of the kth allele at the ℓth locus across the entire population being studied. Large values of $\mu_{\ell k}$, relative to $\mu_{\ell k'}$, make it more likely that $z_{sip\ell k}$ will be larger than $z_{sip\ell k'}$, and so the kth allele will be more prevalent than the (k')th allele. We note that the model (1)–(2) is invariant to a shift in all $\mu_{\ell k}$, as the likelihood is a function of the contrasts $z_{sip\ell k} - z_{sip\ell k'}$, and not the actual values of $z_{sip\ell k}$. Thus, if $\mu_{\ell k}$ were replaced by $\mu_{\ell k} + c$ for $k = 1, 2, \ldots, K_\ell$ and some constant c, then the likelihood of the observed allele data would remain unchanged. To maintain model identifiability, we fix $\mu_{\ell 1} = 0$ for $\ell = 1, 2, \ldots, L$ because only the relative differences (contrasts) in $\mu_{\ell k}$ are identifiable.

The second term in the mean of (2) is $\eta_{s\ell k}$, which is a spatially varying random effect that allows the allele frequencies $\mathbf{p}_{s\ell}$ to vary over the spatial range of the species. Consider

$$\boldsymbol{\eta}_{\ell k} = \begin{bmatrix} \eta_{1\ell k} & \eta_{2\ell k} & \cdots & \eta_{n\ell k} \end{bmatrix}' \sim N\big(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta})\big), \tag{3}$$

where $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is the spatial covariance matrix of the spatially referenced locations for which we have observations, parameterized by $\boldsymbol{\theta}$.

We adopt a Bayesian approach to modeling, and specify prior distributions for all parameters in (1)–(3):

$$\mu_{\ell k} \sim N\big(0, \sigma_\mu^2\big), \qquad \ell = 1, 2, \ldots, L, k = 2, 3, \ldots, K_\ell, \tag{4}$$

$$\boldsymbol{\theta} \sim [\boldsymbol{\theta}], \tag{5}$$

where the bracket notation "$[\cdot]$" indicates a probability distribution. The prior on $\boldsymbol{\theta}$ will depend on the covariance model used, and we thus leave the prior specification (5) in a general form for now. We note that for $\ell = 1, 2, \ldots, L$, each $\mu_{\ell 1}$ is fixed at 0, and thus in (4) we do not specify prior distributions for these fixed baseline allele intercepts.

Implicit in the multinomial model for observed alleles is the assumption that we have observed all possible alleles at each locus. This is not likely to be the case in most natural populations. Under the assumption that the latent allelic processes are independent ($z_{sip\ell k}$ is independent of $z_{sip\ell k'}$, $k \neq k'$), the latent model (2) is unaffected by not observing some alleles, but the interpretation of $p_{s\ell k}$ changes to be the relative probability of observing the kth allele at the ℓth locus of an individual at the sth spatial location, *given* that one of $\{a_{\ell 1}, a_{\ell 2}, \ldots, a_{\ell K_\ell}\}$ is observed.

4. Sage-grouse landscape genetic analysis.

4.1. *Lek network connectivity model.* In the previous section we specified a multinomial model for spatially referenced microsatellite allele data with latent spatial autocorrelation. Critical to our approach is the choice of covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ for the latent spatial random effects in (3). Leks are persistent centers of sage-grouse breeding activity, and are thus essential to gene flow and spatial genetic variation.

We will consider a covariance that is based on the isolation by resistance (IBR) approach of McRae (2006) and takes into account the importance of leks in sage-grouse gene flow, similar in spirit to the model of Knick and Hanser (2011). The IBR approach envisions space as a graph: a set of spatial nodes (leks in our case) connected by edges, where the edges are resistors in an electric circuit. McRae (2006) showed that if the resistance (edge weights) of the resistors was inversely proportional to migration rates between nodes, and migration could be modeled as a random walk on the graph of nodes, then the resistance distance of the graph [Klein and Randić (1993)] is proportional to linearized F_{st}.

Consider a spatial network with nodes at known sage-grouse leks across the western U.S. and edge weights (conductances in the IBR framework) being a function of the Euclidean distance between leks. For the sth lek, consider the set of "neighboring" leks indexed by $t \in \mathcal{N}(s)$, where $\mathcal{N}(s)$ is the set of leks that are directly connected (neighbors) to the sth lek. We consider lek s and lek t to be neighbors if a sage-grouse could plausibly migrate from lek s to lek t directly, without an intermediate stop at any other lek. This is based on the migration model for movement central to the IBR approach as described by McRae (2006).

Define the edge weight between the sth and tth leks as α_{st}/σ^2, where α_{st} is a decreasing function of the Euclidean distance between the leks and σ^2 is a scaling factor. Under the random walk model for migration in the IBR framework, α_{st}/σ^2 is proportional to the migration rate of sage-grouse between the two leks. We consider edge weights that are decreasing functions $f(d_{st})$ of the Euclidean distance d_{st} between leks, and that are set to 0 for all leks that are more distant than a maximum distance d_{MAX}:

$$\alpha_{st} = \begin{cases} f(d_{st}), & d_{st} \le d_{\mathrm{MAX}}, \\ 0, & d_{st} > d_{\mathrm{MAX}}. \end{cases} \tag{6}$$

We will consider multiple functional forms for $f(d)$ and multiple maximum distances d_{MAX} in Section 4.3, and use information criteria to choose the functional form and maximum distance most appropriate for the sage-grouse genetic data. Setting edge weights equal to zero for leks that are more distant than d_{MAX} indicates that migration (and the resulting flow of genetic information) is only possible between distant leks through the use of intermediate leks in the lek network. This is a Markov assumption, and our resulting random effects $\{\boldsymbol{\eta}_{\ell k}\}$ in (3) are Gaussian Markov random fields.

Hanks and Hooten (2013) showed that the spatial connectivity implied by such a random walk migration model (or an equivalent electric circuit) can be modeled as an intrinsic conditional autoregressive (ICAR) Gaussian Markov random field [Besag (1974), Besag and Kooperberg (1995)]. The spatial precision matrix

(inverse covariance matrix) for an ICAR on the entire lek network is given by

$$\frac{1}{\sigma^2}\mathbf{Q} = \frac{1}{\sigma^2}\begin{bmatrix} \sum_{j\neq 1}\alpha_{1j} & -\alpha_{12} & -\alpha_{13} & \cdots \\ -\alpha_{21} & \sum_{j\neq 2}\alpha_{2j} & -\alpha_{23} & \cdots \\ -\alpha_{31} & -\alpha_{32} & \sum_{j\neq 3}\alpha_{3j} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \tag{7}$$

We do not have observed genetic data for all known lek locations. If we divide the leks into sampled (or observed o) and unobserved (u) leks, the spatial random effects $\boldsymbol{\eta}$ (suppressing the locus ℓ and allele k subscripts) and precision matrix $\mathbf{Q}$ can be partitioned accordingly:

$$\boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{\eta}_o \\ \boldsymbol{\eta}_u \end{bmatrix}, \qquad \mathbf{Q} = \frac{1}{\sigma^2}\left[\begin{array}{c|c} \mathbf{Q}_{oo} & \mathbf{Q}_{ou} \\ \hline \mathbf{Q}_{uo} & \mathbf{Q}_{uu} \end{array}\right]$$

and the covariance matrix $\boldsymbol{\Sigma}(\sigma^2)$ of the observed leks is given by the Schur complement

$$\boldsymbol{\Sigma}(\sigma^2) = \sigma^2\left[\mathbf{Q}_{oo} - \mathbf{Q}_{ou}(\mathbf{Q}_{uu})^{-1}\mathbf{Q}_{uo}\right]^{-1}. \tag{8}$$

We assign a conjugate inverse gamma prior distribution to σ^2:

$$\sigma^2 \sim \mathrm{IG}(r, q), \tag{9}$$

with r and q chosen so that the prior has mean of 10 and variance of 100. We also fixed the hyperparameter $\sigma_\mu^2 = 100$. Inference on the parameters ($\{\mu_{\ell k}\}$ and σ^2) in the full model, given by equations (1)–(4) and (6)–(9), can then be made using a Markov chain Monte Carlo (MCMC) algorithm under a Bayesian statistical paradigm.

4.2. *Reduced-rank spatial model.* Approximately 6000 known sage-grouse leks reside across the western United States, resulting in a large ($\approx 6000 \times 6000$) spatial precision matrix $\mathbf{Q}$ representing pairwise connections between leks. The magnitude of this spatial precision matrix will make model fitting and the comparison of potential retrospective sampling designs very computationally intensive. We thus propose a reduced-rank model for the spatial random effect $\boldsymbol{\eta}_{\ell k}$ in (3). Our reduced-rank approach utilizes a spectral decomposition of the spatial precision matrix $\mathbf{Q}$ and is similar to that of Wikle and Cressie (1999), Berliner, Wikle and Cressie (2000) and others.

For a fully observed lek network, $\boldsymbol{\eta}_{\ell k} \sim N(\mathbf{0}, \sigma^2\mathbf{Q}^-)$, where $\mathbf{Q}/\sigma^2$ is an $n_s \times n_s$ precision matrix. The spectral decomposition of $\mathbf{Q}$ yields

$$\mathbf{Q} = \mathbf{M}\mathbf{D}^-\mathbf{M}',$$

where the columns of $\mathbf{M}$ are the eigenvectors of $\mathbf{Q}$ and $\mathbf{D}$ is a diagonal matrix containing the respective eigenvalues $\{\lambda_1, \lambda_2, \ldots, \lambda_{n_s}\}$ of $\mathbf{Q}^-$. Then the spatial random effect can be expressed as

$$\boldsymbol{\eta}_{\ell k} = \mathbf{M}\boldsymbol{\delta}_{\ell k},$$

where

$$\boldsymbol{\delta}_{\ell k} \sim N(\mathbf{0}, \sigma^2 \mathbf{D}).$$

Then a reduced-rank version of the spatial random effect is given by taking only the first n_e eigenvectors of $\mathbf{Q}^-$ and setting $\lambda_m = 0$ for all $m > n_e$. This approximates the random effect $\boldsymbol{\eta}_{\ell k}$ with a random effect $\tilde{\boldsymbol{\eta}}_{\ell k}$ that captures a portion of the spatial structure in $\mathbf{Q}^-$. In practice, n_e can be chosen so that $\tilde{\boldsymbol{\eta}}_{\ell k} \approx \boldsymbol{\eta}_{\ell k}$ but n_e is still much smaller than n_s, the number of leks in the network. This leads to a computationally efficient representation of the spatial random effect, with (3) becoming

$$\tilde{\eta}_{s\ell k} = \tilde{\mathbf{m}}_s' \tilde{\boldsymbol{\delta}}_{\ell k}, \tag{10}$$

$$\tilde{\boldsymbol{\delta}}_{\ell k} \sim N(\mathbf{0}, \sigma^2 \tilde{\mathbf{D}}), \tag{11}$$

where $\tilde{\mathbf{m}}_s'$ is the sth row of $\tilde{\mathbf{M}}$, the $n_s \times n_e$ matrix of the first n_e eigenvectors of $\mathbf{Q}^-$, and $\tilde{\mathbf{D}}$ is the $n_e \times n_e$ diagonal matrix of the first n_e eigenvalues of $\mathbf{Q}^-$.

The complete hierarchical statistical model we have described for the multinomial allele model with latent reduced-rank spatial random effects is

$$\begin{aligned}
y_{sip\ell k} &= \begin{cases} 1, & z_{sip\ell k} = \max\{z_{sip\ell a}, a = 1, \ldots, K_\ell\}, \\ 0, & \text{otherwise}, \end{cases} \\
z_{sip\ell k} &\sim N(\mu_{\ell k} + \tilde{\eta}_{s\ell k}, 1), \\
\tilde{\eta}_{s\ell k} &= \tilde{\mathbf{m}}_s' \tilde{\boldsymbol{\delta}}_{\ell k}, \\
\tilde{\boldsymbol{\delta}}_{\ell k} &\sim N(\mathbf{0}, \sigma^2 \tilde{\mathbf{D}}), \\
\mu_{\ell k} &\sim N(0, \tau^2), \\
\sigma^2 &\sim \mathrm{IG}(r, q).
\end{aligned}$$

Full-conditional distributions are available for all parameters in this hierarchical model, and are given in the supplemental article [Hanks et al. (2016)].

4.3. *Model fitting*. For our analysis of sage-grouse genetic data, two functional forms for $f(d)$ in (6) were specified, inverse distance [$f_1(d) = 1/d$], and inverse squared distance [$f_2(d) = 1/d^2$]. For each of these functional forms, three different models with d_{MAX} being specified as 18 km, 25 km and 50 km were specified. The six resulting spatial models were each fit to the observed allele data from 830 feather samples collected at 243 distinct leks (Figure 1) across the western U.S. Under a Bayesian statistical paradigm, an MCMC algorithm was used to obtain samples from the posterior distribution of parameters in each network model, conditioned on the observed sage-grouse genotype data. Convergence was assessed visually, with chains showing good mixing.

The models were compared using the deviance information criterion (DIC) of Spiegelhalter et al. (2002) (Table 1). The best model (DIC = 661,844) was the

TABLE 1
Comparison of spatial lek-network models based on DIC

Model	$f(d)$	d_{MAX}	DIC
1	$1/d^2$	25,000.00	661,844
2	$1/d$	50,000.00	662,409
3	$1/d$	25,000.00	664,848
4	$1/d^2$	50,000.00	670,883
5	$1/d$	18,000.00	671,375
6	$1/d^2$	18,000.00	674,832

network model with edge weight between leks inversely proportional to the square of the distance between leks $[f_1(d) = 1/d^2]$ and $d_{\mathrm{MAX}} = 25$ km. We note that d_{MAX} gives the distance at which pairwise edges between leks and corresponding entries of $\mathbf{Q}$ are set to zero, implying conditional independence between two leks further apart than 25 km, conditional on the random effects $\eta_{s\ell k}$ at all other leks.

For the following analysis, we focus only on the model with the best (lowest) DIC. This corresponds to an empirical Bayesian approach to model fitting. An alternative approach would be to assign a prior distribution to the exponent (ξ) in $f(d) = d^{-\xi}$ in (6) and jointly estimate the posterior distribution of ξ with the posteriors for other model parameters. A similar approach could be taken for the cutoff parameter d_{MAX}. However, this approach would be computationally demanding, and in testing leads to poor mixing of MCMC chains. Our empirical Bayes approach is similar in spirit to assigning a discrete uniform prior on the range parameter in a geostatistical spatial model [e.g., Diggle and Ribeiro (2007)].

The posterior distribution for σ^2 in the best model has mean of 0.169 and variance of 0.015. The MCMC standard error based on the consistent batch means estimator [Flegal, Haran and Jones (2008)] for σ^2 is 0.0011. This provides some confidence that our MCMC algorithm has converged to the stationary posterior distribution.

The quantity σ^2 is not easily interpreted directly, but the latent representation of $z_{sip\ell k}$ lends an interpretation to the latent spatial covariance matrix $\tilde{\mathbf{\Sigma}}(\sigma^2)$. From (2),

$$z_{sip\ell k} = \mu_{\ell k} + \tilde{\eta}_{s\ell k} + \varepsilon_{sip\ell k}, \qquad \varepsilon_{sip\ell k} \sim N(0, 1),$$

where $\tilde{\eta}_{\ell k} \sim N(\mathbf{0}, \tilde{\mathbf{\Sigma}}(\sigma^2))$. The spatial covariance matrix $\tilde{\mathbf{\Sigma}}(\sigma^2)$ is nonstationary, and the average of the posterior mean diagonal elements of $\tilde{\mathbf{\Sigma}}(\sigma^2)$ is 8.43, which is much greater than the unit variance of the nonspatial noise (nugget) contained in $\varepsilon_{sip\ell k}$ of (2), leading us to conclude that spatial variability in the genetic data accounts for a much larger proportion of the variability than does nonspatial (within lek) variability. Summaries of posterior distributions for $\{\mu_{\ell k}, \ell = 1, 2, \ldots, 14, k = 1, 2, \ldots, K_\ell\}$ are not shown, but correlate well with empirical allele frequencies taken over all 1136 genetic samples.

5. Optimal retrospective sampling for sage-grouse in the western United States. Having fit the model to microsatellite allele data from sage-grouse feather samples collected in 2009–2012, we now consider a retrospective sampling design for this system. Our goal is to recommend optimal regions for sampling previously unsampled sage-grouse leks, given what is known from the spatially referenced genetic data already collected. This requires the definition of a criterion $\phi(\mathbf{d}_j)$ by which we can compare J potential sampling designs ($\mathbf{d}_j$, $j = 1, \ldots, J$).

5.1. *Latent Gaussian design.* A common criterion to minimize for optimal sampling of Gaussian random variables is the mean squared prediction error (MSPE) at unobserved locations [e.g., Zimmerman (2006)]. The observations $\{y_{sip\ell k}\}$ in our model are not Gaussian, rather they are categorical alleles which we model as arising from a multinomial distribution. However, each $z_{sip\ell k}$ is a latent representation of the true observation $y_{sip\ell k}$, and $z_{sip\ell k}$ is normally distributed such that

$$z_{sip\ell k} = \mu_{\ell k} + \eta_{s\ell k} + \varepsilon_{sip\ell k}, \qquad \varepsilon_{sip\ell k} \sim N(0, 1).$$

This leads us to consider the MSPE of the latent z variables as a design criterion. Full posterior predictive inference on this MSPE would be computationally prohibitive, as it would require integrating over the entire posterior distribution. Instead, in what follows, we will consider the parameters $\{\mu_{\ell k}\}$ and σ^2 to be fixed and known, and we will focus on finding optimal designs conditional on the posterior mean values of model parameters. Our development of a design criterion first focuses on the design for the latent Gaussian $\mathbf{z}$, and is reminiscent of Kriging [e.g., Cressie (1993)]. We will then consider a design criterion that takes into account the categorical nature of the data by integrating over the latent Gaussian $\mathbf{z}$.

For the kth allele at the ℓth locus, we can divide the spatial locations, and their corresponding z variables, into observed ($\mathbf{z}_{\ell k}^{(o)} = \{z_{sip\ell k} : s \text{ is observed}\}$) and unobserved ($\mathbf{z}_{\ell k}^{(u)} = \{z_{sip\ell k} : s \text{ is unobserved}\}$) categories. The "observed" set corresponds to the genetic samples obtained from the sage-grouse leks, while the "unobserved" set will be a single unsampled sage-grouse at each unsampled lek. The joint distribution of $\mathbf{z}_{\ell k}$ at all leks is Gaussian:

$$\mathbf{z}_{\ell k} = \begin{pmatrix} \mathbf{z}_{\ell k}^{(o)} \\ \mathbf{z}_{\ell k}^{(u)} \end{pmatrix} \sim N\left(\mu_{\ell k}\mathbf{1}, \begin{pmatrix} \mathbf{\Psi}_{oo} & \mathbf{\Psi}_{ou} \\ \mathbf{\Psi}_{uo} & \mathbf{\Psi}_{uu} \end{pmatrix}\right), \tag{12}$$

where the covariance matrix of $\mathbf{z}_{\ell k}$ is $\mathbf{\Psi} = \sigma^2\mathbf{\Sigma} + \mathbf{I}$ and has been partitioned according to observed and unobserved locations.

If $\mathbf{z}^{(o)}$ are known (observed), then the mean of the distribution of $\mathbf{z}_{\ell k}^{(u)}$ (conditional on $\mathbf{z}^{(o)}$) is given by

$$\hat{\mathbf{z}}_{\ell k}^{(u)} = \mu_{\ell k}\mathbf{1} + \mathbf{\Psi}_{uo}\mathbf{\Psi}_{oo}^{-1}\big(\mathbf{z}_{\ell k}^{(o)} - \mu_{\ell k}\mathbf{1}\big), \tag{13}$$

and the corresponding conditional Gaussian covariance matrix is given by

$$\boldsymbol{\Phi} = E\big[(\mathbf{z}_u - \hat{\mathbf{z}}_u)(\mathbf{z}_u - \hat{\mathbf{z}}_u)'\big] = \boldsymbol{\Psi}_{uu} - \boldsymbol{\Psi}_{uo}\boldsymbol{\Psi}_{oo}^{-1}\boldsymbol{\Psi}_{ou}. \tag{14}$$

Each diagonal entry of $\boldsymbol{\Phi}$ contains the MSPE for z at an unobserved location, and we choose the sum of these entries ["A-optimality" in Harville (2008)] for our latent Gaussian design criterion:

$$\phi_{\mathrm{LG}}(\mathbf{d}) = \mathrm{tr}(\boldsymbol{\Phi}) = \mathrm{sum}\big(\mathrm{diag}(\boldsymbol{\Phi})\big). \tag{15}$$

Other possible alternatives include "D-optimality," in which the design criterion consists of the product of the diagonal elements of $\boldsymbol{\Phi}$. While D-optimality has some desirable properties when considering optimal covariate designs, we focus on A-optimality, which is the prevailing approach in spatial sampling [e.g., Hooten et al. (2009), Wikle and Royle (2005), Zimmerman (2006)] and is interpretable with the design criterion being the mean square prediction error averaged over all unobserved spatial locations.

It is notable that neither the covariance matrix $\boldsymbol{\Phi}$ nor the design criterion $\phi_{\mathrm{LG}}(\mathbf{d})$ depend on $\mathbf{z}_{\ell k}^{(o)}$. Additionally, this design criterion would be identical for any locus ℓ and allele k, as we assume in (3) that the spatial correlation is shared by all latent allelic processes $\{\mathbf{z}_{\ell k}\}$. As the design criterion ϕ_{LG} depends only on the covariance matrix $\boldsymbol{\Phi}$, computation of this design criterion is very efficient, as changing a design (e.g., adding a new lek to the set of previously sampled leks) only involves evaluating (14) and (15) for a different partition of locations into "observed" and "unobserved" in (12).

5.2. *Design for categorical observations*. The genetic observations are not Gaussian, and it is not clear whether the latent Gaussian design criterion (15) is appropriate for categorical data. We thus consider a categorical design criterion. Consider predicting the first (of ploidy 2) observed allele at locus ℓ for an individual sage-grouse at an unsampled spatial location u, that is, we want to predict $\mathbf{y}_{u\ell} = (y_{u11\ell 1}, \ldots, y_{u11\ell K})'$, which will be a vector with one entry equal to 1 (indicating the observed allele) and all other entries equal to zero. Note that, due to our latent Gaussian model (1)–(3), predictions on $\mathbf{y}_{u\ell}$ can be obtained by integrating over predictions on $(z_{u11\ell 1}, \ldots, z_{u11\ell K_\ell})$. As we did in Section 5.1, we will consider prediction and sampling design conditioned on the posterior mean values of $\{\mu_{\ell k}\}$, σ^2 and the latent $\{\mathbf{z}_{\ell k}^{(o)}\}$. Our approach could be extended to full posterior predictive inference by integrating over the posterior distribution of these parameters, but this is computationally prohibitive in the case of the sage-grouse study.

If we condition on $\{\mathbf{z}_{\ell 1}^{(o)}, \ldots, \mathbf{z}_{\ell K_\ell}^{(o)}\}$, then the mean of the latent Gaussian $(z_{u11\ell 1}, \ldots, z_{u11\ell K_\ell})$ is $(\hat{z}_{u11\ell 1}, \ldots, \hat{z}_{u11\ell K_\ell})$, with each entry obtained using (13). Our prediction of $\mathbf{y}_{u\ell}|\{\mathbf{z}_{\ell 1}^{(o)}, \ldots, \mathbf{z}_{\ell K_\ell}^{(o)}\}$ is then $y_{u11\ell\hat{k}_{u\ell}} = 1$, where $\hat{z}_{u11\ell\hat{k}_{u\ell}} > \hat{z}_{u11\ell a}$, $a \neq \hat{k}_{u\ell}$, with all other entries in $\mathbf{y}_{u\ell}$ equal to zero.

To obtain a design criterion, we note that the MSPE of $\mathbf{y}_{u\ell}|\{\mathbf{z}_{\ell 1}^{(o)}, \ldots, \mathbf{z}_{\ell K_\ell}^{(o)}\}$ is equal to the probability that the index $k_{u\ell}$ of the maximum entry of $(z_{u11\ell 1}, \ldots, z_{u11\ell K_\ell})$ is different from the index $\hat{k}_{u\ell}$ of the maximum entry of $(\hat{z}_{u11\ell 1}, \ldots, \hat{z}_{u11\ell K_\ell})$.

Each $z_{u11\ell a}$ is marginally distributed as

$$z_{u11\ell a}|\{\mathbf{z}_{\ell a}^{(o)}\} \sim N(\hat{z}_{u11\ell a}, \psi_u^2),$$

where $\hat{z}_{u11\ell a}$ is given by (13) and ψ_u^2 is the uth diagonal element of $\mathbf{\Phi}$ in (14). While the latent z random variables are correlated in space, they are uncorrelated between alleles; that is, $z_{u11\ell a}$ is independent of $z_{u11\ell a'}$ for $a \neq a'$, and $\mathbf{z}_{u\ell} = (z_{u11\ell 1}, z_{u11\ell 2}, \ldots, z_{u11\ell K_\ell})'$ is distributed

$$\mathbf{z}_{u\ell}|\{\mathbf{z}_{\ell a}^{(o)}, a = 1, \ldots, K_\ell\} \sim N(\hat{\mathbf{z}}_{u\ell}, \psi_u^2 \mathbf{I}). \tag{16}$$

Without loss of generality, suppose that the index $\hat{k}_{u11\ell}$ of the maximum $\hat{z}_{u11\ell a}$ is $\hat{k}_{u11\ell} = K_\ell$; that is, the observed allele is the last allele in the list. Then the probability of an incorrect categorical allele prediction is

$$\begin{aligned}
\theta_{u\ell} &= P(k_{u11\ell} \neq \hat{k}_{u11\ell}|\{\mathbf{z}_{\ell a}^{(o)}, a = 1, \ldots, K_\ell\}) \\
&= 1 - P(k_{u11\ell} = \hat{k}_{u11\ell}|\{\mathbf{z}_{\ell a}^{(o)}, a = 1, \ldots, K_\ell\}) \\
&= 1 - P(z_{u11\ell 1} - z_{u11\ell K_\ell} < 0, z_{u11\ell 2} - z_{u11\ell K_\ell} < 0, \ldots, \\
&\qquad z_{u11\ell K_\ell - 1} - z_{u11\ell K_\ell} < 0|\{\mathbf{z}_{\ell a}^{(o)}, a = 1, \ldots, K_\ell\}),
\end{aligned}$$

which is an orthant probability of the multivariate normal $\mathbf{w}_{u\ell} \sim N(\mathbf{L}\hat{\mathbf{z}}_{u\ell}, \psi_u^2 \mathbf{L}\mathbf{L}')$, defined as

$$\mathbf{w}_{u\ell} = \begin{bmatrix} w_{u11\ell 1} \\ w_{u11\ell 2} \\ \vdots \\ w_{u11\ell K_\ell - 1} \end{bmatrix} = \mathbf{L}\mathbf{z}_{u\ell} = \begin{bmatrix} 1 & 0 & 0 & \cdots & -1 \\ 0 & 1 & 0 & \cdots & -1 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \cdots & 1 & -1 \end{bmatrix} \begin{bmatrix} z_{u11\ell 1} \\ z_{u11\ell 2} \\ \vdots \\ z_{u11\ell K_\ell} \end{bmatrix}, \tag{17}$$

with $\mathbf{L}$ a contrast matrix of dimension $(K_\ell - 1) \times K_\ell$. Our proposed categorical design criterion is then the sum of this probability for each locus at each unobserved location:

$$\begin{aligned}
\phi_{\text{cat}}(\mathbf{d}) &= \sum_u \sum_\ell \theta_{u\ell} \\
&= \sum_u \sum_\ell (1 - P(w_{u11\ell 1} < 0, w_{u11\ell 2} < 0, \ldots, w_{u11\ell K_\ell - 1} < 0)).
\end{aligned} \tag{18}$$

Note that this design criterion considers prediction error at each locus for one individual at each unobserved sample location.

The calculation of each of the multivariate normal orthant probabilities $\{\theta_{u\ell}\}$ is simplified, as the correlation matrix can be written as $\mathbf{L}\mathbf{L}' = \mathbf{I} + \mathbf{1}\mathbf{1}'$. In this special

case, each multivariate orthant probability can be expressed as a one-dimensional integral on [0, 1] [Genz and Bretz (2009), pages 3, 16–17, Marsaglia (1963)]:

$$\theta_{u\ell} = 1 - \int_0^1 \prod_{k=1}^{K_\ell - 1} \left(1 - F\left(\frac{\hat{z}_{u\ell K_\ell} - \hat{z}_{u\ell k}}{\sqrt{\psi_u^2}} - F^{-1}(t)\right)\right) \mathrm{d}t,$$

where F is the CDF of a standard normal random variable. We approximated this one-dimensional integral using numerical quadrature. This provides an efficient approach to calculating the categorical design criterion ϕ_{cat} (18).

5.3. *Optimal sampling for reduced-rank spatial models.* Computing the design criteria (15) and (18) requires calculation of the inverse $\boldsymbol{\Psi}_{oo}^{-1}$ which has computational cost of $O(n_d^3)$, where n_d is the number of locations in the proposed design. In the sage-grouse study, $n_d = 1177$ leks were sampled during the years of 2009–2012. For our retrospective design, we consider adding leks to the sampling design, which will increase this number. Inverting a matrix of this size is computationally demanding, especially if we wish to compare a large number of potential sampling designs. However, we can take advantage of the reduced-rank spatial model presented in Section 4.2 to significantly reduce the computational cost of computing the design criterion by replacing the covariance matrix $\boldsymbol{\Phi} = \boldsymbol{\Psi}_{uu} - \boldsymbol{\Psi}_{uo}\boldsymbol{\Psi}_{oo}^{-1}\boldsymbol{\Psi}_{ou}$ (14) with a reduced-rank approximation $\tilde{\boldsymbol{\Phi}} = \tilde{\boldsymbol{\Psi}}_{uu} - \tilde{\boldsymbol{\Psi}}_{uo}\tilde{\boldsymbol{\Psi}}_{oo}^{-1}\tilde{\boldsymbol{\Psi}}_{ou}$.

Note that the reduced-rank covariance matrix $\tilde{\boldsymbol{\Phi}}$ of all observed and unobserved nodes can be decomposed and written in block form:

$$\begin{aligned}\tilde{\boldsymbol{\Phi}} &= \sigma^2 \tilde{\mathbf{M}} \tilde{\mathbf{D}} \tilde{\mathbf{M}}' + \mathbf{I} \\ &= \sigma^2 \cdot \begin{bmatrix} \tilde{\mathbf{M}}_o \\ \tilde{\mathbf{M}}_u \end{bmatrix} \tilde{\mathbf{D}} \begin{bmatrix} \tilde{\mathbf{M}}_o' & \tilde{\mathbf{M}}_u' \end{bmatrix} + \begin{bmatrix} \mathbf{I}_o & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_u \end{bmatrix},\end{aligned}$$

which leads to a reduced-rank version of the Kriging covariance matrix:

$$\tilde{\boldsymbol{\Phi}} = \tilde{\boldsymbol{\Psi}}_{uu} - \tilde{\boldsymbol{\Psi}}_{uo}\tilde{\boldsymbol{\Psi}}_{oo}^{-1}\tilde{\boldsymbol{\Psi}}_{ou} \tag{19}$$

$$\begin{aligned}&= \sigma^2 \tilde{\mathbf{M}}_u \tilde{\mathbf{D}} \tilde{\mathbf{M}}_u' + \mathbf{I}_u \\ &\quad - (\sigma^2 \tilde{\mathbf{M}}_u \tilde{\mathbf{D}} \tilde{\mathbf{M}}_o')(\sigma^2 \tilde{\mathbf{M}}_o \tilde{\mathbf{D}} \tilde{\mathbf{M}}_o' + \mathbf{I}_o)^{-1}(\sigma^2 \tilde{\mathbf{M}}_o \tilde{\mathbf{D}} \tilde{\mathbf{M}}_u').\end{aligned} \tag{20}$$

This decomposition requires the inverse of an $n_d \times n_d$ matrix, but the inverse in question can be expressed using the Sherman–Morrison–Woodbury identity [e.g., Gentle (2007), page 221]

$$(\tilde{\mathbf{M}}_o(\sigma^2 \tilde{\mathbf{D}})\tilde{\mathbf{M}}_o' + \mathbf{I}_o)^{-1} = \mathbf{I}_o - \tilde{\mathbf{M}}_o\left(\frac{1}{\sigma^2}\tilde{\mathbf{D}}^{-1} + \tilde{\mathbf{M}}_o' \mathbf{I}_o \tilde{\mathbf{M}}_o\right)^{-1} \tilde{\mathbf{M}}_o'. \tag{21}$$

The resulting expression requires only the inverse of a matrix of dimensionality $n_e \times n_e$, where n_e is equal to the number of eigenvectors retained in the reduced-rank version of $\mathbf{Q}$ and typically $n_e \ll n_d$. This formulation allows us to take advantage of the reduced-rank approximation to the spatial covariance in the computation of the sampling criteria $\phi_{\mathrm{LG}}(\mathbf{d})$ and $\phi_{\mathrm{cat}}(\mathbf{d})$. Similar reduced-rank approximations have been used for prediction [e.g., Wikle and Cressie (1999)], but not, to our knowledge, in optimal sampling design.

5.4. *Recommendations for retrospective sage-grouse sampling effort.* To make recommendations for the retrospective sampling effort of sage-grouse leks, we identify optimal retrospective designs under both the latent Gaussian MSPE sampling criterion (15) and the categorical sampling criterion (18). As described in Section 4.3, genetic data from 830 individuals at 243 distinct leks (Figure 1) were used to fit the statistical model proposed in Section 4.2. Each of these leks were considered to be observed for our purpose of identifying an optimal retrospective design. We also included 934 additional leks where feather samples had been collected during 2009–2012 at these additional leks, but had not been genotyped before the resumption of sampling. The genetic information from these samples could thus not be used to estimate the spatial covariance parameters in our statistical model, but knowing that samples have been obtained from these additional sampling locations can aid in identifying regions of high sampling importance for the retrospective sampling effort. This results in a total of $243 + 934 = 1177$ leks which we consider to be observed, or sampled, in the existing design.

Given this set of observed leks, we consider adding R additional leks to the existing design. The motivation behind this type of design is that resources may exist to collect genetic information (e.g., sage-grouse feathers) at only a subset of the known leks across the western United States. Under such a constraint, we chose this retrospective design of R additional leks so that the sampling design criterion is optimized. For illustrative purposes, we consider $R = 10$.

A complete search of the space of retrospective designs would be computationally prohibitive, thus we constructed an iterative search algorithm to identify a pseudo-optimal retrospective design. We first considered 10,000 random retrospective designs of ten additional leks randomly chosen from all leks where feathers had not been collected (unsampled leks), and computed the latent Gaussian sampling criterion (15) and categorical criterion (18) for each. We also considered multiple manually specified initial designs chosen so that the ten new leks were regularly spaced in geographic space. The best six designs from this initial search were used as initial designs in an iterative exchange algorithm [Royle (1998)]. The switching algorithm is detailed in Algorithm 1.

We ran the iterative algorithm for $N_{\mathrm{switch}} = 1000$ iterations for each initial design. A local minimum was found for each initial design; the resulting designs were then compared, and the design with the smallest design criterion was chosen. This process was repeated once for the latent Gaussian criterion (15) and once for

Algorithm 1 Sampling design switching algorithm

Set the initial design $\mathbf{d} = (s_1, s_2, \ldots, s_R)$ ▷ R = the size of the design
Compute the initial design criterion $\phi(\mathbf{d})$
Set N_{switch} = the number of iterations
Set d_{MAX}
for iter in $1{:}N_{\text{switch}}$ **do**
    for v in $1{:}R$ **do**
        Draw s_v^* uniformly from $\{s : \|s - s_v\| < d_{\text{MAX}}\}$
        Compute the design criterion $\phi(\mathbf{d}^*)$ for $\mathbf{d}^* = (s_1, \ldots, s_{v-1}, s_v^*, s_{v+1}, \ldots, s_R)$
        if $\phi(\mathbf{d}^*) < \phi(\mathbf{d})$ **then**
            $\mathbf{d} \leftarrow \mathbf{d}^*$
return d ▷ **d** is a locally-optimal design

the categorical criterion (18). The resulting optimal designs for the two criteria are shown in Figure 2. This retrospective design is based on the estimated spatial covariance for a network model of lek connectivity and provides guidance on how states can prioritize future sage-grouse lek sampling efforts by identifying 10 leks of high sampling importance in a study designed to characterize sage-grouse genetic diversity.

5.5. *Simulation study.* We conducted a simulation study to compare how effective the latent Gaussian design criterion ϕ_{LG} and the categorical design criterion ϕ_{cat}^* are at identifying important sampling locations. Using posterior mean parameters from our results in Section 4, we simulated spatially correlated categorical data on the entire sage-grouse lek network using the model described in (1)–(3). We considered simulating one locus with 10 alleles for one individual at each lek in the spatial network. We considered predicting the simulated categorical observation conditional on a fixed, true value of σ^2 and the fixed, true latent Gaussian $z_{sip\ell k}$ at each lek in the optimal design. The latent $z_{sip\ell k}$ at leks not in the optimal design (where predictions are required) were treated as unknown. This mimics the case in which a researcher has estimated σ^2 and $z_{sip\ell k}$ at the locations in the optimal design from existing data, such as through the posterior mean after model fitting as described in Section 4. In practice, there will be uncertainty about these parameter estimates; fixing them at their true values for this simulation study allows us to study the relative merits of the latent Gaussian and full categorical design criteria in a "best case" scenario. As always, poor estimation of model parameters resulting from biased samples or poor choices of prior distributions could negatively affect predictive error. For optimal designs, we used the optimal retrospective designs from Section 5.4, which are shown in Figure 2. This will allow us

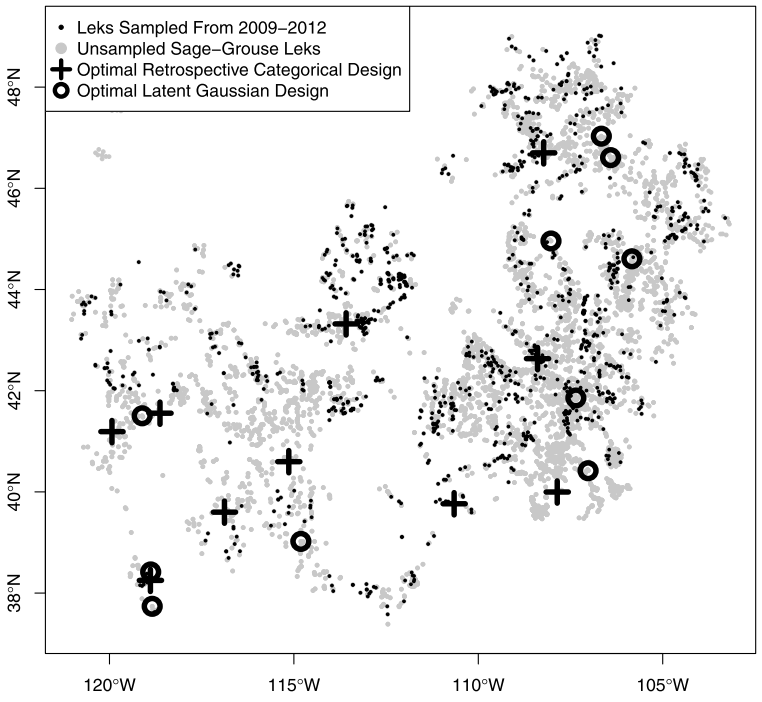

FIG. 2. *Sage-grouse lek sampling recommendations based on latent Gaussian MSPE. Leks shown as grey circles indicate leks that have not been sampled, while leks shown as black dots have been sampled in* 2009–2012. *Leks shown as crosses are the optimal categorical retrospective lek sampling locations under the constraint that no more than* 10 *additional leks are sampled in* 2013. *Leks shown as open circles are the optimal latent Gaussian retrospective lek sampling locations under the same constraint.*

to study the ability of the latent Gaussian design criterion to approximate the full categorical criterion. The predicted categorical observations were compared with the true simulated categorical observations, and the MSPE was computed as the number of leks where the predicted allele was not the true simulated allele. This process was repeated 1000 times, with results shown in Figure 3.

In Figure 3(a), we show the categorical and latent Gaussian design criteria for 300 random designs. It is clear that the categorical and latent Gaussian design criteria are not always in agreement. This is evident in Figure 2 as well, where there is very little overlap between the optimal categorical and latent Gaussian designs. Figure 3(b) shows the simulation MSPEs for the optimal latent Gaussian design and the optimal categorical design for all runs in the simulation study. While the optimal categorical design (MSPE = 2451) does, on average, predict better than the optimal latent Gaussian design (MSPE = 2459), the difference is slight. This indicates that the latent Gaussian design criterion may be an effective substitute for the more computationally intensive categorical design criterion.

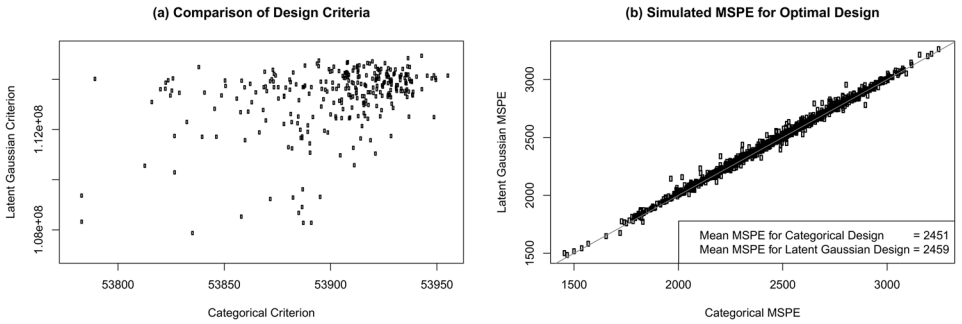

FIG. 3. *Simulation study comparing the categorical* (ϕ_{cat}) *and latent Gaussian* (ϕ_{LG}) *design criteria. The latent Gaussian design criterion* ϕ_{LG} *shows little correlation with* ϕ_{cat} (a). *The optimal designs from Section* 5.4 *were compared by simulating categorical observations on the entire lek network and comparing predictions from the optimal design with actual simulated observations. While the optimal categorical design results in better predictions than the latent Gaussian design* (b), *the difference is slight.*

6. Discussion. We have proposed a multinomial model with latent spatial random effects for microsatellite allele data and illustrated how optimal retrospective spatial designs can be obtained for landscape genetic studies. We defined spatial covariance between leks based on a graphical lek network, with edge weights a function of Euclidean distance between leks. This model was developed to account for the lekking behavior of greater sage-grouse that results in a spatial constellation of breeding nodes [Knick and Hanser (2011)], and will be appropriate for other species which have clearly defined breeding areas or population locations. Alternately, a landscape graph could be specified where the edge weights are defined by the local landscape features [e.g., Cushman and Landguth (2010), Cushman et al. (2006), McRae (2006), McRae and Beier (2007), McRae et al. (2008), Spear et al. (2010)], and connectivity is defined under an isolation by resistance or least-cost path (LCP) approach. Both the IBR and LCP approaches imply a nonstationary spatial covariance, with the nonstationarity a function of landscape characteristics hypothesized to facilitate or impede gene flow. Hanks and Hooten (2013) parameterized the edge weights in an ICAR spatial model based on landscape covariates, and made inference on parameters related to the resistance of different landscape covariates using a Bayesian approach.

We have not included fixed covariate effects in our latent model (2), rather we have modeled only population-level allele prevalence ($\mu_{\ell k}$) and latent spatial random variation $\eta_{s\ell k}$. If the loci used in the analysis are from genes influenced by selection for specific landscape features, then a linear predictor $\mathbf{x}_s'\boldsymbol{\beta}_{\ell k}$ could be included in the mean of (2), where $\mathbf{x}_s$ is a vector containing landscape characteristics at the sth spatial location, and $\boldsymbol{\beta}_{\ell k}$ are allele-specific regression parameters which provide indications of the effect of selection on each allele. In the case where the loci are selectively neutral, there is no reason to assume that allele frequency would

be linked to the landscape features of the spatial locations where the individuals are observed. Loci could be classified as being selectively neutral if a regression analysis reveals no significant relationships between allele prevalence and the landscape characteristics in $\mathbf{x}$.

We also assumed in (3) that the latent allelic spatial effects $\boldsymbol{\eta}_{la}$ have the same spatial covariance matrix $\boldsymbol{\Psi}(\boldsymbol{\theta})$ for each locus and allele. This implies that the processes driving spatial variation in allele frequencies are the same for each locus. If the loci in question have similar mutation rates, then this assumption is likely to be met, as the remaining microevolutionary processes driving spatial variation in allele frequencies (e.g., mating, survival and movement) should be shared across loci. If mutation rates are highly variable between loci, then (3) could be generalized to allow for loci-specific covariance matrices [e.g., $\boldsymbol{\eta}_{\ell k} \sim N(\mathbf{0}, \boldsymbol{\Psi}_{\ell})$].

While genetic data were collected from 2009 to 2012, we did not include a temporal component in our model. The main assumption implicit in collapsing over time is that the mean lek-specific allele probabilities $\{p_{s\ell k}\}$ are constant from 2009–2012. One could consider allowing the spatial allele probabilities to vary over time, but because no leks were visited more than once, we proceeded with this simplifying assumption. Additionally, before the start of the 2013 season, genotyped feather samples were only available for leks in the southwestern half of the sage-grouse range. When genotype data are available from the northeastern half of the range, the analysis could be performed with more complete observations.

Other considerations reflecting our a priori understanding of the metapopulation structure could also be considered in the sampling design. If population delineation is likely to be small across the range of the species, then regular spatial coverage might be preferred. If directional climate change or habitat change is shifting the species distribution, areas at both the trailing and leading edge may be important to sample, as selective pressures may be greater in these locations. Even in a stable species distribution, edge or peripheral populations tend to have smaller effective population sizes, and thus are subject to more change due to random processes, potentially making them of higher importance for sampling [Schwartz et al. (2003)]. Thus, any process that could cause a lek to be small and subject to drift (e.g., close to energy development or agricultural tillage), or to be large and under unique selective pressures, could be a cause for prioritization of that lek in the sampling process. Similarly, the large geographic range of sage-grouse means that different populations may be subject to different selective pressures. Areas that are in ecologically unique habitats may be important to sample to assess if unique genetic adaptations (e.g., local adaptation) have occurred. Cryptic species, subspecies or distinct populations segments can be found by sampling these areas. The design criteria proposed in this manuscript focus on minimizing the global uncertainty in our understanding of population genetics. Future work will consider local design criteria that emphasize understanding changes on the fringes of a population.

The use of a network model assumes that we know the location of all sage-grouse leks across the study region. In many situations, it would be more realistic to assume that there are subpopulations at many unknown locations across

the study region. In this case, the latent spatial processes could be modeled using a continuous Gaussian random field instead of the discrete-space network model used in this paper. Nonstationary spatial covariance could be modeled using spatial deformation approaches [e.g., Schmidt and O'Hagan (2003)] or convolution-based covariance functions [e.g., Calder and Cressie (2007)]. Hierarchical spatial modeling of landscape genetic data provides new opportunities to provide inference for landscape effects on gene flow within a formal statistical framework.

Acknowledgments. We appreciate the collaborative effort by 11 states in the Western Association of State Agencies in contributing sage-grouse data and expertise, and collecting feathers for this study. We are also grateful for the helpful suggestions provided by Karen Kafadar, an anonymous Associate Editor and two anonymous reviewers. Their careful reading of this manuscript strengthened it in many ways. Any use of trade, firm or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

SUPPLEMENTARY MATERIAL

Full conditional distributions (DOI: 10.1214/16-AOAS929SUPP; .pdf). Here we provide full conditional distributions for all parameters in the hierarchical statistical model with latent reduced-rank spatial random effects.

REFERENCES

CUSHMAN, S. A. and LANDGUTH, E. L. (2010). Scale dependent inference in landscape genetics. *Landsc. Ecol.* **25** 967–979.

ALBERT, J. H. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88** 669–679. MR1224394

BERLINER, L. M., WIKLE, C. K. and CRESSIE, N. (2000). Long-lead prediction of Pacific SSTs via Bayesian dynamic modeling. *J. Climate* **13** 3953–3968.

BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **36** 192–236. With discussion by D. R. Cox, A. G. Hawkes, P. Clifford, P. Whittle, K. Ord, R. Mead, J. M. Hammersley, and M. S. Bartlett and with a reply by the author. MR0373208

BESAG, J. and KOOPERBERG, C. (1995). On conditional and intrinsic autoregressions. *Biometrika* **82** 733–746. MR1380811

CALDER, C. A. and CRESSIE, N. (2007). Some topics in convolution-based spatial modeling. In *Proceedings of the* 56*th Session of the International Statistics Institute* 22–29. Instituto Nacional de Estatistica, Lisbon, Portugal.

CONNELLY, J. W., HAGEN, C. A. and SCHROEDER, M. A. (2011). Characteristics and dynamics of greater sage-grouse populations. In *Greater Sage-Grouse*: *Ecology and Conservation of a Landscape Species and Its Habitats*. *Studies in Avian Biology* **38** 53–67. Univ. California Press, Oakland.

CONNELLY, J. W., SCHROEDER, M. A., SANDS, A. R. and BRAUN, C. E. (2000). Guidelines to manage sage-grouse populations and their habitats. *Wildl. Soc. Bull.* **28** 967–985.

CRESSIE, N. A. C. (1993). *Statistics for Spatial Data*. Wiley, New York. MR1239641

CROOKS, K. R. and SANJAYAN, M. A. (2006). *Connectivity Conservation*. Cambridge Univ. Press, Cambridge.

CUSHMAN, S. A., MCKELVEY, K. S., HAYDEN, J. and SCHWARTZ, M. K. (2006). Gene flow in complex landscapes: Testing multiple hypotheses with causal modeling. *Amer. Nat.* **168** 486–499.

DALKE, P. D., PYRAH, D. B., STANTON, D. C., CRAWFORD, J. E. and SCHLATTERER, E. F. (1963). Ecology, productivity, and management of sage-grouse in Idaho. *J. Wildl. Manag.* **27** 811–841.

DIGGLE, P. J. and RIBEIRO, P. J. JR. (2007). *Model-Based Geostatistics*. Springer, New York. MR2293378

DURAND, E., JAY, F., GAGGIOTTI, O. E. and FRANÇOIS, O. (2009). Spatial inference of admixture proportions and secondary contact zones. *Mol. Biol. Evol.* **26** 1963–1973.

FLEGAL, J. M., HARAN, M. and JONES, G. L. (2008). Markov chain Monte Carlo: Can we trust the third significant figure? *Statist. Sci.* **23** 250–260. MR2516823

GARTON, E. O., CONNELLY, J. W., HORNE, J. S., HAGEN, C. A., MOSER, A. and SCHROEDER, M. A. (2011). Greater sage-grouse population dynamics and probability of persistence. In *Greater Sage-Grouse*: *Ecology and Conservation of a Landscape Species and Its Habitats*. *Studies in Avian Biology* **38** 293–381. Univ. California Press, Oakland.

GENTLE, J. E. (2007). *Matrix Algebra*: *Theory*, *Computations*, *and Applications in Statistics*. Springer, New York. MR2337395

GENZ, A. and BRETZ, F. (2009). *Computation of Multivariate Normal and t Probabilities*. *Lecture Notes in Statistics* **195**. Springer, Dordrecht. MR2840595

GUILLOT, G., ESTOUP, A., MORTIER, F. and COSSON, J. F. (2005). A spatial statistical model for landscape genetics. *Genetics* **170** 1261–1280.

GUILLOT, G., LEBLOIS, R., COULON, A. and FRANTZ, A. C. (2009). Statistical methods in spatial genetics. *Mol. Ecol.* **18** 4734–4756.

HANKS, E. M. and HOOTEN, M. B. (2013). Circuit theory and model-based inference for landscape connectivity. *J. Amer. Statist. Assoc.* **108** 22–33. MR3174600

HANKS, E. M., HOOTEN, M. B., KNICK, S. T., OYLER-MCCANCE, S. J., FIKE, J. A., CROSS, T. B. and SCHWARTZ, M. K. (2016). Supplement to "Latent spatial models and sampling design for landscape genetics." DOI:10.1214/00-AOAS929SUPP.

HARVILLE, D. A. (2008). *Matrix Algebra from a Statistician's Perspective*. Springer, New York.

HOOTEN, M. B., WIKLE, C. K., SHERIFF, S. L. and RUSHIN, J. W. (2009). Optimal spatio-temporal hybrid sampling designs for ecological monitoring. *J. Veg. Sci.* **20** 639–649.

KLEIN, D. J. and RANDIĆ, M. (1993). Resistance distance. *J. Math. Chem.* **12** 81–95. Applied graph theory and discrete mathematics in chemistry (Saskatoon, SK, 1991). MR1219566

KNICK, S. T. and HANSER, S. E. (2011). Connecting pattern and process in greater sage-grouse populations and sagebrush landscapes. In *Greater Sage-Grouse*: *Ecology and Conservation of a Landscape Species and Its Habitats*. *Studies in Avian Biology* **38** 383–406. Univ. California Press, Oakland.

MARSAGLIA, G. (1963). Expressing the normal distribution with covariance matrix $A + B$ in terms of one with covariance matrix A. *Biometrika* **50** 535–538. MR0181061

MCKELVEY and SCHWARTZ (2005). dropout: A program to identify problem loci and samples for noninvasive genetic samples in a capture-mark-recapture framework. *Molecular Ecology Notes* **5** 716–718.

MCRAE, B. H. (2006). Isolation by resistance. *Evolution* **60** 1551–1561.

MCRAE, B. H. and BEIER, P. (2007). Circuit theory predicts gene flow in plant and animal populations. *Proc. Natl. Acad. Sci. USA* **104** 19885–19890.

MCRAE, B. H., DICKSON, B. G., KEITT, T. H. and SHAH, V. B. (2008). Using circuit theory to model connectivity in ecology, evolution, and conservation. *Ecology* **89** 2712–2724.

PATTERSON, R. L. (1952). *The Sage-Grouse in Wyoming*. Sage Books, Los Angeles.

ROYLE, J. (1998). An algorithm for the construction of spatial coverage designs with implementation in S-PLUS. *Comput. Geosci.* **24** 479–488.

SCHMIDT, A. M. and O'HAGAN, A. (2003). Bayesian inference for non-stationary spatial covariance structure via spatial deformations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 743–758. MR1998632

SCHROEDER, M. A., ALDRIDGE, C. L., APA, A. D., BOHNE, J. R., BRAUN, C. E., BUNNELL, S. D., CONNELLY, J. W., DEIBERT, P. A., GARDNER, S. C., HILLIARD, M. A. et al. (2004). Distribution of sage-grouse in North America. *Condor* **106** 363–376.

SCHWARTZ, M. K., MILLS, L. S., ORTEGA, Y., RUGGIERO, L. F. and ALLENDORF, F. W. (2003). Landscape location affects genetic variation of Canada lynx (*Lynx canadensis*). *Mol. Ecol.* **12** 1807–1816.

SELANDER, R. K. (1970). Behavior and genetic variation in natural populations. *Am. Zool.* **10** 53–66.

SLATKIN, M. (1987). Gene flow and the geographic structure of natural populations. *Science* **236** 787–792.

SMITH, J. T., FLAKE, L. D., HIGGINS, K. F., KOBRIGER, G. D. and HOMER, C. G. (2005). Evaluating lek occupancy of greater sage-grouse in relation to landscape cultivation in the Dakotas. *West. N. Am. Nat.* **65** 310–320.

SMOUSE, P. E. and PEAKALL, R. (1999). Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. *Heredity* **82** 561–573.

SOKAL, R. R., ODEN, N. L. and WILSON, C. (1991). Genetic evidence for the spread of agriculture in Europe by demic diffusion. *Nature* **351** 143–145.

SPEAR, S. F., BALKENHOL, N., FORTIN, M. J., MCRAE, B. H. and SCRIBNER, K. (2010). Use of resistance surfaces for landscape genetic studies: Considerations for parameterization and analysis. *Mol. Ecol.* **19** 3576–3591.

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. and VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 583–639. MR1979380

TAYLOR, P. D., FAHRIG, L., HENEIN, K. and MERRIAM, G. (1993). Connectivity is a vital element of landscape structure. *Oikos* **68** 571–573.

WIKLE, C. K. and CRESSIE, N. (1999). A dimension-reduced approach to space-time Kalman filtering. *Biometrika* **86** 815–829. MR1741979

WIKLE, C. K. and ROYLE, J. A. (2005). Dynamic design of ecological monitoring networks for non-Gaussian spatio-temporal data. *Environmetrics* **16** 507–522. MR2147540

WILEY, R. H. (1973). Territoriality and non-random mating in sage-grouse, *Centrocercus urophasianus*. *Anim. Behav. Monogr.* **6** 87–169.

ZIMMERMAN, D. L. (2006). Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction. *Environmetrics* **17** 635–652. MR2247174

E. M. HANKS
DEPARTMENT OF STATISTICS
PENNSYLVANIA STATE UNIVERSITY
UNIVERSITY PARK, PENNSYLVANIA 16802
USA
E-MAIL: hanks@psu.edu

M. B. HOOTEN
U.S. GEOLOGICAL SURVEY
COLORADO COOPERATIVE FISH
AND WILDLIFE RESEARCH UNIT
DEPARTMENT OF FISH, WILDLIFE,
AND CONSERVATION BIOLOGY
DEPARTMENT OF STATISTICS
COLORADO STATE UNIVERSITY
FORT COLLINS, COLORADO 80523
USA

S. T. KNICK
U.S. GEOLOGICAL SURVEY
FOREST AND RANGELAND ECOSYSTEM
SCIENCE CENTER
BOISE, IDAHO 83706
USA

S. J. OYLER-MCCANCE
J. A. FIKE
FORT COLLINS SCIENCE CENTER
2150 CENTRE AVE BLDG C
FORT COLLINS, COLORADO 80526
USA

T. B. CROSS
M. K. SCHWARTZ
USFS ROCKY MOUNTAIN RESEARCH STATION
NATIONAL GENOMICS CENTER FOR WILDLIFE
AND FISH CONSERVATION
800 E. BECKWITH AVE
MISSOULA, MONTANA 59801
USA