

# LATENT SPATIAL MODELS AND SAMPLING DESIGN FOR LANDSCAPE GENETICS

BY EPHRAIM M. HANKS,<sup>\*</sup> MEVIN B. HOOTEN,<sup>†,‡,§</sup> STEVEN T. KNICK,<sup>¶</sup>  
SARA J. OYLER-McCANCE,<sup>||</sup> JENNIFER A. FIKE,<sup>||</sup> TODD B.  
CROSS,<sup>\*\*,††</sup> AND MICHAEL K. SCHWARTZ,<sup>\*\*</sup>

*Department of Statistics, The Pennsylvania State University<sup>\*</sup>, U. S.  
Geological Survey, Colorado Cooperative Fish and Wildlife Research Unit<sup>†</sup>,  
Department of Fish, Wildlife, and Conservation Biology, Colorado State  
University<sup>‡</sup>, Department of Statistics, Colorado State University<sup>§</sup>, U. S.  
Geological Survey, Forest and Rangeland Ecosystem Science Center, Boise,  
ID, U.S.A.<sup>¶</sup>, U. S. Geological Survey, Fort Collins Science Center, Fort  
Collins, CO, U.S.A.<sup>||</sup> U. S. Forest Service, Rocky Mountain Research  
Station, Missoula, MT, U.S.A.<sup>\*\*</sup>, Wildlife Biology Program, University of  
Montana, Missoula, MT, U.S.A.<sup>††</sup>*

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    **1. Introduction.** Landscape connectivity is “the degree to which the  
2    landscape facilitates or impedes movement among resource patches” (Taylor  
3    et al., 1993). Connectivity among subpopulations or habitat patches is  
4    an important factor in maintaining biodiversity, understanding the spread  
5    of infectious diseases, and allocating resources for conservation. Many man-  
6    agement efforts devote considerable resources to maintaining or restoring  
7    landscape connectivity (e.g., Crooks and Sanjayan, 2006). Natural and an-  
8    thropogenic disturbance can alter the landscape, affecting connectivity and  
9    increasing uncertainty about the functioning of current ecosystems. Data  
10    driven decision-making can lead to effective management of species and  
11    ecosystems.

Spatially-referenced genetic data are commonly used to study functional connectivity (e.g., Slatkin, 1987; Guillot et al., 2005; Cushman et al., 2006; McRae, 2006; Durand et al., 2009). 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 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., Selander, 1970; Smouse and Peakall, 1999; Sokal, Oden and Wilson, 1991; Guillot et al., 2009; Hanks and Hooten, 2013). 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 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. (Patterson, 1952; Connelly et al., 2000). 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; Wiley, 1973; Smith et al., 2005).

Greater sage-grouse were historically widely distributed across 12 western states in the U.S. and three Canadian provinces, but have undergone both a

51 population decline (Garton et al., 2011) and a range contraction (Schroeder  
52 et al., 2004). Quantifying the distribution of sage-grouse genetic variability  
53 across the range of the species could be used to identify population struc-  
54 ture and characteristics of gene flow that provide insight into regions where  
55 management actions could have the greatest impact.

56 *2.1. Sample Collection.* From 2009 to 2012, sage-grouse feathers were  
57 collected from sage-grouse leks during the spring breeding season. An op-  
58 portunistic sample of all the feathers observed on the lek was collected by  
59 collecting feathers observed while traversing the lek. Feathers were placed  
60 in paper envelopes with the spatial coordinates for the lek. Following collec-  
61 tion, envelopes were stored in cool, dry facilities out of direct sunlight. All  
62 envelopes and feathers collected from any given lek were pooled by lek, as it  
63 was impossible to determine before a lab analysis which feathers could have  
64 come from the same sage-grouse.

65 *2.2. Sample Extraction and Genotyping.* DNA was extracted from feath-  
66 ers using QIAGENs DNeasy Blood and Tissue Kit and the user developed  
67 protocol for purification of total DNA from nails, hair, or feathers. Each  
68 sample was genotyped across a panel of 14 variable microsatellite loci, all of  
69 which have been redesigned specifically to optimize efficacy with low quality  
70 and quantity DNA acquired from non-invasively collected feather samples.

71 Each sample was amplified multiple times at each microsatellite locus  
72 to screen for genotyping error (e.g., allelic drop out, false alleles, scoring  
73 error, etc.). If there were any variation between the genotypes generated for  
74 each sample, samples were genotyped another two to six times to confirm  
75 genotype accuracy. As genetic material was coming from feathers, many of  
76 which were significantly weathered, and some samples failed at some loci. We  
77 removed from the analysis any individual where samples failed at more than  
78 1/3 of the loci. The program DROPOUT (McKelvey and Schwartz 2005)  
79 was used to identify duplicate sample genotypes and screen for genotyping  
80 error.

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

86 **3. Latent Spatial Models for Landscape Genetics.** A microsatel-  
87 lite is a portion of repetitive DNA in which certain short DNA sequences  
88 are repeated multiple times. In transcription, it is possible for the number

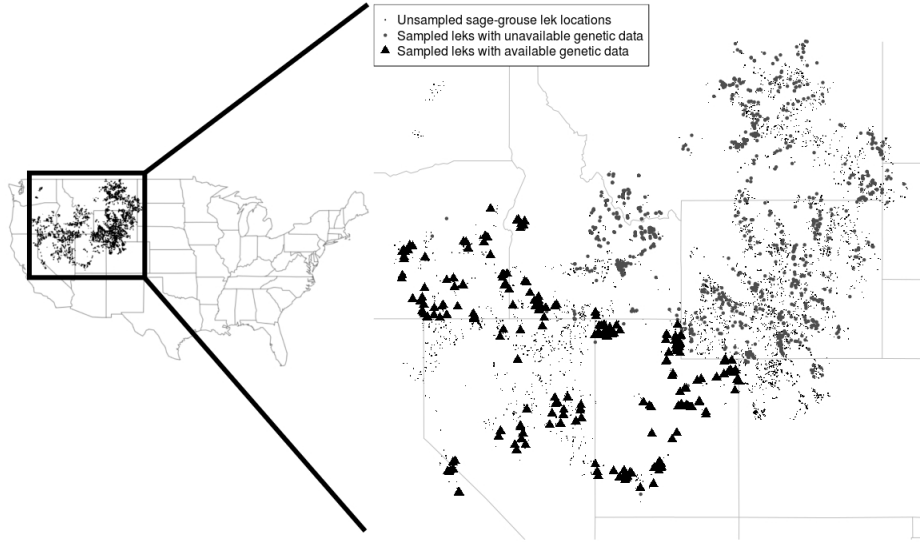


FIG 1. Sage-grouse lek locations in the western United States. Feather samples were collected from 1177 sage-grouse leks between 2009 and 2012. Microsatellite allele data from 243 of these leks were available for analysis in March, 2013.

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  $\ell^{\text{th}}$  locus, for  $\ell = 1, 2, \dots, L$ , we denote the list of all distinct observed alleles from all individuals in the study as  $\{a_{\ell 1}, a_{\ell 2}, \dots, a_{\ell K_\ell}\}$ . 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  $\mathbf{p}_{s\ell} \equiv (p_{s\ell 1} \ p_{s\ell 2} \ \dots \ p_{s\ell K_\ell})'$ , where  $s \in \{1, 2, \dots, S\}$  indexes the spatial location.

Let  $\mathbf{y}_{sipel}$  be a vector coded such that the  $k$ -th entry  $y_{sipelk} = 1$  if the  $p^{\text{th}}$  (indexing ploidy) observed allele at the  $\ell^{\text{th}}$  locus is  $a_{\ell k}$  for the  $i^{\text{th}}$  individual sage-grouse at the  $s^{\text{th}}$  spatial lek location, and  $y_{sipelk} = 0$  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

$$(1) \quad y_{sipelk} = \begin{cases} 1 & , \ z_{sipelk} = \max\{z_{sipel a}, \ a = 1, \dots, K_\ell\} \\ 0 & , \text{ otherwise} \end{cases}$$

where

$$(2) \quad z_{sipelk} \stackrel{iid}{\sim} N(\mu_{\ell k} + \eta_{s\ell k}, 1).$$

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  $p_{s\ell k} = P(z_{sipelk} = \max\{z_{sipel a}, \ a = 1, \dots, K_\ell\})$ .

The mean of the latent variable  $z_{sipelk}$  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  $k^{\text{th}}$  allele at the  $\ell^{\text{th}}$  locus across the entire population being studied. Large values of  $\mu_{\ell k}$ , relative to  $\mu_{\ell k'}$  make it more likely that  $z_{sipelk}$  will be larger than  $z_{sipelk'}$ , and so the  $k^{\text{th}}$  allele will be more prevalent than the  $(k')^{\text{th}}$  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  $z_{sipelk} - z_{sipelk'}$ , and not the actual values of  $z_{sipelk}$ . Thus, if  $\mu_{\ell k}$  were replaced by  $\mu_{\ell k} + c$  for  $k = 1, 2, \dots, K_\ell$  and some constant  $c$ , the likelihood of the observed allele data would remain unchanged. To maintain model identifiability, we fix  $\mu_{\ell 1} = 0$  for  $\ell = 1, 2, \dots, L$ , 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

$$(3) \quad \boldsymbol{\eta}_{\ell k} = [\eta_{1\ell k} \ \eta_{2\ell k} \ \dots \ \eta_{n\ell k}]' \sim N(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta})),$$

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

139 We adopt a Bayesian approach to modelling, and specify prior distribu-  
 140 tions for all parameters in (1)-(3)

$$(4) \quad \mu_{\ell k} \sim N(0, \sigma_{\mu}^2), \ell = 1, 2, \dots, L, k = 2, 3, \dots, K_{\ell},$$

141

$$(5) \quad \boldsymbol{\theta} \sim [\boldsymbol{\theta}],$$

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

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

#### 155 4. Sage-Grouse Landscape Genetic Analysis.

156 4.1. *Lek Network Connectivity Model.* In the previous section we spec-  
 157 ified a multinomial model for spatially referenced microsatellite allele data  
 158 with latent spatial autocorrelation. Critical to our approach is the choice of  
 159 covariance matrix  $\boldsymbol{\Sigma}(\boldsymbol{\theta})$  for the latent spatial random effects in (3). Leks are  
 160 persistent centers of sage-grouse breeding activity, and are thus essential to  
 161 gene flow and spatial genetic variation.

162 We will consider a covariance that is based on the isolation by resistance  
 163 (IBR) approach of [McRae \(2006\)](#) and takes into account the importance  
 164 of leks in sage-grouse gene flow, similar in spirit to the model of [Knick  
 165 and Hanser \(2011\)](#). The IBR approach envisions space as a graph: a set of  
 166 spatial nodes (leks in our case) connected by edges, where the edges are  
 167 resistors in an electric circuit. [McRae \(2006\)](#) showed that if the resistance  
 168 (edge weights) of the resistors was inversely proportional to migration rates  
 169 between nodes, and migration could be modeled as a random walk on the  
 170 graph of nodes, then the resistance distance of the graph ([Klein and Randić,  
 171 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  $s^{\text{th}}$  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  $s^{\text{th}}$  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  $s^{\text{th}}$  and  $t^{\text{th}}$  leks as  $\alpha_{st}/\sigma^2$ , where  $\alpha_{st}$  is a decreasing function of the Euclidean distance between the leks and  $\sigma^2$  is a scaling factor. Under the random walk model for migration in the IBR framework,  $\alpha_{st}/\sigma^2$  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  $d_{\text{MAX}}$ :

$$(6) \quad \alpha_{st} = \begin{cases} f(d_{st}) & , d_{st} \leq d_{\text{MAX}} \\ 0 & , d_{st} > d_{\text{MAX}} \end{cases}.$$

We will consider multiple functional forms for  $f(d)$  and multiple maximum distances  $d_{\text{MAX}}$  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  $d_{\text{MAX}}$  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

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

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  $\ell$  and allele  $k$  subscripts)  
and precision matrix  $\mathbf{Q}$  can be partitioned accordingly

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

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

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

We assign a conjugate inverse gamma prior distribution to  $\sigma^2$ :

$$(9) \quad \sigma^2 \sim IG(r, q)$$

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  $\sigma^2$ ) 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 6,000 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 compu-  
tationally 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 de-  
composition 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  $\{\lambda_1, \lambda_2, \dots, \lambda_{n_s}\}$  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:

$$(10) \quad \tilde{\eta}_{s\ell k} = \tilde{\mathbf{m}}'_s \tilde{\boldsymbol{\delta}}_{\ell k}$$

236

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

where  $\tilde{\mathbf{m}}'_s$  is the  $s^{\text{th}}$  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

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

Full-conditional distributions are available for all parameters in this hierarchical model, and are given in the Supplement.

**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 18km, 25km, and 50km 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 network model with edge weight between leks inversely proportional to the square of the distance between leks ( $f_1(d) = 1/d^2$ ) and  $d_{MAX} = 25\text{km}$ . 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 25km, conditional on the random effects  $\eta_{slk}$  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 ( $\xi$ ) in  $f(d) = d^{-\xi}$  in (6) and jointly estimate the posterior distribution of  $\xi$  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  $\sigma^2$  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  $\sigma^2$  is 0.0011. This provides some confidence that our MCMC algorithm has converged to the stationary posterior distribution.

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

$$z_{siplk} = \mu_{\ell k} + \tilde{\eta}_{slk} + \epsilon_{siplk} \quad , \quad \epsilon_{siplk} \sim N(0, 1)$$

where  $\tilde{\eta}_{\ell k} \sim N(\mathbf{0}, \tilde{\Sigma}(\sigma^2))$ . The spatial covariance matrix  $\tilde{\Sigma}(\sigma^2)$  is nonstationary, and the average of the posterior mean diagonal elements of  $\tilde{\Sigma}(\sigma^2)$  is 8.43, which is much greater than the unit variance of the non-spatial noise (nugget) contained in  $\epsilon_{siplk}$  of (2) leading us to conclude that spatial variability in the genetic data accounts for a much larger proportion of the variability than does non-spatial (within lek) variability. Summaries of posterior distributions for  $\{\mu_{\ell k}, \ell = 1, 2, \dots, 14, k = 1, 2, \dots, K_{\ell}\}$  are not shown, but correlate well with empirical allele frequencies taken over all 1,136 genetic samples.

TABLE 1  
*Comparison of Spatial Lek-Network Models Based on DIC.*

Model	$f(d)$	$d_{MAX}$	DIC
1	$1/d^2$	25000.00	661844
2	$1/d$	50000.00	662409
3	$1/d$	25000.00	664848
4	$1/d^2$	50000.00	670883
5	$1/d$	18000.00	671375
6	$1/d^2$	18000.00	674832

288 **5. Optimal Retrospective Sampling for Sage-Grouse in the West-**  
 289 **ern United States.** Having fit the model to microsatellite allele data from  
 290 sage-grouse feather samples collected in 2009-2012, we now consider retro-  
 291 spective sampling design for this system. Our goal is to recommend optimal  
 292 regions for sampling previously unsampled sage-grouse leks, given what is  
 293 known from the spatially referenced genetic data already collected. This re-  
 294 quires the definition of a criterion  $\phi(\mathbf{d}_j)$  by which we can compare  $J$  potential  
 295 sampling designs  $(\mathbf{d}_j, j = 1, \dots, J)$ .

296 **5.1. Latent Gaussian Design.** A common criterion to minimize for opti-  
 297 mal sampling of Gaussian random variables is the mean squared prediction  
 298 error (MSPE) at unobserved locations (e.g., [Zimmerman, 2006](#)). The obser-  
 299 vations  $\{y_{sipelk}\}$  in our model are not Gaussian; rather, they are categorical  
 300 alleles which we model as arising from a multinomial distribution. However,  
 301 each  $z_{sipelk}$  is a latent representation of the true observation  $y_{sipelk}$ , and  $z_{sipelk}$   
 302 is normally distributed such that

$$z_{sipelk} = \mu_{\ell k} + \eta_{s\ell k} + \epsilon_{sipelk} \quad , \quad \epsilon_{sipelk} \sim N(0, 1).$$

303 This leads us to consider the MSPE of the latent  $z$  variables as a design cri-  
 304 terion. Full posterior predictive inference on this MSPE would be computa-  
 305 tionally prohibitive, as it would require integrating over the entire posterior  
 306 distribution. Instead, in what follows, we will consider the parameters  $\{\mu_{\ell k}\}$   
 307 and  $\sigma^2$  to be fixed and known, and we will focus on finding optimal designs  
 308 conditional on the posterior mean values of model parameters. Our devel-  
 309 opment of a design criterion first focuses on design for the latent Gaussian  
 310  $\mathbf{z}$ , and is reminiscent of Kriging (e.g., [Cressie, 1993](#)). We will then consider  
 311 a design criterion that takes into account the categorical nature of the data  
 312 by integrating over the latent Gaussian  $\mathbf{z}$ .

313 For the  $k$ -th allele at the  $\ell$ -th locus, we can divide the spatial loca-  
 314 tions, and their corresponding  $z$  variables, into observed  $(\mathbf{z}_{\ell k}^{(o)} = \{z_{sipelk} :$

315  $s$  is observed } ) and unobserved ( $\mathbf{z}_{\ell k}^{(u)} = \{z_{s\ell k} : s \text{ is unobserved}\}$ ) cate-  
 316 gories. The “observed” set corresponds to the genetic samples obtained from  
 317 the sage-grouse leks, while the “unobserved” set will be a single unsampled  
 318 sage-grouse at each unsampled lek. The joint distribution of  $\mathbf{z}_{\ell k}$  at all leks  
 319 is Gaussian

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

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

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

$$(13) \quad \hat{\mathbf{z}}_{\ell k}^{(u)} = \mu_{\ell k} \mathbf{1} + \Psi_{uo} \Psi_{oo}^{-1} (\mathbf{z}_{\ell k}^{(o)} - \mu_{\ell k} \mathbf{1})$$

324 and the corresponding conditional Gaussian covariance matrix is given by:

$$(14) \quad \Phi = E [(\mathbf{z}_u - \hat{\mathbf{z}}_u)(\mathbf{z}_u - \hat{\mathbf{z}}_u)'] = \Psi_{uu} - \Psi_{uo} \Psi_{oo}^{-1} \Psi_{ou}.$$

325 Each diagonal entry of  $\Phi$  contains the MSPE for  $z$  at an unobserved location,  
 326 and we choose the sum of these entries (“A-optimality” in [Harville \(2008\)](#))  
 327 for our latent Gaussian design criterion:

$$(15) \quad \phi_{\text{LG}}(\mathbf{d}) = \text{tr}(\Phi) = \text{sum}(\text{diag}(\Phi)).$$

328 Other possible alternatives include “D-optimality”, in which the design cri-  
 329 terion consists of the product of the diagonal elements of  $\Phi$ . While D-  
 330 optimality has some desirable properties when considering optimal covariate  
 331 designs, we focus on A-optimality which is the prevailing approach in spa-  
 332 tial sampling (e.g., [Wikle and Royle, 2005](#); [Zimmerman, 2006](#); [Hooten et al.,](#)  
 333 [2009](#)), and is interpretable with the design criterion being the mean square  
 334 prediction error averaged over all unobserved spatial locations.

335 It is notable that neither the covariance matrix  $\Phi$  or the design criterion  
 336  $\phi_{\text{LG}}(\mathbf{d})$  depend on  $\mathbf{z}_{\ell k}^{(o)}$ . Additionally, this design criterion would be identical  
 337 for any locus  $\ell$  and allele  $k$ , as we assume in (3) that the spatial correlation  
 338 is shared by all latent allelic processes  $\{\mathbf{z}_{\ell k}\}$ . As the design criterion  $\phi_{\text{LG}}$   
 339 depends only on the covariance matrix  $\Phi$ , computation of this design cri-  
 340 terion is very efficient, as changing a design (e.g., adding a new lek to the  
 341 set of previously sampled leks) only involves evaluating (14) and (15) for a  
 342 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  $\ell$  for an individual sage-grouse at an unsampled spatial location  $u$ . That is, we want to predict  $\mathbf{y}_{u\ell} = (y_{u11\ell 1}, \dots, y_{u11\ell K})'$ , 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  $(z_{u11\ell 1}, \dots, z_{u11\ell K})$ . As we did in Section 5.1, we will consider prediction and sampling design conditioned on the posterior mean values of  $\{\mu_{\ell k}\}$ ,  $\sigma^2$ , 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  $\{\mathbf{z}_{\ell 1}^{(o)}, \dots, \mathbf{z}_{\ell K_\ell}^{(o)}\}$ , then the mean of the latent Gaussian  $(z_{u11\ell 1}, \dots, z_{u11\ell K_\ell})$  is  $(\hat{z}_{u11\ell 1}, \dots, \hat{z}_{u11\ell K_\ell})$ , with each entry obtained using (13). Our prediction of  $\mathbf{y}_{u\ell} | \{\mathbf{z}_{\ell 1}^{(o)}, \dots, \mathbf{z}_{\ell K_\ell}^{(o)}\}$  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  $\mathbf{y}_{u\ell} | \{\mathbf{z}_{\ell 1}^{(o)}, \dots, \mathbf{z}_{\ell K_\ell}^{(o)}\}$  is equal to the probability that the index  $k_{u\ell}$  of the maximum entry of  $(z_{u11\ell 1}, \dots, z_{u11\ell K_\ell})$  is different from the index  $\hat{k}_{u\ell}$  of the maximum entry of  $(\hat{z}_{u11\ell 1}, \dots, \hat{z}_{u11\ell K_\ell})$ .

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  $\psi_u^2$  is the  $u^{\text{th}}$  diagonal element of  $\Phi$  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  $\mathbf{z}_{u\ell} = (z_{u11\ell 1}, z_{u11\ell 2}, \dots, z_{u11\ell K_\ell})'$  is distributed

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

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:

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

373 which is an orthant probability of the multivariate normal  $\mathbf{w}_{u\ell} \sim N(\mathbf{L}\hat{\mathbf{z}}_{u\ell}, \psi_u^2 \mathbf{L}\mathbf{L}')$ ,  
 374 defined as  
 (17)

$$\mathbf{w}_{u\ell} = \begin{bmatrix} w_{u11\ell1} \\ w_{u11\ell2} \\ \vdots \\ w_{u11\ell K_\ell-1} \end{bmatrix} = \mathbf{L}\mathbf{z}_{u\ell} = \begin{bmatrix} 1 & 0 & 0 & \dots & -1 \\ 0 & 1 & 0 & \dots & -1 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \dots & 1 & -1 \end{bmatrix} \begin{bmatrix} z_{u11\ell1} \\ z_{u11\ell2} \\ \vdots \\ z_{u11\ell K_\ell} \end{bmatrix}$$

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

(18)

$$\phi_{\text{cat}}(\mathbf{d}) = \sum_u \sum_\ell \theta_{u\ell} = \sum_u \sum_\ell (1 - P(w_{u11\ell1} < 0, w_{u11\ell2} < 0, \dots, w_{u11\ell K_\ell-1} < 0)).$$

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

380 The calculation of each of the multivariate normal orthant probabilities  
 381  $\{\theta_{u\ell}\}$  is simplified as the correlation matrix can be written as  $\mathbf{L}\mathbf{L}' = \mathbf{I} + \mathbf{1}\mathbf{1}'$ .  
 382 In this special case, each multivariate orthant probability can be expressed  
 383 as a one-dimensional integral on  $[0,1]$  (Marsaglia, 1963; Genz and Bretz,  
 384 2009, pp3,16-17)

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

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

388 **5.3. Optimal Sampling for Reduced Rank Spatial Models.** Computing the  
 389 design criteria (15) and (18) requires calculation of the inverse  $\Psi_{oo}^{-1}$  which  
 390 has computational cost of  $O(n_d^3)$ , where  $n_d$  is the number of locations in the  
 391 proposed design. In the sage-grouse study,  $n_d = 1177$  leks were sampled dur-  
 392 ing the years of 2009-2012. For our retrospective design, we consider adding  
 393 leks to the sampling design, which will increase this number. Inverting a  
 394 matrix of this size is computationally demanding, especially if we wish to  
 395 compare a large number of potential sampling designs. However, we can take  
 396 advantage of the reduced rank spatial model presented in Section 4.2 to sig-  
 397 nificantly reduce the computational cost of computing the design criterion  
 398 by replacing the covariance matrix  $\Phi = \Psi_{uu} - \Psi_{uo}\Psi_{oo}^{-1}\Psi_{ou}$  (14) with a  
 399 reduced-rank approximation  $\tilde{\Phi} = \tilde{\Psi}_{uu} - \tilde{\Psi}_{uo}\tilde{\Psi}_{oo}^{-1}\tilde{\Psi}_{ou}$ .

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

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

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

(19)

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

(20)

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

403 This decomposition requires the inverse of a  $n_d \times n_d$  matrix, but the  
 404 inverse in question can be expressed using the Sherman-Morrison-Woodbury  
 405 identity (e.g., [Gentle, 2007](#), p. 221)

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

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

414 *5.4. Recommendations for Retrospective Sage-Grouse Sampling Effort.* To  
 415 make recommendations for retrospective sampling effort of sage-grouse leks,  
 416 we identify optimal retrospective designs under both the latent Gaussian  
 417 MSPE sampling criterion (15) and the categorical sampling criterion (18).  
 418 As described in Section 4.3, genetic data from 830 individuals at 243 dis-  
 419 tinct leks (Figure 1) were used to fit the statistical model proposed in Section  
 420 4.2. Each of these leks were considered to be observed for our purpose of  
 421 identifying an optimal retrospective design. We also included 934 additional  
 422 leks where feather samples had been collected during 2009-2012 at these  
 423 additional leks, but had not been genotyped before the resumption of sam-  
 424 pling. The genetic information from these samples could thus not be used

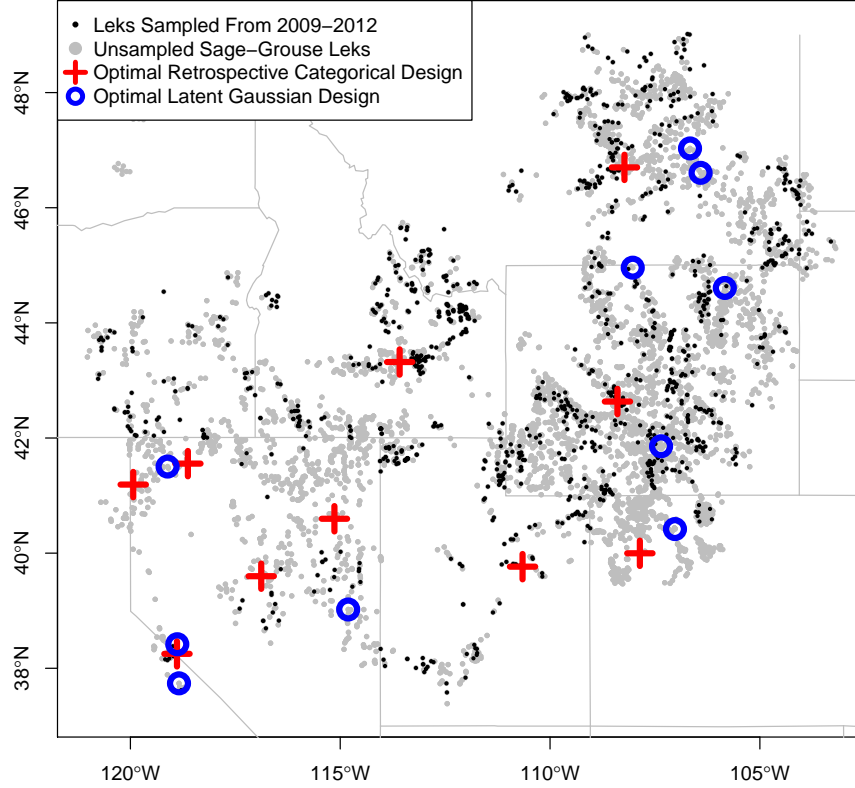


FIG 2. Sage-grouse lek sampling recommendations based on latent Gaussian MSPE. Leks shown as grey circles indicate leks that have not been sampled, while leks shown as black dots have been sampled in 2009–2012. Leks shown as crosses are the optimal categorical retrospective lek sampling locations under 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 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.

---

**Algorithm 1** Sampling Design Switching Algorithm

---

```

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

```

---

We ran the iterative algorithm for  $N_{\text{switch}} = 1000$  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 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  $\phi_{LG}$  and the categorical design criterion  $\phi_{cat}^*$  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 the latent Gaussian  $z_{sipelk}$  at each lek in the optimal design. In practice, one could fix each of these at the posterior mean after model fitting as described in Section 4. For optimal designs, we used the optimal retrospective designs from Section 5.4, which are shown in Figure 2. This will allow us 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.

**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

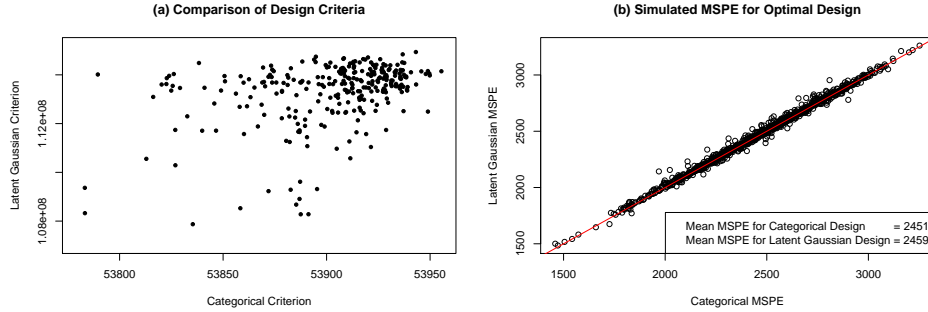


FIG 3. *Simulation study comparing the categorical ( $\phi_{cat}$ ) and latent Gaussian ( $\phi_{LG}$ ) design criteria. The latent Gaussian design criterion  $\phi_{LG}$  shows little correlation with  $\phi_{cat}$  (a). The optimal designs from Section 5.4 were compared by simulating categorical observations 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.*

(e.g., [McRae, 2006](#); [McRae and Beier, 2007](#); [McRae et al., 2008](#); [Cushman et al., 2006](#); [Cushman and Landguth, 2010](#); [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  $s^{\text{th}}$  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  $\eta_{la}$  have

the same spatial covariance matrix  $\Psi(\theta)$  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.,  $\eta_{\ell k} \sim N(\mathbf{0}, \Psi_{\ell})$ ).

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 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.

**Acknowledgements.** 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. Funding for this project was provided by U.S. Geological Survey RWO 98. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

## References.

- ALBERT, J. H. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88** 669–679.
- BERLINER, L. M., WIKLE, C. K. and CRESSIE, N. (2000). Long-lead prediction of Pacific SSTs via Bayesian dynamic modeling. *Journal of Climate* **13** 3953–3968.
- BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)* **36** 192–236.
- BESAG, J. and KOOPERBERG, C. (1995). On conditional and intrinsic autoregressions. *Biometrika* **82** 733.
- CALDER, C. A. and CRESSIE, N. (2007). Some topics in convolution-based spatial modeling. *Proceedings of the 56th Session of the International Statistics Institute* 22–29.
- CONNELLY, J. W., HAGEN, C. A. and SCHROEDER, M. A. (2011). Characteristics and dynamics of greater sage-grouse populations. *Greater sage-grouse: ecology and conservation of a landscape species and its habitats. Studies in Avian Biology*. **38** 53–67.
- CONNELLY, J. W., SCHROEDER, M. A., SANDS, A. R. and BRAUN, C. E. (2000). Guidelines to manage sage-grouse populations and their habitats. *Wildlife Society Bulletin* **28** 967–985.
- CRESSIE, N. (1993). *Statistics for Spatial Data*. Wiley-Interscience.
- CROOKS, K. R. and SANJAYAN, M. (2006). *Connectivity conservation*. Cambridge University Press.
- CUSHMAN, S. A. and LANDGUTH, E. L. (2010). Scale dependent inference in landscape genetics. *Landscape Ecology* **25** 967–979.
- CUSHMAN, S. A., MCKELVEY, K. S., HAYDEN, J. and SCHWARTZ, M. K. (2006). Gene flow in complex landscapes: testing multiple hypotheses with causal modeling. *The American Naturalist* **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. *The Journal of Wildlife Management* **27** 811–841.
- DIGGLE, P. and RIBEIRO, P. J. (2007). *Model-based geostatistics*. Springer.
- DURAND, E., JAY, F., GAGGIOTTI, O. E. and FRANÇOIS, O. (2009). Spatial inference of admixture proportions and secondary contact zones. *Molecular Biology and Evolution* **26** 1963–1973.
- FLEGAL, J. M., HARAN, M. and JONES, G. L. (2008). Markov chain Monte Carlo: Can we trust the third significant figure? *Statistical Science* 250–260.
- 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. *Greater sage-grouse: ecology and conservation of a landscape species and its habitats. Studies in Avian Biology* **38** 293–381.
- GENTLE, J. E. (2007). *Matrix algebra: theory, computations, and applications in statistics*. Springer Verlag.
- GENZ, A. and BRETZ, F. (2009). *Computation of multivariate normal and t probabilities* **195**. Springer Science & Business Media.
- 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. *Molecular Ecology* **18** 4734–56.
- HANKS, E. M. and HOOTEN, M. B. (2013). Circuit Theory and Model-Based Inference for Landscape Connectivity. *Journal of the American Statistical Association* **108** 22–33.
- HARVILLE, D. A. (2008). *Matrix algebra from a statistician's perspective*. Springer-Verlag.
- HOOTEN, M. B., WIKLE, C. K., SHERIFF, S. L. and RUSHIN, J. W. (2009). Optimal spatio-temporal hybrid sampling designs for ecological monitoring. *Journal of Vegetation Science* **20** 639–649.
- KLEIN, D. and RANDIĆ, M. (1993). Resistance distance. *Journal of Mathematical Chemistry* **12** 81–95.
- KNICK, S. T. and HANSER, S. E. (2011). *Greater Sage-Grouse: Ecology and Conservation of a Landscape Species and Its Habitats. Studies in Avian Biology* **Volume 38** Ch 16, pp. 383–406.
- MARSAGLIA, G. (1963). Expressing the normal distribution with covariance matrix  $A+B$  in terms of one with covariance matrix  $A$ . *Biometrika* 535–538.
- 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. *Proceedings of the National Academy of Sciences of the United States of America* **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.
- NAVIDI, W., ARNHEIM, N. and WATERMAN, M. (1992). A multiple-tubes approach for accurate genotyping of very small DNA samples by using PCR: statistical considerations. *American Journal of Human Genetics* **50** 347.
- PATTERSON, R. L. (1952). *The sage-grouse in Wyoming*. Sage Books.
- ROYLE, J. (1998). An algorithm for the construction of spatial coverage designs with implementation in S-PLUS. *Computers & Geosciences* **24** 479–488.
- SCHMIDT, A. M. and O'HAGAN, A. (2003). Bayesian inference for non-stationary spatial covariance structure via spatial deformations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65** 743–758.

- 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. *The 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*). *Molecular Ecology* **12** 1807–1816.
- SELANDER, R. K. (1970). Behavior and genetic variation in natural populations. *American Zoologist* **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. *Western North American Naturalist* **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. *Molecular Ecology* **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. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64** 583–639.
- 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.
- WIKLE, C. K. and ROYLE, J. A. (2005). Dynamic design of ecological monitoring networks for non-Gaussian spatio-temporal data. *Environmetrics* **16** 507–522.
- WILEY, R. H. (1973). Territoriality and non-random mating in sage-grouse, *Centrocercus urophasianus*. *Animal Behaviour Monographs* **6** 87–169.
- ZIMMERMAN, D. L. (2006). Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction. *Environmetrics* **17** 635–652.

## SUPPLEMENT: FULL CONDITIONAL DISTRIBUTIONS

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

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

685 Full-conditional distributions are available for all parameters in this hier-  
 686 archical model. As noted by [Albert and Chib \(1993\)](#), the full conditional for  
 687  $z_{sipelk}$  is truncated normal:

$$z_{sipelk} | \cdot \sim \text{TN}_{\mathcal{D}_{sipelk}}(\mu_{\ell k} + \tilde{\eta}_{s\ell k}, 1)$$

688 where  $\text{TN}_{\mathcal{D}}(\cdot, \cdot)$  indicates a truncated normal distribution with support  $\mathcal{D}$   
 689 and

$$\mathcal{D}_{sipelk} = \begin{cases} (0, \infty) & y_{sipelk} = 1 \\ (-\infty, 0) & y_{sipelk} = 0 \end{cases}.$$

690 The full-conditional distribution for  $\tilde{\boldsymbol{\delta}}_{\ell k}$  is normally distributed:

$$\tilde{\boldsymbol{\delta}}_{\ell k} \sim N(\mathbf{A}^{-1} \mathbf{b}_{\ell k}, \mathbf{A}^{-1})$$

691 where

$$\mathbf{A} = 2\tilde{\mathbf{M}}' \mathbf{K}' \mathbf{K} \tilde{\mathbf{M}} + \frac{1}{\sigma^2} \tilde{\mathbf{D}}^{-1}$$

692 and

$$\mathbf{b}_{\ell k} = \tilde{\mathbf{M}}' \mathbf{K}' (\mathbf{z}_{1\ell k} + \mathbf{z}_{2\ell k} - 2\mu_{\ell k}).$$

693 The matrix  $\mathbf{K}$  is a design matrix linking the individual sage-grouse to the  
 694 lek. Then  $\mathbf{K}' \mathbf{K}$  is a diagonal matrix with the  $s^{\text{th}}$  diagonal entry equal to the  
 695 number of animals sampled at the  $s^{\text{th}}$  lek.

696 The full-conditional distribution for  $\mu_{\ell k}$  is also normally distributed

$$\mu_{\ell k} | \cdot \sim N\left(\frac{h_{\ell k}}{c}, \frac{1}{c}\right)$$



697 with

$$\frac{1}{c} = \frac{1}{\tau^2} + \sum_s \sum_i \sum_p 1$$

698 and

$$h_{\ell k} = \sum_s \sum_i \sum_p (z_{sip\ell k} - \tilde{\eta}_{s\ell k}),$$

699 where the sums are taken over all possible values of  $s$ ,  $i$ , and  $p$ . Thus, for  
 700 example,  $\sum_s \sum_i \sum_p 1$  is the total number of observations at one locus, and  
 701 is equal to two times the sum total number of individuals at all leks, as there  
 702 are two allele observations for each individual (diploid) sage-grouse.

703 The full-conditional distribution for  $\sigma^2$  is inverse-gamma:

$$\sigma^2 | \cdot \sim IG \left( \left( \frac{1}{r} + \frac{1}{2} \sum_{\ell} \sum_k \tilde{\delta}'_{\ell k} \tilde{\mathbf{D}}^{-1} \tilde{\delta}_{\ell k} \right)^{-1}, q + \frac{1}{2} \sum_{\ell} N_{\ell} \right).$$

704

CONTACT:  
 EPHRAIM M. HANKS  
 323 THOMAS BUILDING  
 UNIVERSITY PARK, PA 16802  
 E-MAIL: [hanks@psu.edu](mailto:hanks@psu.edu)