Environmental and Ecological Statistics (2021) 28:567–586
https://doi.org/10.1007/s10651-021-00502-7

Site reduction in redundant ecosystem sampling schemes

Spencer Hays[1] · Bandana Kumari[2] · Ben Stewart-Koster[3] · Edward L. Boone[4] · Fran Sheldon[5]

Received: 24 February 2020 / Revised: 8 December 2020 / Accepted: 22 April 2021 /
Published online: 7 May 2021

Abstract

Data collection for freshwate regions of The Ecosystem Health Monitoring Program (EHMP), in southeast Queensland, Australia, involves the sampling of over 130 sites among 19 catchments twice per year and has been ongoing for over ten years. The sampling design was derived following an exhaustive process of indicator and site selection to develop a composite indicator that represented aquatic ecosystem health. After 13 years of implementation, there was an interest in identifying redundancies in sampling to reduce sampling costs without making a substantial impact on the integrity of the program and its capacity to report on ecosystem health. This paper focuses on identifying a subset of sites and times that could be removed from sampling with a minimal impact on the subsequent ecosystem health scores. Herein, Mixed models are employed to assess a variance structure from which optimality criteria are utilized to identify the scheme. Integer programs are then used to ensure specific practical constraints are observed.

Keywords A-optimal · Ecosystem-health · Freshwater · Integer programming

Handling Editor: Luiz Duczmal

✉ Spencer Hays
sphays@iu.edu

[1] Department of Statistics, Indiana University, Bloomington, IN, USA

[2] Systems Modeling and Analysis, Virginia Commonwealth University, Richmond, VA, USA

[3] Australian Rivers Institute, Griffith University, Brisbane, Australia

[4] Statistical Sciences and Operations Research, Virginia Commonwealth University, Richmond, VA, USA

[5] Sciences, Griffith University, Brisbane, Australia

1 Introduction

Comprehensive sampling is essential in monitoring the health of an ecosystem. However, the frequency, duration, and breadth of the data collected can be problematic in the practical sense of financial resources required and the physical difficulties of accessing remote sites. This paper considers the scenario of a large-scale monitoring program with a sampling scheme already in place where data are collected multiple times per year over a large number of locations within distinct regions or strata. With too few sample sites, the data are not representative of the ecosystem. That solution is straightforward: increase the number of geographic sites from which to sample. This of course, requires additional expense, both literal and figurative, so the question becomes, given an existing large-scale monitoring program, is there now redundancy in that program? And is there a way to refine and optimize the sampling scheme to reduce that redundancy? Addressed herein is exactly that: for a large-scale monitoring program, we develop a method to identify and remove redundancy in sampling while still retaining a maximal amount of information from the optimized sampling scheme. Our proposed method synthesizes concepts from Operations Research, Experimental Design, and Linear Mixed Models to reduce a large number of collection sites to a smaller representative subset.

The use of mixed models accompanied by design of experiments' (DOE) optimality measures have been used for site selection has been employed before. For example, in an agricultural setting Sebolai et al. (2005) proposes an A-optimality criterion based on mixed models' covariance matrices. Here, however, the data are simulated and represent a pre-specified fixed grid plot of an agricultural tract. Further, the A-optimality component assesses a pre-specified pattern of collections sites within the grid. While this is common in agricultural studies, it is not readily applicable in the present setting due to a physical geography that constrains the possibility of a fixed grid, and further restricts selection to a pre-specified pattern. In another agricultural study, Clarke and Stefanova (2011) relax the fixed grid constraint and also considers DOE optimality measures for sites, however the method proposed there does not employ mixed models. A comprehensive treatment of mixed models and DOE optimality measures is found in Schmelter (2007), however the application closest to ecological studies is again an agricultural setting. While the use of optimality measures of mixed models used in DOE is well-documented, these methods lack the additional step presented in our proposed method which is the use of linear integer programming to complete the site selection. Further, the application of these methods is typically restricted to agricultural studies and does not directly translate to ecological studies where the natural geography can present difficulties using a fixed-grid layout.

A novel component of our proposed method is linking a linear integer program (LP or LIP) to optimality measures on resulting covariance matrices from mixed models on data from a set of sites. The use of LIP's in site selection is also not new in the literature and specifically not new in the ecological literature. Beyer and Watts (2016) employ LIPs for site selection and do so in a setting similar to the setting that motivates our proposed method. Specifically, their method considers constraints imposed both by geography and budgetary concerns. The LIP there is the primary selection method and as such is far more complex than the one we employ here. Our use of mixed models and

optimality lessens the dependence on a complex linear program where we constrain only the total number of sites selected and that at least one site is selected within each predefined region. Moilanen (2008) also employs linear programs as a comparison to so-called "heuristic" methods. Using the terminology of this application, our approach is a hybrid of LPs and heuristic approaches by again combining the LP as a final step in site selection using mixed models and optimality measures. Methods similar to Moilanen (2008) are found in Vanderkam et al. (2007), Csuti et al. (1997), and Pressey et al. (1996), but again focus on the LP approach.

What is evident in the literature is that there are approaches using mixed models with or without a DOE optimality measure. However, these methods do not always directly apply to non-agricultural geographies. It is also evident that there is a fair amount of literature employing LPs for site selection for regions like the ecosystem motivating our development but rather compare these methods against a heuristic model, rather than combining both approaches as we do here. Therefore we differentiate our proposed approach to others that have been used for similar problems by implementing a three-step procedure of a mixed model, a DOE optimality measure, and an LIP selection algorithm.

This paper is organized in the following manner. Section 2 details the proposed method beginning with the available data then culminating in the synthesis of mixed models, A- or D-optimality, and integer programming. In Sect. 3, results are presented on data from freshwater sites sampled in the Ecosystem Health Monitoring Program (Bunn et al. 2010), illustrating model performance. In the penultimate section (Sect. 4) a simulation study is presented to illustrate the performance of the method described herein. The paper concludes with Sect. 5, which summarises key findings, addresses considerations for further implementation, and offers exciting potential for future and continued work.

2 Methods

The overarching goal is to develop a method to reduce redundancy in a sampling site scheme. Specifically, for a large number of sites, stratified across multiple regions, with data collected annually or intra-annually, we propose a method to decrease the number of sites sampled and the frequency at which they are sampled, while still retaining a maximal amount of information. Further, to ensure the sampled data is representative of the entire survey area, we can employ constraints such that each region of stratification is included. The method comprises multiple steps. First, of the variables/indices collected at each site a subset of these is selected through standard linear regression methods. Next, with the reduced set of explanatory variables, a linear mixed effect model is estimated to obtain the model variance matrix. Then, drawing from the design of experiments literature, the covariates and the variance matrix can be used to select the sites which maximize an optimality criterion within each region/catchment. Finally, an integer program is employed to enforce the aforementioned constraints. To develop the method, we use as a motivating example—which is in-fact the genesis of this method—the concept and design of the Healthy Waterways Ecosystem Health Monitoring Program (EHMP) (Stewart-Koster et al. 2014)

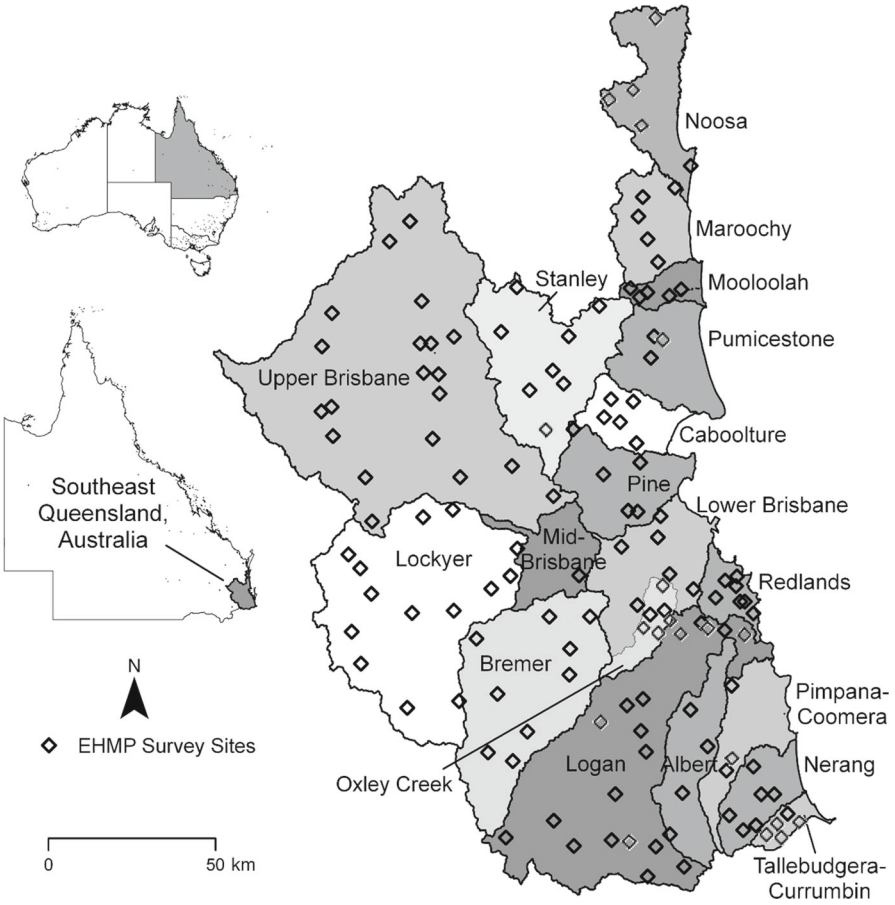

Fig. 1 Sheldon et al. (2012). Map of the region of interest relative to the continent. Currently sampled sites (131) are indicated by a diamond symbol and comprise 19 regions. The algorithm described in Section 2 is designed to select 60 sites such that at least one site exists in each of the 19 regions

described in Sect. 2.1. Subsequent subsections of this section detail the components of the proposed approach.

2.1 Data

The dataset in this application comes from the Southeast Queensland freshwater Ecosystem Health Monitoring Program (EHMP). The EHMP is a comprehensive program that assesses stream ecosystem health based on an average of 16 indicators from five indicator groups (Abal et al. 2005). The program has been running since 2002 and involves sampling 131 locations across 19 catchments (see Fig. 1), twice per year (in the austral spring and autumn) to derive the annual ecosystem health score for each catchment (Bunn et al. 2010). The five indicator groups are water quality,

macroinvertebrate assemblages, fish assemblages, nutrient concentrations and ecosystem processes. The observed data for each indicator is scored from 0 to 1 against an ideal, or reference condition, and subsequently averaged up to derive a final score for each catchment (EHMP 2008). This final score serves as the response variable for the mixed model discussed in the following sections with explanatory variables comprising the raw input variables to the various indices.

2.2 Mixed models

We apply a regression/mixed model based approach first for variable selection. Then once the final model is identified, we use that mixed model to develop A-optimality measures for site selection. Keeping in mind the EHMP scheme detailed in Sect. 2.1 as a motivating example, we use the following notation for indices

$$\begin{aligned} \text{Site} &: i = 1, \ldots, n, \\ \text{Time} &: j = t_i, \ldots, T_i, \\ \text{Variables} &: k = 1, \ldots, p. \end{aligned} \tag{1}$$

Note the allowance that each site may have a unique beginning and ending sample date, but we assume that all sites are sampled at the same frequency with evenly and identically equally spaced intervals. The method is easily extended for deviations from this assumption and these considerations are addressed further in Sect. 5. Going forward, we will reference the number of sequential time points for site i as $r_i \equiv T_i - t_i + 1$.

Our dependent variable y_{ij} can be viewed as the annual or seasonal score referenced in Sect. 2.1, which is a summary measure of the raw data collected at each site i at the jth sampling time. Each variable collected is represented by x_{ijk}: the kth predictor at site i during sampling time j, with coefficients $\beta_1, \ldots \beta_p$. Explanatory variables used for the EHMP are included in Sect. 1; these raw variables are the explanatory variables and consist of various measures of freshwater attributes. To create the score y_{ij}, measured variables are categorized and within each category are averaged to create an index scaled to the [0,1] interval. From the sub-indices an overall index of freshwater health is created.

We denote ε_{ij} and ν_{ij} as the errors associated with site i at sampling time j, which collectively form an autoregressive process of order one (AR(1)). With δ_i representing the random effect associated with site i, our model is then of the form

$$\begin{aligned} y_{ij} &= \sum_{k=1}^{p} \beta_k x_{ijk} + \delta_i + \varepsilon_{ij}, \\ \varepsilon_{ij} &= \rho \varepsilon_{i,j-1} + \nu_{ij}, \end{aligned} \tag{2}$$

with $\varepsilon_{ij} \sim AR(1)$ such that $\nu_{ij} \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$ so $COV(\varepsilon_{ij}, \varepsilon_{ij'}) = \sigma_\varepsilon^2 \rho^{|j-j'|}$. We assume for the random effect that $\delta_i \overset{iid}{\sim} N(0, \sigma_\delta^2)$ and that ν_{ij} is independent of

δ_i for all $j = 1, \ldots, T_i$; $i = 1, \ldots, n$. For ease of notation, it is assumed that $x_{ij1} = 1$ for all i, j so that $\beta_1 x_{ij1} = \beta_1$ is a model intercept. Note that model is easily extended to include random effects for region as well (fixed effects for region can be among the x_{ijk}); for now our focus is on Model (2).

Based on the usual mixed-model data and error assumptions, Model (2) is estimated via Maximum Likelihood. For the variable selection stage, initially all covariates can be considered for the model, then criteria such as The Bayesian Information Criterion (BIC) and/or Akaike Information Criterion (AIC) can be applied to the likelihood expression to determine the number of variables to retain in the final model. For both the selection model(s) and final model, performance can be assessed via cross validation. However, note that a typical ten-fold cross validation in this case can present difficulties. Recall one the primary goals of the method: that at least one site in each region is sampled. Without some sort of constraint, entire sites and even regions can be eliminated from the training set and removing an entire site/region would make a prediction for that site/region questionable. Therefore, one possible alternative is a random selection of observations for a test set and the remainder used for training the model, ensuring the training set contains at least one observation from each site. Then this procedure can be replicated. For example, we can generate, say, 200 random test sets of 100 observations each to run a modified cross validation. To measure the fit of the model the ratio of mean square error (MSE) of the fitted model to the mean square prediction error (MSPE) of the test set indicates the degree to which the error is larger in the prediction of the validation set as opposed to the training set. Values near one provide evidence that the model can successfully predict out-of-sample or, alternatively, that the model is not overfitted nor overly influenced by influential observation(s) should they exist.

2.3 Optimality

After variable selection, to begin to develop a method for site selection, we can draw on concepts from the design of experiments (DOE) literature. Specifically, one goal of design of experiments is to choose or find the "optimal" levels of treatments for experimentation given a fixed set of treatment options, which is analogous here to selecting the sites to sample based on the data collected at the site. One definition of an optimal design/site-selection would be one that minimizes the variance/covariance of the parameters estimated for $\beta_1, \ldots, \beta_p$ in Model (2). However, with p variables we have $\frac{1}{2}p(p-1)$ variance and covariance terms. Thus we require a criterion which results in the minimization of some sense of all of these terms. To this end, we can consider various optimality criteria; popular choices from DOE are *A-optimality* or *D-optimality* (see Martin 1986; Kiefer 1974, e.g., respectively), and other measures exist (such as *U*-, *G*-, *I*-, or *S-optimality*); any of which provide a scalar measure of variance/covariance that we can minimize to provide an optimal site-selection. Presently, we will use the A-optimal criterion to illustrate the method (with some notes on the D-optimal criterion) to select the sites that account for the maximal information in score measures; any of the other optimisation criteria can easily be interchanged and a discussion on the use of multiple criteria is addressed in Sect. 5. Regardless, once

the final model is determined via Sect. 2.2, to determine the variance matrix structure, consider the matrix formulation of the model introduced in that section:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\boldsymbol{\delta} + \boldsymbol{\varepsilon}, \tag{3}$$

where $\mathbf{y}_{N\times 1}$ is a vector containing the scores y_{ij} for time j at site i; $\mathbf{X}_{N\times p}$ comprises the fixed effects; $\mathbf{U}_{N\times n}$ is a matrix of indicator variables specifying the whether or not the observation belongs to site i; $\boldsymbol{\delta}_{n\times 1}$ as the random effects; and $\boldsymbol{\varepsilon}_{N\times 1}$ as the model errors for $N \equiv \sum_{i=1}^{n}(T_i - t_i + 1)$ total observations (further detail on Equation (3) components is found in Sect. 1). Based on this formulation and the error assumptions from Sect. 2.2, the variance of $\mathbf{y}$ is expressed as

$$\mathbf{V} \equiv \text{var}(\mathbf{y}) = \mathbf{U}\mathbf{G}\mathbf{U}' + \mathbf{R}, \tag{4}$$

where $\mathbf{G} \equiv E[\boldsymbol{\delta}\boldsymbol{\delta}']$ and $\mathbf{R} \equiv E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}']$. The first term matrix in Eq. (4) is a function of σ_δ^2; the latter is a block diagonal matrix with diagonal blocks defined by the covariance structure noted in Sect. 2.2: for each site i, we denote the ith block of the $\mathbf{R}$ matrix as R_i which are of the Toeplitz form with dimensions $r_i \times r_i$ with for $j, j' \in \{1, \ldots, r_i\}$, have jth row, j'th column elements $\sigma_\varepsilon^2 \rho^{|j-j'|}$. As an example, for a "balanced" design where $(T_i - t_i + 1) \equiv T\ \forall i = 1, \ldots n$ (same number of timepoints within each site), then $R_i \equiv R_{T\times T}$ for all i and with the error assumptions stated in Sect. 2.2,

$$\begin{aligned} \mathbf{R} &= \sigma_\varepsilon^2(\mathbf{I}_n \otimes R) \\ \mathbf{U}\mathbf{G}\mathbf{U}' &= \sigma_\delta^2(\mathbf{I}_n \otimes \mathbf{J}_T), \end{aligned} \tag{5}$$

with $\mathbf{J}_T$ defined as a $T \times T$ matrix with entries all equal to unity. Note that both the $\mathbf{U}\mathbf{G}\mathbf{U}'$ and $\mathbf{R}$ matrices resolve to a dimension nT block diagonal matrix with n blocks of $T \times T$ dimension with blocks $\sigma_\delta^2\mathbf{J}_T$ and $\sigma_\varepsilon^2 R$, respectively. This example generalizes to the present setting where the number of observations per site varies (see Appendix Sect. 1).

Based on Eqs. (3)– (5), it is a standard result (see McCulloch and Searle 2001, e.g.) to show that the variance/covariance matrix for the estimated parameters $\hat{\boldsymbol{\beta}}$ is expressed as

$$\mathbf{W} \equiv \text{var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}. \tag{6}$$

Measures such as the determinant or a function of the trace of the $p \times p$ matrix $\mathbf{W}$ provide a scalar measure of the "size" of the matrix and thus are criteria by which we can solve a minimization problem of the variance. In fact, A-optimality is defined by the minimization of the trace of $\mathbf{W}$, A = tr($\mathbf{W}$), while D-optimality requires the minimization of the determinant of $\mathbf{W}$, D = $|\mathbf{W}|$; In the latter case this is equivalent to minimization of the product of eigenvalues of $\mathbf{W}$; the former case requires a minimization of the sum of eigenvalues of $\mathbf{W}$. The idea behind either and all of these methods is to minimize some sense of "average" variance of the coefficients. Again, here we focus on A-optimality as an example, noting that any other optimality criterion is equally applicable.

The optimality criterion can be calculated *for each site*, then ranked according to their respective values. Specifically, recall we have defined the number of sequential time points for site i as $r_i \equiv T_i - t_i + 1$, then let $\mathbf{x}_i$ with dimensions $r_i \times p$ denote the covariate observations associated with site i. Then we can construct the $r_i \times r_i$ covariance matrix associated with that site and the corresponding optimality measures (again any of the measures noted earlier in this section). With the block diagonal covariance matrix $\mathbf{V}$ as shown in Eqs. (4) and (15), with blocks $\mathbf{V}_i$ defined as

$$\mathbf{V}_i = \sigma_\varepsilon^2 R_i + \sigma_\delta^2 \mathbf{J}_{r_i}, \tag{7}$$

a well-known result (see Greene 1994, p. 33. e.g.) is that the inverse is composed of the inverses of the individual blocks:

$$\mathbf{V}^{-1} = diag\{\mathbf{V}_i^{-1}\}_{i=1}^n. \tag{8}$$

Then the $r_i \times r_i$ covariance matrices associated with each site i and the corresponding A-optimality measure is as follows:

$$\begin{aligned} \mathbf{w}_i &\equiv (\mathbf{x}_i' \mathbf{V}_i^{-1} \mathbf{x}_i)^{-1}, \\ \mathrm{A}_i &\equiv \mathrm{tr}(\mathbf{w}_i). \end{aligned} \tag{9}$$

Again, we continue to use A-optimality as the illustrative criterion and the per-site measure applies analogously for other optimality criteria. For example, were we to use a D-optimality criterion, we could calculate $\mathrm{D}_i \equiv |\mathbf{w}_i|$. Within each catchment/region, the optimality measure (here, A_i) can be calculated in order to rank the sites according to their variability to determine the optimal site(s) in each catchment. Again noting that the goal is to select one or more optimal sites within each catchment such that a fixed total of, say, M sites are selected, we can employ the optimality measures in an integer program to achieve this goal.

2.4 Integer programming

The mixed model of Sect. 2.2 provides a method by which variables can be selected from the data collected at each site to form a refined model. The resulting covariance matrix is then used to rank sites within each region according to their contribution to overall variability as measured by the optimality criterion introduced in Sect. 2.3. The final component to our method is to quantitatively employ practical constraints to achieve the overarching goal of reducing the number of sites sampled to a fixed number M (redundancy) while ensuring that the remaining sampled sites are representative (each region/catchment is sampled). Here we rely on the method of Linear Integer Programming: the optimality criterion (OC) values of each site are known using the method described in Sect. 2.3; the objective of the integer program is to choose sites among all possible sites which will minimize the overall OC value of the model subject to these constraints.

To develop the motivation behind a linear integer program, we will focus on the A-optimality criterion; the application to other measures is immediate. We wish to find an optimal subset of sites from which to collect data, and we have defined optimal here to mean sites with the lowest OC (A-optimal) values. Going back to multi-variable calculus, a constrained optimization entails maximizing or minimizing an objective function by calculating then solving systems of derivatives to find the "z_i" values that optimize the objective function. In our present setting, the "z_i" variables simply indicate whether or not a site is selected, and we weight each of these according to their A_i value. So our objective function—the one that we would like to be as small as possible—is $\mathrm{A}_1 z_1 + \cdots + \mathrm{A}_i z_i + \cdots + \mathrm{A}_n z_n$. As constraints, we just require that $z_1 + \cdots + z_i + \cdots + z_n = \mathrm{M}$ and that within each region at least one site is represented; for example, for a region with sites s, q, and r, we require $z_s + z_q + z_r > 0$. However, note that our objective function is *linear* in the arguments z_i over which we would like to minimize. Further, the variables are *integer*-valued (binary, actually). Therefore, the usual tools of calculus are unavailable to us so we must numerically minimize our linear objective function. This is how we arrive at a linear integer program to determine the sites to be selected. Based on all possible combinations of sites to be selected, our constraints confine a boundary subset of sites that satisfy these. Then within this space, the optimal sites are chosen.

Formalizing these ideas, we can express the linear integer program in the following manner. We will continue to use the index $i = 1, \ldots, n$ to denote site i and introduce the index $h = 1, \ldots, L$ to denote a sequence of *sets* representing the sites in the total of L catchments or regions. That is, we create L partitions of the $i = 1, \ldots, n$ sites so that each site i is an element of some catchment set h: $i \in h$. We define a sequence of $i = 1, \ldots, n$ variables z_i as

$$z_i = \begin{cases} 1 & \text{if site } i \text{ is selected for sample.} \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

and the indicator function $1_{\{i \in h\}}$ as

$$1_{\{i \in h\}} = \begin{cases} 1 & \text{if site } i \text{ is in region } h, \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

Using A_i as defined in Equation (9), we then have the following linear integer program:

$$\min_{\mathbf{z}} \sum_{i=1}^{n} \mathrm{A}_i z_i \quad \text{subject to:} \quad \begin{cases} \sum_{i=1}^{n} z_i = \mathrm{M}, \\ \sum_{i=1}^{n} 1_{\{i \in h\}} z_i > 0 \text{ for } h = 1, \ldots, L, \end{cases} \tag{12}$$

which is a minimization over the sites selected $\mathbf{z} \equiv [z_1, \ldots, z_n]'$.

The extension to a seasonal analysis is straightforward. Continuing with the EHMP in mind as a motivating example, we consider semi-annual data collection in the Spring and Autumn, though consideration of other frequencies (quarterly, monthly, etc.) should follow easily from the following specification. We use the additional subscript ; q with $q \in \{v, a\}$ to indicate Spring (vernal) and Autumn, respectively, for

the defined terms in Eqs. (9), (10), and (11). For example, $A_{i;v}$ and $A_{i;a}$ indicates the respective Spring and Autumn A-optimality measure for site i. That is, in much the same way that a measure A_i can be calculated on all observations/timepoints for site i as in Sect. 2.3 and Eq. 9, we can calculate the measure $A_{i;v}$ based on all of the *Spring* time points and data associated with site i. Similarly, we define

$$z_{i;v} = \begin{cases} 1 & \text{if site } i \text{ is selected for sample in Spring,} \\ 0 & \text{otherwise.} \end{cases}$$

Using this notation, and again focusing on A-optimality, the optimization is of the following form:

$$\min_{\mathbf{z}_v, \mathbf{z}_a} \left[\sum_{i=1}^{n} A_{i;v} z_{i;v} + \sum_{i=1}^{n} A_{i;a} z_{i;a} \right] \quad \text{subject to:} \quad \begin{cases} \sum_{i=1}^{n} z_{i;v} = M, \\ \sum_{i=1}^{n} z_{i;a} = M, \\ \sum_{i=1}^{n} 1_{\{i \in h\}} z_{i;v} > 0 \text{ for } h = 1, \ldots, L \\ \sum_{i=1}^{n} 1_{\{i \in h\}} z_{i;a} > 0 \text{ for } h = 1, \ldots, L. \end{cases} \tag{13}$$

The linear integer program is the capstone to the method introduced in this paper. Mixed models are used to select variables and determine a parsimonious model to predict the dependent variable. Based on the variance/covariances estimates from that model, site-specific OC values are determined. Finally, these values are inputs to a linear program which determines a subset of sites that account for maximal information in an effort to reduce the number of sites sample and decrease redundancy. In the next section, the method is applied to the EHMP data described in Sect. 2.1.

3 Results

In this section, we apply the method of Sect. 2 to the EHMP data and design discussed in its Sect. 2.1. Of the 131 sites dispersed among 19 regions, the specific goal is to select a subset of $M = 60$ sites while at least one site in each region is selected. The analysis is performed using a combination of software packages: SAS 9.3 for mixed models and integer programming; R 3.2.2 for calculating A-optimality measures as well as providing graphics.

Using the notation of Sect. 2.2, our dependent variable y_{ij} is the *logit transformation* of the raw semi-annual score described in Sect. 2.1, which we denote here as s_{ij} so that

$$y_{ij} = \ln \left[\frac{s_{ij}}{1 - s_{ij}} \right],$$

with values of $s_{ij} = 1$ set to the value $s_{ij} = 0.999$; the distributions of the raw and transformed data are shown in Fig. 2.

The logit transformation is employed to satisfy the normality constraints of the mixed model presented in Section 2.2. The regressors x_{ijk} consist of a constant term, the raw data collected at each site (see Sects. 2.1 and 1) and a seasonal indicator

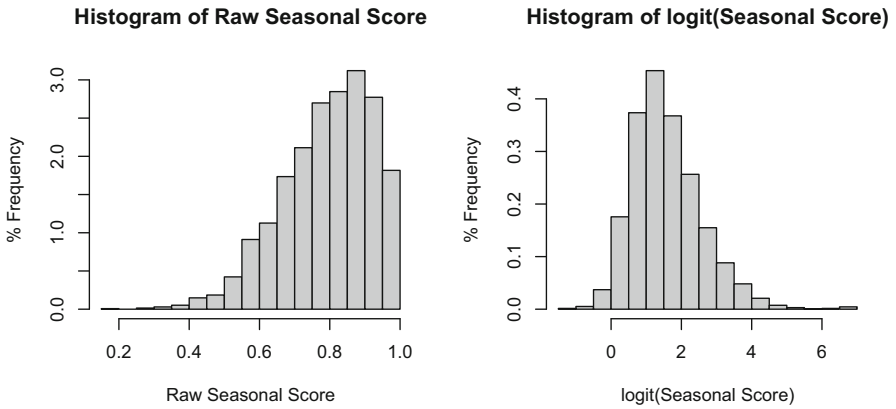

Fig. 2 Raw and logit-transformed seasonal score

(Autumn = 1). The error specifications are those noted in 2.2 and the model applied then is that of Eq. (2). The original data consists 2890 observations on 131 sites among 19 catchments. Sixteen sites were excluded from the analysis due to one or more of the explanatory variables being missing for the entire time series of that site resulting in a reduction of 124 observations. Additionally, 68 observations were removed due to a missing dependent variable; per site the data were either completely missing, or would require extrapolation for imputation. Per site, missing values for independent and dependent variables were imputed via linear interpolation using neighboring values. The resulting analysis data set thus consists of 2698 observations for 19 catchments and 115 sites, which are depicted in Fig. 1; the total number of observations per site varies from 1 to 22 consecutive semi-annual time points. From the original data, only 192 observations were removed resulting in a loss of less than 7%.

After the mixed model is estimated, we employ the integer program described in Sect. 2.4 and Eq. (12) with $n = 115$, $L = 19$, and M = 60. The results of the method are shown Fig. 3 with black dots indicating the selected sites and white diamonds denoting those that were not selected. Each region has at least one site selected, and a total of sixty sites were selected; hence the result satisfies all the requirements.

At first glance, it appears that the method results in the selection of sites that are closely clustered together geographically. However, recall that the method controls for optimal selection within each catchment. Therefore while it may appear that the selected sites cluster around specific geographical regions, these sites actually belong to distinct catchments and represent the optimal site(s) for that catchment; the algorithm is indifferent to proximity *between* catchments and focuses on optimality *within* catchments. Further, using distance as the sole delineator among sites disregards other aspects such as elevation, climate, and/or proximity to urban areas which clearly would distinguish quite distinct ecosystems.

Figure 4 illustrates this feature; depicted therein are the EHMP sites—selected and otherwise—with additional geographical and geological illustration. It can be seen that in the Brisbane area (upper inset) that four selected sites in relatively close proximity exist in actually four distinct catchments with rather different ecosystems.

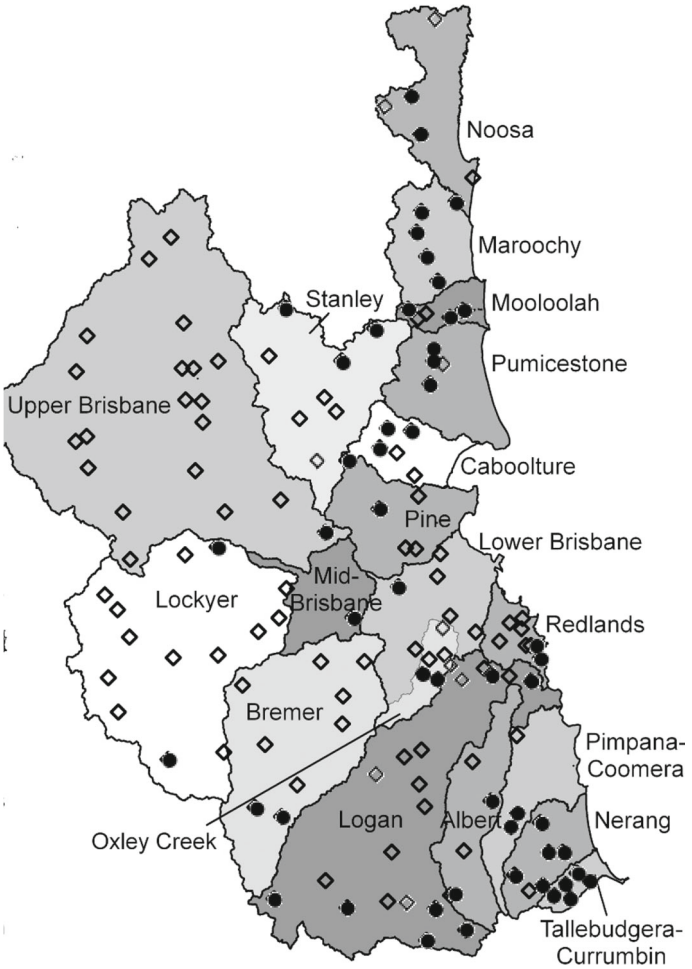

Fig. 3 Current EHMP survey sites and candidates for optimality criterion selection. Selected sites: black dots indicate sites selected via an A-optimality criterion application of the method proposed in Sect. 2. White diamonds denote unselected sites

Another example (lower inset) displays five selected sites in close proximity. One of these belongs to the separate Albert catchment; the remaining four are all in the Logan catchment. Among those four it is clear from the figure that the terrain varies significantly within a small geographic area and so it makes sense to sample from each. It is also worth noting that a nearby site (upper-left of lower inset) *very* near a selected site is not chosen so that redundancy is removed. Currently our method does not explicitly incorporate spatial/terrain components and a further iteration method could incorporate some spatial constraints should proximity be a concern; this is noted in Sect. 5.

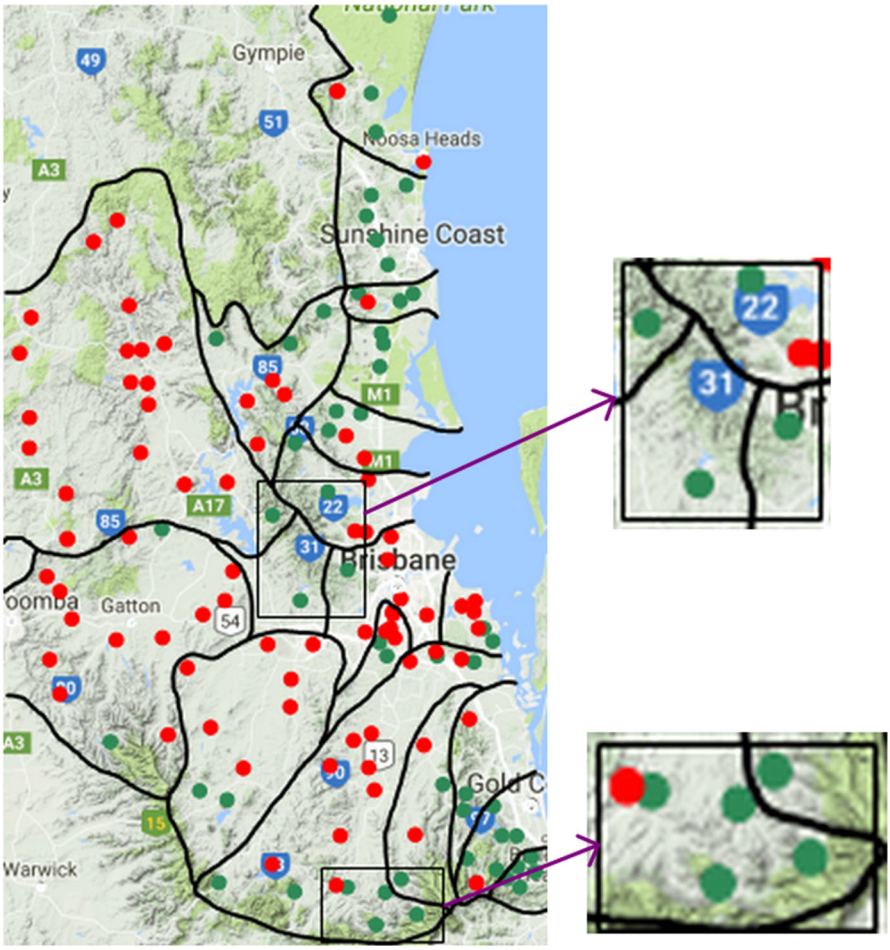

Fig. 4 Sites within and among catchments: Inspection of Fig. 3 may appear to indicate that some catchments dominate the sites selected and/or that selected sites are clustered. Selected sites appear as green dots and unselected sites are red dots. This map includes geographic/geological features to illustrate the intuitive reason for the closeness/farness of selected sites. Upper right: four nearby selected sites are located in distinct catchments *and* separated by significant geological features. Lower right: five selected sites. Only four are in the Logan catchment; despite their proximity, geological features illustrate how they are distinct

4 Simulation

To further assess the proposed method, we design a simulation study to compare this method with an alternative method. The simulation design is discussed first and then this section concludes with the simulation results. For consistency, we employ the notation introduced in Sects. 2.2 and 2.4. It is noted that to generate various quantities, we use the uniform distribution so that simulated values have greater dispersion than a distribution like the normal.

4.1 Simulation design

The following simulation was performed 500 times in two settings and can certainly be implemented for more. Because of a relatively large number of parameters to consider—as we will see below—some parameters were selected to remain fixed for the simulation study while other parameters were specified to vary from one simulation to the next.

The first component of each simulation is to define the regions and the site locations. To define a geography for the simulation we defined the continuous subspace $[0, 1] \times [0, 1] \subset \mathbf{R}^2$ as the area containing $m = 25$ regions which were then defined as

$$[0, 0.2) \times [0, 0.2), [0.2, 0.4) \times [0.2, 0.4), \ldots, [0.8, 1.0) \times [0.8, 1.0]. \quad (14)$$

$m = 25$ was chosen to reflect the types of studies similar to the EHMP setting that motivated our method.

Next, for each simulation, locations for $n = 400$ sites were randomly assigned by generating 20 random numbers from a uniform[0,1] distribution as a longitudinal coordinate then also 20 random numbers from the same distribution to generate a latitudinal coordinate. A large number of sites was selected to motivate the purpose of the proposed method: to select a much smaller sub-sample of sites that adequately explain the variation observed in all sites. The number of sites to select is specified as $M = 50$ to reflect a manageable subset of the $n = 400$ to use for continued data collection.

Finally for the time points of data collection $j = t_i, \ldots, T_i$ we select $j = 1, \ldots, T = 20$.The reason for such a selection is that 20 monthly observations is a reasonable choice for sampling designs like the EHMP described in Sect. 1 which motivated the proposed method.

For the simulated data we selected $p = 5$ explanatory variables $x_1, \ldots, x_5$. Initial values of these variables were generated from 5 non-overlapping uniform distributions ((0, 2), (2, 4) etc.). Then to account for a temporal component over $T = 20$ time points the subsequent values for $x_1, \ldots x_5$ were specified as an $AR(1)$ process with autoregressive parameters generated from a uniform distribution with boundaries 0.20 and 0.50 to reflect a moderate amount of correlation. The error component of each $AR(1)$ process is specified as normal $N(0, \sigma = 0.25)$ to reflect a moderate level of variability. The score variable y for each site i at each time j was then created as an index of the explanatory variables and the sum of two error components. Refer to the model specified in Eq. (2). One error component δ_i is generated from a normal distribution $N(0, \sigma_\delta)$ with a variance parameter unique to each region. This introduces a spatial component in that values of y_{ij} within the same region follow an identical variance component. Finally, the second error component ε_{ij} is specified as an $AR(1)$ process with error generated from an $N(0, \sigma_\varepsilon = 0.25)$ and AR parameter $\rho = 0.15$ which was close to the parameter observed in the application from Sect. 3.

We also considered another simulation setting with an additional seasonal cycle with period 12. To the error process for y four seasonal indicators were added to reflect three-month components of four seasons. The additional explanatory variable was specified as $cos(\frac{2\pi j}{12})$ as a continuous cycle with period 12. For ease of distinction

in the results section we refer to this setting as "with cycle" and the former setting as "without cycle."

4.2 Simulation assessment

For each simulation, the method described in Sect. 2 was applied to the simulated data selecting $M = 50$ sites based on that algorithm and using A-optimality as the criterion (other criteria may be used; see Sect. 5). For a model of comparison, two sites within each of the $m = 25$ regions were randomly selected. This was implemented because in the absence of any selection method but with an imperative that the sites must be reduced to M, a stratified random sample from regions is a reasonable choice. Going forward we will refer to this as the "random" method; we refer to our proposed method as the "LP" method.

To compare the two algorithms, three methods of assessment were considered. The first compares the root mean squared error (RMSE) from the mixed model (Sect. 2.2) for the selected sites from both algorithms for each simulation. We examine the average of that RMSE over all simulations as well as confidence bounds for each algorithm. Because single model is used to examine all 400 simulations over 20 times points, we expect a relatively large RMSE. It is also noteworthy that this is only a comparison on selected sites.

Notably a thorough assessment will be more reliant on examining non-selected sites; the motivation for the proposed method is to select sites that still adequately retain the information from sites that will be discontinued. Therefore, our second method of assessment is to compare each algorithm on the sites not selected. To do this, a separate mixed model (described above) is fit on the selected sites for each of the LP method and the random method. Then the non-selected sites for either algorithm are scored with their respective models. The mean over all mean squared prediction errors (MSPE) is compared as well as prediction bounds. The preferred algorithm will have a lower MSPE than the previous assessment method in that separate models are fit for each of the LP and random selection. This measure accounts for an aggregate ability to retain the information from sites that will be discontinued.

The third and final assessment provides a regional assessment of MSPE for the LP and random methods. For each simulation and each region ($500 \times 25 = 12{,}500$) MSPE is calculated. The number of times the LP MSPE is less than the random MSPE divided by the total number of per site per simulation (1250) creates a proportion estimate $\hat{p}_{LP}$. An analogous proportion $\hat{p}_r$ is also calculated. Because the regions are from a fictitious geography and the site locations are randomly assigned we need not examine each quintet any more closely than the method described here. As an assessment we consider the proportions themselves, a test of the hypothesis $H_0 : p \leq 0.50$ and an associated one-sided 95% confidence region.

4.3 Simulation results

Results for each of the three assessments discussed above are contained in Tables 1, 2, and 3. For the initial assessment on selected sites for both the noncyclic and cyclic

Table 1 Average RMSE on selected sites for two simulation types (Section 4) with +/- 2× standard deviation.)

Simulation	Method	Lower	Mean RMSE	Upper
Without cycle	LP	3.315	3.871	4.426
	Random	3.460	3.866	4.272
With cycle	LP	7.786	8.598	9.411
	Random	7.975	8.616	9.257

Table 2 Average MSPE on non-selected sites for two simulation types (Section 4) with +/- 2× standard deviation.)

Simulation	Method	Lower	Mean MSPE	Upper
Without cycle	LP	0.1856	0.1939	0.2022
	Random	0.1865	0.1942	0.2019
With cycle	LP	0.4189	0.4340	0.4490
	Random	0.4191	0.4341	0.4490

Table 3 Per region proportion of lower MSPE. For each site for each method (Section 4) MSPE is calculated for non selected sites. The proportion is calculated as the number of times the LP MSPE is lower divided by the number of simulations (500) × the number of regions (25)

Simulation	Method	Proportion	Lower	Upper	p-value
Without cycle	LP	0.517	0.510	1.0	< 0.0001
	Random	0.482	0.475	1.0	0.9999
With cycle	LP	0.510	0.503	1.0	0.0112
	Random	0.490	0.4823	1.0	0.9887

95% confidence intervals and hypothesis test $H_0 : p \leq 0.50$

settings we see very little difference between the LP and random methods. Notably this analysis is restricted to the 50 of 400 sites selected for each method. Further, because the data is simulated the optimality measures for all sites are much closer than those observed in the actual EHMP data. Extensions and alterations of the simulation method are addressed in Sect. 5.

The second assessment results are presented in Table 2. The mean MSPE for the LP method is lower than that for the randomly selected sites though the margin is relatively small. However this does show some promise of the LP method to select sites that represent the behavior on the non-selected sites. The LP method is relatively "cheap" in terms of implementation and computation. In addition, the conceptual framework of the method considers a selection procedure more rigorous than simple stratified sample and produces comparable if not better results.

Table 3 illustrates the third and final assessment of the LP and random methods described in the previous section. For both the cyclic and the noncyclic settings the proportion of times that the LP is outperforms the random method is significantly (statistically) greater than half. Proceeding along the reasoning in the previous paragraph, the LP model is preferable to random assignment or at worst is at least as good. Further, for an easily implementable method the cost is relatively small and presents a method

more intuitive. Extensions and further work for the overall method are discussed in the next section including specific considerations for a broader set of simulation studies.

5 Conclusion

Throughout the paper, various modifications or further work have been noted and those are addressed here. We then close with an overall conclusion regarding our proposed method.

One future consideration concerns the moving of sites. Our work with the EHMP data had the benefit of multiple years of data on many *stationary* sites. As long as those sites remain and remain in the same location, the model presented here can easily be updated as new data becomes available and optimized at any time. However, should sites move or be discontinued, this presents a continuity issue with the current method. Future work therefore will consider through real and simulated data the artificial removal or moving of sites to determine its impacts on the method.

While simulation results did show some promise as opposed to random selection, those results were admittedly marginal. As noted in the initial paragraph of Sect. 4.2 the proposed method relies on many assumptions and tuning parameters for a simulation. It is difficult to vary all of these in a simulation. Future work will include simulations varying subsets of these. Generating an appropriate error process for simulated data is straightforward. However the design of dependent and explanatory variables that resemble a setting like the EHMP can present some challenges. A further iteration of the simulation study will investigate even more realism in the simulated data and consider explicit spatial correlation.

Another model extension to consider is relaxing some of the mixed model assumptions presented in Sect. 2.2, specifically in regard to the model variance structure. There it was assumed that all sites were sampled at identical frequency and interval, which motivated the use of the AR(1) variance structure. In the case of missed observations, it is straightforward to apply either the AR(1) or Toeplitz variance structures to account for missed observations provided the assumption of identical intervals; it is even possible to incorporate heteroscedastic error within these structures. Going a step further if both identical frequency *and* interval are lost, it is possible to use the first order ex-ante dependence model to account for this. Any of these modifications are easily incorporated into the method should the particular data exhibit these structures.

Finally, for illustration purposes, our optimality criterion of choice was A-optimality and this was the method employed for our demonstration on the EHMP data and the simulation. There are several other optimality criteria that exist, such as D-, S-, U-, G-, and I-optimality. An interesting analysis would be to compare multiple optimality criteria and determine if the selected sites are consistent among the various criterion; this is an aspect we will consider for future work.

In conclusion, in this paper we have presented an algorithm for site selection that employs three methods from disparate areas of statistics and operations research. In order to reduce the number of sites within regions from which data is collected while retaining maximal information, we use a mixed model approach to derive the variability per site, an optimality criterion from the DOE literature to essentially rank

each site by the amount of information it contains, and finally a linear integer program to select the "best" site(s) within each region subject to a maximum total number of sites selected. The demonstration of the method was applied to the EHMP data to illustrate its effectiveness and it was shown that the method works and is an adaptable algorithm for the desired goal. A simulation study was also employed that showed some positive results but requires further exploration. It is important to note that the method is viable and applicable for any case where redundancy in monitoring programs is suspected and/or a concern.

Funding Not applicable.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Appendix

Annual/seasonal score components

Specific data collected at freshwater sites:

1. pH, Cond, Temp and DO are averaged to obtain the Water Quality Indicator,
2. DelC, R24 and GPP are averaged to form the Ecosystem Process indicator,
3. MacroRich, PET, and SIGNAL are averaged to form the Macroinvertebrate indicator,
4. PONSE, FishOE and PropAlien are averaged to form the Fish indicator, and
5. The index DelN by itself forms the Nutrient indicator.

Matrix model notation for mixed models

Section 2.3 introduces the matrix notation in Eq. 3 for Model (2) described in Sect. 2.2. This section illustrates the explicit components of the $\mathbf{y}$, $\mathbf{X}$, $\boldsymbol{\beta}$, $\mathbf{U}$, $\boldsymbol{\delta}$, and $\boldsymbol{\varepsilon}$, matrices and vectors in Eq. 3 in relation to the scalar notation and indices contained in both of Sects. 2.2 and 2.3. Based on Indices (1) and Model (2), let

$$\begin{aligned}\mathbf{y}_i &= [y_{i,t_i}, \ldots, y_{i,T_i}]',\\ \boldsymbol{\varepsilon}_i &= [\varepsilon_{i,t_i}, \ldots, \varepsilon_{i,T_i}]',\end{aligned}$$

denote the $r_i \times 1$ vectors for the score y and error terms ε for site i at times $t_i, \ldots, T_i$ for r_i adjacent and evenly spaced time points. Then the $N \times 1$ vectors $\mathbf{y}$ and $\boldsymbol{\varepsilon}$ displayed in Eq. (3) for sites $i = 1, \ldots, n$ are defined as

$$\begin{aligned}\mathbf{y} &= [\mathbf{y}_1', \ldots, \mathbf{y}_n']',\\ \boldsymbol{\varepsilon} &= [\boldsymbol{\varepsilon}_1', \ldots, \boldsymbol{\varepsilon}_n']'.\end{aligned}$$

The random effects and regressor coefficients, respectively, are denoted as

$$\begin{aligned}\boldsymbol{\delta}_{n\times 1} &= [\delta_i, \ldots, \delta_n]',\\ \boldsymbol{\beta}_{p\times 1} &= [\beta_1, \ldots, \beta_p]'.\end{aligned}$$

If we denote the $r_i \times p$ regressor matrix for site i as $\mathbf{x}_i$, then the full regressor matrix $\mathbf{X}$ in Eq. (3) is represented by the vertical concatenation of these matrices:

$$\mathbf{X} = \begin{bmatrix}\mathbf{x}_1\\ \vdots\\ \mathbf{x}_n\end{bmatrix}_{N\times p}.$$

Finally, the $\mathbf{U}_{N\times n}$ matrix consists of 0/1 entries based on whether or not observation y_{ij} is observed in site i. If we denote an $r_i \times n$ matrix $\mathbf{u}_i$ for site i as a matrix with the ith column equal to unity and all other entries equal to zero, then the $\mathbf{U}$ matrix is represented by the vertical concatenation of these matrices:

$$\mathbf{U} = \begin{bmatrix}\mathbf{u}_1\\ \vdots\\ \mathbf{u}_n\end{bmatrix}_{N\times n}.$$

General covariance matrix for y

Consider $r_i \equiv T_i - t_i + 1$, and $N = \sum_{i=1}^{n} r_i$ for sites $i = 1, \ldots, n$ with the matrix $\mathbf{J}$ and matrices R_i of dimension $r_i \times r_i$ defined as in Sect. 2.2. Then the resulting covariance matrix is expressed as

$$\begin{aligned}\mathbf{V} &= \mathbf{U}\mathbf{G}\mathbf{U}' + \mathbf{R}\\ \Leftrightarrow \begin{bmatrix}\mathbf{V}_1 & & \mathbf{0}\\ & \ddots & \\ & \mathbf{V}_i & \\ & & \ddots\\ \mathbf{0} & & \mathbf{V}_n\end{bmatrix}_{N\times N} &= \sigma_\delta^2 \begin{bmatrix}\mathbf{J}_{r_1} & & \mathbf{0}\\ & \ddots & \\ & \mathbf{J}_{r_i} & \\ & & \ddots\\ \mathbf{0} & & \mathbf{J}_{r_n}\end{bmatrix}\\ &\quad + \sigma_\varepsilon^2 \begin{bmatrix}R_1 & & \mathbf{0}\\ & \ddots & \\ & R_i & \\ & & \ddots\\ \mathbf{0} & & R_n\end{bmatrix}. \qquad (15)\end{aligned}$$

References

Beyer HL, Watts ME (2016) Solving conservation planning problems with integer linear programming. Ecol Model 328:14–22

Bunn S, Abal E, Smith M, Choy S, Fellows C, Harch B, Kennard M, f. Sheldon, (2010) Healthy Waterways, Healthy Catchments: Making the Connection in South East Queensland, Moreton Bay and Catchments Partnership, Brisbane Queensland, 222 p. Ecological Applications 55:223–240

Clarke PG, Stefanova KT (2011) Optimal design for early-generation plant-breeding trials with unreplicated or partially replicated test lines. Austral N Z J Stat 53:461–480

Csuti B, Polasky S, Williams PH (1997) A comparison of reserve selection algorithms using data on terrestrial vertebrates in Oregon. Biol Conserv 80:83–97

Greene WH (1994) Econometric analysis. Prentice Hall, Upper Saddle River

Kiefer J (1974) General equivalence theory for optimum designs (approximate theory). Ann Stat 2:849–879

Martin R (1986) On the design of experiments under spatial correlation. Biometrika 73:247–277

McCulloch CE, Searle SR (2001) Generalized, linear, and mixed models. Wiley, New York

Moilanen A (2008) Two paths to a suboptimal solution—once more about optimality in reserve selection. Biol Conserv 141:1919–1923

Pressey RL, Possingham HP, Margules CR (1996) Optimality in reserve selection algorithms: when does it matter and how much? Biol Conserv 76:259–267

Schmelter T (2007) The optimality of single-group designs for certain mixed models. Metrika 65:183–193

Sebolai B, Pedersen J, Marx D, Boykin D (2005) Effect of control plot density, control plot arrangement, and assumption of random or fixed effects on nonreplicated experiments for germplasm screening using spatial models. Crop Sci 45:1978–1984

Sheldon F, Peterson EE, Boone EL, Sippel S, Bunn SE, Harch BD (2012) Identifying the spatial scale of land use that most strongly influences overall river ecosystem health score. Ecol Appl 22:2188–2203

Stewart-Koster B, Boone EL, Sheldon F (2014) Statistical Investigation for Optimisation of the Healthy Waterways, Ecosystem Health Monitoring Program (EHMP)

Vanderkam RPD, Wiersma YF, King DJ (2007) Heuristic algorithms vs. linear programs for designing efficient conservation reserve networks: evaluation of solution optimality and processing time. Biol Conserv 137:349–358

Spencer Hays is a clinical associate professor at Indiana University in the department of statistics. He is also the director of the Indiana Statistical Consulting Center at the university. Previously her served as an assistant professor of statistics at Virginia Commonwealth University where he taught consulting, theory, and computing as well as participated in numerous research collaborations primarily in pediatric studies and ecology. Prior to his time at VCU, he served as a scientist for the U.S. Department of Energy's Pacific Northwest National Laboratory; located in southeastern Washington state. There he pursued a number of projects in applied areas ranging from power systems engineering to the detection of hazardous materials at border crossings. In 2011 Dr. Hays achieved his Ph.D. in statistics from The University of North Carolina at Chapel Hill (UNC); the thesis titled "Functional Dynamic Factor Models." In general, Hays' research focuses on the development of statistical methods for dynamic data; especially those suited to functional data analysis. His techniques draw from an education in theoretical probability and mathematical statistics, combined with practical experience in psychiatry, economics, and statistical consulting. In his time spent at UNC, Dr. Hays worked as a research assistant for the UNC School of Medicine Department of Psychiatry. There he worked with senior biostatistics faculty and researching neuroscientists investigating the effects of antipsychotic medications on schizophrenic patients and fMRI images of at-risk autistic juveniles. Prior to his time at UNC, Hays worked as a statistical consultant in industry for 6 years; for both Capital One Corporation in Richmond, VA and Toronto, then Wachovia (now Wells Fargo) in Charlotte, NC. Hays received a B.A. in economics from Indiana University in 1997, with minors in mathematics and German and followed by an M.A. in economics from Michigan State University in 1999.