Statistical Optimisation of Site Sampling to Remove Sample Redundancy.

Spencer Hays
Virginia Commonwealth University, Richmond (VA), USA - shays@vcu.edu

Bandana Kumari*
Virginia Commonwealth University, Richmond (VA), USA - kumarib@vcu.edu

Ben Stewart-Koster
Australian Rivers Institute, Griffith University, Brisbane, Australia - b.stewart-koster@griffith.edu.au

Edward L. Boone
Virginia Commonwealth University, Richmond (VA), USA - elboone@vcu.edu

Fran Sheldon
Australian Rivers Institute, Griffith University, Brisbane, Australia - f.sheldon@griffith.edu.au

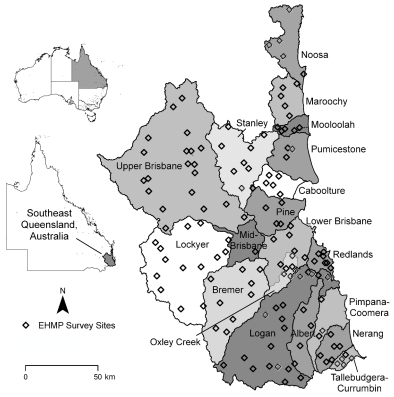
Abstract

Data collection for fresh water regions of The Ecosystem Health Monitoring Program (EHMP), in south-east 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: Ecosystem Health, Fresh water, A-optimal, Integer Programming.

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 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 optimise 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 information from the optimised sampling scheme. Our proposed method

Figure 1: Sheldon et al. (2012). Map of region of interest relative to continent.



synthesizes concepts from Operations Research, Experimental Design, and Linear Mixed Models to reduce a large number of collection sites to a smaller representative subset.

This paper is organised in the following manner. Section 2 details the proposed method beginning with the available data then culminating in the synthesis of mixed models, A- optimality, and integer programming. In Section 3, results are presented on data from freshwater sites sampled in the Ecosystem Health Monitoring Program (Bunn et al., 2010) illustrating model performance. The paper concludes with Section 4 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 sites 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 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) described in Section . 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, 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 indictor is scored from 0-1 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 Section 2.1 as a motivating example, we use the following notation for indices: Site $i = 1, \ldots, n,$; Time $j = t_i, \ldots, T_i,$; and Variables $k = 1, \ldots, p$. 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 Section . 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 Section 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 $\beta_1, \ldots \beta_p$. 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

$$\begin{aligned} y_{ij} &= \sum_{k=1}^{p} \beta_k x_{ijk} + \delta_i + \varepsilon_{ij}, \\ \varepsilon_{ij} &= \rho \varepsilon_{i-1,j} + \nu_{ij}, \end{aligned} \tag{0.1}$$

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 (0.1).

Based on the usual mixed-model data and error assumptions, Model (0.1) 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.

2.3 Optimality. We draw on concepts from the design of experiments (DOE) literature; specifically, the goal to choose or find the "optimal" levels of treatments for experimentation. This is analogous to selecting the sites to sample based on the data collected at the site. Presently, we will use the A-optimal criterion to illustrate the method to select the sites that account for the maximal information in score measures; any of the other optimisation criteria can easily be interchanged. Once the final model is determined via Section 2.2, to determine the variance matrix structure, consider the matrix formulation of the model introduced in that section:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\boldsymbol{\delta} + \boldsymbol{\varepsilon}, \tag{0.2}$$

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. Based on this formulation and the error assumptions from Section , the variance of $\mathbf{y}$ is expressed as

$$\mathbf{V} \equiv \text{var}(\mathbf{y}) = \mathbf{U}\mathbf{G}\mathbf{U}' + \mathbf{R}, \tag{0.3}$$

where $\mathbf{G} \equiv E[\boldsymbol{\delta}\boldsymbol{\delta}']$ and $\mathbf{R} \equiv E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}']$. The first term matrix in Equation (0.3) is a function of σ_δ^2; the latter is a block diagonal matrix with diagonal blocks defined by the covariance structure noted in Section : 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 Section ,

$$\begin{aligned} \mathbf{R} &= \sigma_\varepsilon^2(\mathbf{I}_n \otimes R) \\ \mathbf{UGU}' &= \sigma_\delta^2(\mathbf{I}_n \otimes \mathbf{J}_T), \end{aligned} \tag{0.4}$$

with $\mathbf{J}_T$ defined as a $T \times T$ matrix with entries all equal to unity. Note that both the $\mathbf{UGU}'$ 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.

Based on Equations (0.2)- (0.4), 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 $\mathbf{W} \equiv \text{var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$. 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. 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 Equation (0.3), with blocks defined as $\mathbf{V}_i = \sigma_\varepsilon^2 R_i$. It can be shown that the inverse is composed of the inverses of the individual blocks: $\mathbf{V}^{-1} = diag\{\mathbf{V}_i^{-1}\}_{i=1}^n$. Then the the $r_i \times r_i$ covariance matrices associated with each site i and the corresponding A-optimality measure is as follows:

$$\begin{aligned} \mathbf{w}_i &\equiv (\mathbf{x}_i'\mathbf{V}_i^{-1}\mathbf{x}_i)^{-1}, \\ \mathrm{A}_i &\equiv \text{tr}(\mathbf{w}_i). \end{aligned} \tag{0.5}$$

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 Section 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 Section 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 while ensuring that the remaining sampled sites are representative. 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 Section 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 values. We numerically minimise our linear objective function and arrive at a linear integer program to determine the sites selected; the constraints of the program determine a boundary subset of sites that satisfy said constraints from which optimal sites can be 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

$$z_i = \begin{cases} 1 & \text{if site } i \text{ is selected for sample.} \\ 0 & \text{otherwise.} \end{cases} \tag{0.6}$$

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

$$1_{\{i \in h\}} = \begin{cases} 1 & \text{if site } i \text{ is in region } h, \\ 0 & \text{otherwise.} \end{cases} \tag{0.7}$$

Using A_i as defined in Equation (0.5), we then have the following linear integer program:

$$\min_{\mathbf{z}} \sum_{i=1}^{n} \mathrm{A}_i z_i \quad \text{subject to:} \quad \begin{cases} \sum_{i=1}^{n} z_i = \mathrm{M}, \\ \sum_{i=1}^{n} 1_{\{i \in h\}} z_i > 0 \text{ for } h = 1, \ldots, L, \end{cases} \tag{0.8}$$

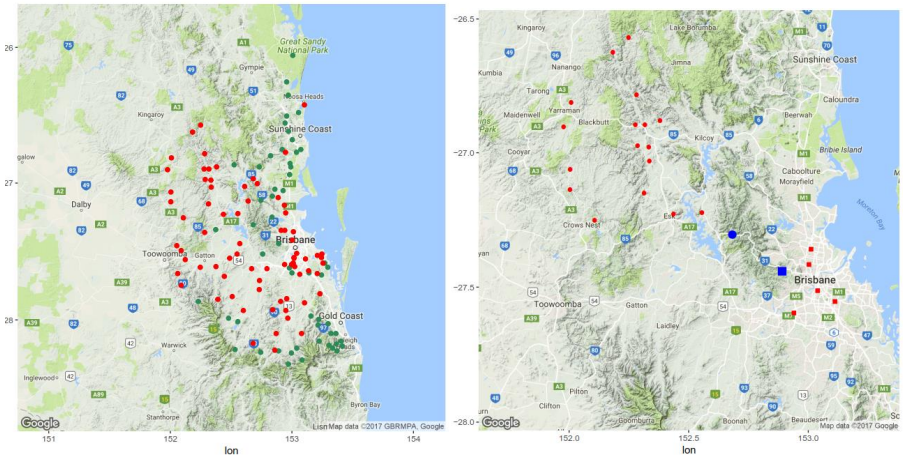
which is a minimization over the sites selected $\mathbf{z} \equiv [z_1, \ldots, z_n]'$.

The extension to a seasonal analysis is straightforward and reserved for a later paper. Using the notation of Section 2.2, our dependent variable y_{ij} is the *logit* transformation of the raw semi-annual score described in Section , 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 logit transformation is employed to satisfy the normality constraints of the mixed model presented in Section . The regressors x_{ijk} consist of a constant term, the raw data collected at each site, and a seasonal indicator (Autumn = 1). The error specifications are those noted in and the model applied then is that of Equation (0.1). The original data consists 2,890 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 2,698 observations for 19 catchments and 115 sites, which are depicted in Figure 2; 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 Section and Equation (0.8) with $n = 115$, $L = 19$, and $\mathrm{M} = 60$. Then the results of the method are shown Figure 2 with green dots indicating the selected sites and red dotes 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 3 illustrates this feature; depicted therein are the EHMP sites for the Lower Brisbane (squares) and Upper Brisbane (dots) catchments with selected sites in blue and unselected sites in red. While the selected sites are relatively close in proximity, each belongs to a distinct catchment with rather different ecosystems. A further iteration of the present method could incorporate some spatial constraints should proximity be a concern, and this is noted in Section .

Figure 2: Selected sites.



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

One future consideration concerns the moving of sites. Another model extension to consider is relaxing some of the mixed model assumptions presented in Section 2.2. 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. There are several other optimality criterion that exist, such as D-, S-, U-, G-, and I-optimality. 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. .

References

Bunn, S., Abal, E., Smith, M., Choy, S., Fellows, C., Harch, B., Kennard, M., and 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.

McCulloch, C. E. and Searle, S. R. (2001), *Generalized, Linear, and Mixed Models*, Wiley: New York.

Sheldon, F., Peterson, E. E., Boone, E. L., Sippel, S., Bunn, S. E., and Harch, B. D. (2012), "Identifying the spatial scale of land use that most strongly influences overall river ecosystem health score," *Ecological Applications*, 22, 2188–2203.

Stewart-Koster, B., Boone, E. L., and Sheldon, F. (2014), "Statistical Investigation for Optimisation of the Healthy Waterways," *Ecosystem Health Monitoring Program (EHMP)*.