Environ Ecol Stat (2014) 21:751–773
DOI 10.1007/s10651-014-0279-2

Sampling designs on stream networks using the pseudo-Bayesian approach

Matthew G. Falk · James M. McGree · Anthony N. Pettitt

Received: 3 June 2013 / Revised: 27 January 2014 / Published online: 28 February 2014

Abstract Monitoring stream networks through time provides important ecological information. The sampling design problem is to choose locations where measurements are taken so as to maximise information gathered about physicochemical and biological variables on the stream network. This paper uses a pseudo-Bayesian approach, averaging a utility function over a prior distribution, in finding a design which maximizes the average utility. We use models for correlations of observations on the stream network that are based on stream network distances and described by moving average error models. Utility functions used reflect the needs of the experimenter, such as prediction of location values or estimation of parameters. We propose an algorithmic approach to design with the mean utility of a design estimated using Monte Carlo techniques and an exchange algorithm to search for optimal sampling designs. In particular we focus on the problem of finding an optimal design from a set of fixed designs and finding an optimal subset of a given set of sampling locations. As there are many different variables to measure, such as chemical, physical and biological measurements at each location, designs are derived from models based on different types of response variables: continuous, counts and proportions. We apply the methodology to a synthetic example and the Lake Eacham stream network on the Atherton Tablelands in Queensland, Australia. We show that the optimal designs depend very much on the choice of utility function, varying from space filling to clustered designs and mixtures of these, but given the utility function, designs are relatively robust to the type of response variable.

Handling Editor: Pierre Dutilleul.

M. G. Falk (✉) · J. M. McGree · A. N. Pettitt
Mathematical Sciences, Science and Engineering Faculty, Queensland University of Technology (QUT), Brisbane, QLD 4000, Australia
e-mail: m.falk@qut.edu.au

Keywords Exchange algorithm · Pseudo-Bayesian design · Stream network · Utility function

1 Introduction

Choosing sampling locations on stream networks is of critical importance to the efficient estimation and prediction of important physicochemical and biological variables on the network. The design problem is to choose sampling locations where measurements are taken so as to maximise information gathered about variables observed on the stream network to facilitate accurate predictions for the whole network and precise estimates of model parameters. The moving average approach for spatial statistical models of stream networks (Peterson and Ver Hoef 2010; Ver Hoef and Peterson 2010; Ver Hoef et al. 2006; Cressie et al. 2006) uses a spatial covariance based on the shortest distance between two locations measured along the network (stream distance) instead of the traditional Euclidean distance. The stream network models account for the branching structure of the network, flow direction and network weighting (e.g. volume of flowing water) to estimate covariances and predict variables of interest on the network. We use these underlying models to compare and determine sampling designs for stream networks.

A substantial review of stream network sampling approaches for monitoring, categorised into probability-based and model-based designs, is presented in Dobbie et al. (2008). A summary of design for random fields is given in Müller (1998), and Caselton and Zidek (1984) consider optimal design of a monitoring network as a decision problem. We are specifically concerned here with the pseudo-Bayesian assessment of model-based designs, comparing specific fixed designs and finding optimal designs by approximating the average utility for a design using Monte Carlo integration. We also consider the so-called retrospective design of finding an optimal subset of a given set of sampling locations for which we may have data on a stream network. This type of retrospective design is considered by Diggle and Lophaven (2006) and Diggle and Ribeiro (2007) for geostatistical problems. For spatial modelling, optimal designs are usually determined for accurate predictions or estimating covariance and other parameters, noting that the two approaches can lead to vastly different designs (Zimmerman 2006), the former typically yielding evenly spaced sampling locations (space filling designs) and the latter involving some clustering. The work by Diggle and Lophaven (2006) and Zhu and Stein (2006) proposes that spatial prediction is the ultimate aim with accurate parameter estimation mainly a means to achieve this in a geostatistical context. Both Li (2009) (for stream networks) and Zimmerman (2006) (in a geostatistical context) consider three main classes of utility functions for optimal design: prediction (with known covariance parameters), parameter estimation and empirical prediction (covariance parameters unknown). These authors show that these utility function classes tend to give space filling, clustered and a mix of both, respectively, as optimal designs. This paper aims to extend these approaches to stream networks by incorporating prior information on spatial covariance model parameters for simulation based optimal design, the so-called pseudo-Bayesian approach to design. In pseudo-Bayesian designs the utility function is generally a function of the parameter only. It

is important to find pseudo-Bayesian sampling designs, as opposed to designs based on point estimates of a single covariance function, to mitigate against deviations from the assumed model and parameter values. We consider a variety of types of response variable including continuous, proportions and count data with data being modelled under the general linear and generalized linear mixed model (GLMM) framework. This range of responses allows designs to be compared across different response variable types providing wider scope for the applicability of our methodology and results to real-world problems.

Selecting optimal designs via simulation is reviewed by Muller (1999), discussing strategies including prior simulation, smoothing of Monte Carlo simulations, Markov Chain Monte Carlo (MCMC) and simulated annealing to maximise expected utility with respect to some design parameter. Recently, a greedy exchange algorithm (for example, Evangelou and Zhu 2012), has been an approach used for optimal design.

This paper proposes a version of this exchange algorithm to search for optimal retrospective designs on stream networks using Monte Carlo approximations of average utility values. This involves swapping locations in and out of a subset of locations (from a fixed set of locations) in order to find a subset (of fixed size) which optimizes the average utility. As there are different aims of stream network sampling, we consider four utility functions for finding optimal designs: (1) prediction with known covariance, (2) empirical parameter estimation, (3) a function of the Fisher information matrix for parameter estimation and (4) a hybrid of prediction and estimation involving prediction with unknown covariance parameters. To simplify computation we consider a discrete set of possible locations on the network for sampling. As there are different kinds of measurements taken, such as chemical, physical or biological measurements at each location, designs are developed based on different spatial network GLMMs (Zhang 2002).

The aim of the paper is to use a pseudo-Bayesian approach to provide algorithms to approximate average utilities in order to compare designs and to find optimal subsets for the retrospective design problem. Optimal designs across a range of utility functions are found and robustness of designs to both choice of utility function and type of response variable is investigated. We use both synthetic stream network data and data from the Lake Eacham stream network to illustrate the methodology.

Our paper is structured as follows. Section 2 outlines the statistical models used, Sect. 3 gives details of various utility functions and Sect. 4 develops the algorithms and methods. The results are presented in Sects. 5 and 6. Firstly, in Sect. 5.1 the pseudo-Bayesian approach for utility evaluation is applied to compare different fixed designs in terms of the four different utility functions. We then utilize the proposed exchange algorithm approach to find the optimal subset of locations for each utility function for a synthetic example and the Lake Eacham case study in Sect. 5.2. Section 6 gives results for designs for proportions and counts. The paper concludes with a discussion in Sect. 7.

2 Spatial statistical models for stream networks

Spatial statistical models for stream networks use moving averages and are summarised by Ver Hoef and Peterson (2010). We provide limited details here relevant to our

paper. Let $y(s_i)$ and $y(s_j)$ denote the observations taken from a spatial domain $\mathbb{S}$ at two locations s_i and s_j on a network. Assume a Gaussian random field $\mathbf{Z}$ exists over $\mathbb{S}$ and observations are conditionally independent given the value of the random field with a distribution from the exponential family. We define $\mathbf{Y}|\mu \sim$ exponential family with $E(\mathbf{Y}|\mu) = \mu = g^{-1}(\mathbf{X}\boldsymbol{\beta} + \mathbf{z}_u + \mathbf{z}_d)$, terms defined below, and build models for $\text{cov}(y(s_i), y(s_j))$ using covariance matrices for correlated random effects. Consider a stream network design with data y with S_N as the set of all possible sampling locations, S_n is a subset of these sampling locations ($S_n \in S_N$) and S_p is the set of prediction locations. In the case of evaluating fixed designs (Sect. 5.1), S_p is a general set of prediction locations on the network. However, when considering retrospective design (Sect. 5.2), we predict at locations not in the subset S_n, that is $S_p = S_N - S_n$.

General linear models for stream networks are constructed using moving averages (Ver Hoef and Peterson 2010). 'Tail-up' models are those where a moving average starts at a location on the stream network and is non-zero upstream of the location. This necessitates splitting of the moving average at stream junctions to maintain stationarity. 'Tail-down' models, alternatively, are those which are only non-zero downstream from a location. Spatial covariance is based on stream distance which is the shortest distance between two points along the stream network itself (Ver Hoef et al. 2014). We let h denote the stream distance between two points s_i and s_j on a stream network. The tail-up exponential model is as follows:

$$C_u(s_i, s_j|\theta_u) = \begin{cases} \pi_{i,j}\sigma_u^2 \exp(-3h/\alpha_u) & \text{if } s_i \text{ and } s_j \text{ are flow-connected,} \\ 0 & \text{if } s_i \text{ and } s_j \text{ are flow-unconnected,} \end{cases} \quad (1)$$

where $\sigma_u^2 > 0$ is an overall variance parameter (partial sill), α_u is the range parameter, $\theta_u = (\sigma_u^2, \alpha_u)^T$ and flow-connected sites are those connected by flowing water. To calculate the weights $\pi_{i,j}$ for tail-up models Shreve's stream order (Shreve 1967) is used which allocates a weight of 1 to each of the uppermost segments of the network (i.e. those stream segments with no segments upstream). These values are summed in the downstream direction, so that the weight for a segment downstream from a confluence (junction) is the sum of the two segments joining at the confluence. This is used to create the additive function for stream segments associated with volume, or a proxy thereof. An additive function is formed when moving downstream and the value of the additive function for a point x is $\Omega(x)$, which is equal to the additive function value of the stream segment it lies on. For two flow-connected points, s_i downstream of s_j, the tail-up model weights are:

$$\pi_{i,j} = \sqrt{\frac{\Omega(s_j)}{\Omega(s_i)}}. \quad (2)$$

The tail-down exponential model for flow-connected and flow-unconnected sites is as follows:

$$C_d(s_i, s_j|\theta_d) = \sigma_d^2 \exp(-3h/\alpha_d), \quad (3)$$

where $\sigma_d^2 > 0$, $\alpha_d > 0$ and $\theta_d = (\sigma_d^2, \alpha_d)^T$ are defined similarly to those parameters of the tail-up model.

The stream network general linear model considered is of the form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{z}_u + \mathbf{z}_d + \epsilon, \tag{4}$$

where $\mathbf{X}$ is a design matrix of fixed effects, $\boldsymbol{\beta}$ is the vector of model parameters, $\mathbf{z}_u$ contains spatially-autocorrelated random variables with tail-up autocovariance, $\text{var}(\mathbf{z}_u) = \sigma_u^2\mathbf{R}(\alpha_u)$, $\mathbf{R}(\alpha_u)$ is a correlation matrix that depends on the range parameter α_u, $\mathbf{z}_d$ is similarly defined containing spatially-autocorrelated random variables with tail-down autocovariance and ϵ contains independent random errors with $\text{var}(\epsilon) = \sigma_0^2\mathbf{I}$. The general covariance matrix is

$$\text{cov}(\mathbf{Y}) = \Sigma = \sigma_u^2\mathbf{R}(\alpha_u) + \sigma_d^2\mathbf{R}(\alpha_d) + \sigma_0^2\mathbf{I}. \tag{5}$$

Restricted maximum likelihood (REML) can be used to estimate $\boldsymbol{\beta}$ and the covariance parameter $\boldsymbol{\theta} = (\theta_u^T, \theta_d^T, \sigma_0^2)$. REML is equivalent to estimation of the variance parameters by the posterior mode, having marginalised over fixed effects with an uninformative prior (Harville 1974).

In order to find regression and covariance parameter estimates, pseudo-data are generated for parameter estimates of spatial stream network GLMMs (Wolfinger and O'Connell 1993). When considering designs for GLMMs on spatial stream networks we follow Evangelou and Zhu (2012) who, in a geostatistical context, present an approximation of the prediction uncertainty for GLMMs. We use elements of the spatial stream network models to compute utility functions for evaluating stream network designs. Details of these utility functions for general linear models and GLMMs on spatial stream networks are provided in the following section.

3 Utility functions

Bayesian optimal design is concerned with maximising the expected utility, $U(d) = \mathbb{E}[u(d, \theta, y)]$, with respect to a given design d, for an experiment yielding data y, modelled by the likelihood function $p(y|\theta)$ and prior $p(\theta)$ with $p(\theta, y) = p(y|\theta)p(\theta)$. An optimal design d^* can therefore be expressed as

$$d^* = \arg\max_{d\in D} U(d), \text{ where } U(d) = \int_y \int_\theta u(d, \theta, y)\ p(\theta, y)\ d\theta dy,$$

such that $U(d)$ is the expected utility for design d. The above integrals can be approximated using Monte Carlo integration with independent draws $(\theta^{(m)}, y^{(m)})$ from $p(\theta, y)$, that is,

$$\hat{U}(d) = \frac{1}{M}\sum_{m=1}^{M} u\left(d, \theta^{(m)}, y^{(m)}\right).$$

For pseudo-Bayesian designs we typically have utility $u(d, \theta)$ independent of y but dependent on $p(y|\theta)$, the sampling model. To derive sampling designs, speci-

fication of a utility function to quantify the usefulness of a design is required. The utility functions relevant to stream network sampling may be based on prediction, either with known covariance parameters or estimated parameters. Additionally, the utility function can be based on Fisher information which quantifies parameter information conveyed by a design, or empirical parameter estimation where parameters are unknown.

A utility for determining pseudo-Bayesian designs for prediction is given by

$$u_{\text{pred}}(d,\theta) = \left(\sum_{s_j \in S_p} \text{var}(\hat{y}(s_j)) \right)^{-1},$$

where $\hat{y}(s_j)$ is the prediction at location s_j, $\text{var}(\hat{y}(s_j))$ is the kriging variance for universal kriging (Cressie 1993) with covariances rather than variograms, $d \in D$ is the set of possible designs and θ represents the spatial covariance model parameters. When the utility is calculated with known θ it can be considered a prediction utility. Empirical prediction describes the utility when θ is unknown and calculated using estimated spatial covariance parameters, $\hat{\theta}$, which we denote by $u_{\text{Epred}}(d,\theta)$ (Zhu and Stein 2006). Empirical prediction is a hybrid approach in that the utility encompasses both accurate predictions and precise parameter estimation.

The utility function for parameter estimation can be based on Fisher information. The ijth element of the information matrix associated with restricted maximum likelihood (REML) estimation, $\mathbf{I}_{\text{REML}}(d,\theta)$, is given by

$$\frac{1}{2}\text{tr}\left(P\frac{\partial \Sigma}{\partial \theta_i} P\frac{\partial \Sigma}{\partial \theta_j} \right),$$

where $P = \Sigma^{-1} - \Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}$, Σ is the covariance matrix of the response y and θ is the vector of covariance parameters. Optimal designs seek to maximize

$$u_{\text{FIM}}(d,\theta) = det\left(\mathbf{I}_{\text{REML}}(d,\theta)\right).$$

Alternatively, a utility function can also be defined for empirical parameter estimation, such as

$$u_{\text{Eparm}}(d,\theta) = det\left(\hat{\text{var}}(\hat{\theta}|d)^{-1}\right),$$

where $\hat{\theta}$ represents the estimator of the parameters based on design d and $\hat{\text{var}}$ the estimated variance. For the retrospective design using the Lake Eacham case study, the true value of θ is assumed to be that found using all the sampling locations, denoted by θ_N.

The proposed objective when designing for GLMMs on spatial stream networks is to predict the Gaussian random field $\mathbf{Z}$ over the spatial domain $\mathbb{S}$ for data y at locations S_p. At a location on the network, $\mathbf{Z}$ is defined by $(\mathbf{z}_u + \mathbf{z}_d)$. We use the approximation of prediction uncertainty for GLMMs (Evangelou and Zhu 2012) and seek designs that maximise its inverse, which we denote $u_{\text{GLMM}}(d, \theta)$.

4 Methods

In this section we firstly introduce the algorithms for utility function estimation and determining retrospective designs. The algorithms and methods are then developed for their application to the examples.

4.1 Algorithms

The first algorithm is a pseudo-Bayesian approach to estimate the utility function for sampling designs on stream networks. For a given design, the algorithm firstly draws from a prior distribution on covariance parameters. If required, the data are simulated, a stream network model fitted and predictions generated. A utility is then calculated (Sect. 3) and the process is repeated. The estimated utility is found as the average of the utility from the Monte Carlo simulations. The algorithm is outlined below.

Algorithm 1 Utility function estimation for sampling designs

for $d = 1 : D$ ##sample designs## **do**
  for $m = 1 : M$ ##Monte Carlo simulations## **do**
    Draw $\theta^{(m)} \sim p(\theta)$, where $p(\theta)$ is a prior on parameters
    Simulate $y^{(m)} \sim p(y|\theta^{(m)})$ at design locations S_n (if required)
    Fit spatial stream network model to estimate parameters $\hat{\theta}^{(m)}$ (if required)
    Generate $\hat{y}^{(m)}$ at prediction sites S_p (if required)
    Calculate utility $u^{(m)} = u(d, \theta^{(m)})$ from Section 3, as required.
  end for
  Estimated utility $\hat{U}(d) = \frac{1}{M} \sum_{m=1}^{M} u^{(m)}$
end for

A greedy exchange algorithm is also proposed to determine a subset of sampling locations S_n from possible locations in S_N, which is known as a retrospective design (Diggle and Lophaven 2006). The greedy exchange algorithm begins with a random sample to form the initial subset, S_n, then seeks to swap sites from S_N so as to maximise the utility function, estimated via Algorithm 1, until there is no improvement. Exchange algorithms for design problems are reviewed by Royle (2002) and it is acknowledged that the algorithm does not necessarily find the optimal design but finds a reasonable design quickly (Evangelou and Zhu 2012).

Let d^{ij} denote the design where the ith location in S_n is swapped for the jth location in $S_N - S_n$. The search is repeated K times in the hope of avoiding local optima and is presented in Algorithm 2.

Algorithm 2 Greedy exchange algorithm

for $k = 1 : K$ ##number of reps.## **do**
  Set d as a random configuration of n sampling locations
  Evaluate $\hat{U}(d)$ as per Algorithm 1
  Initialise $\hat{U}(d^{ij}) = \hat{U}(d)$
  while $\max(\hat{U}(d^{ij})) \geq \hat{U}(d)$ **do**
    for $i = 1 : n$ **do**
      for $j = 1 : (N - n)$ **do**
        Evaluate $\hat{U}(d^{ij})$ as per Algorithm 1
      end for
    end for
    Find locations which maximise utility, $(\hat{U}(d^{ij}))$, and swap
    if $\max(\hat{U}(d^{ij})) > \hat{U}(d)$ **then**
      $\max(\hat{U}(d^{ij})) \rightarrow \hat{U}(d)$
      $d^{ij} \rightarrow d$
    end if
  end while
end for

4.2 Illustrative examples

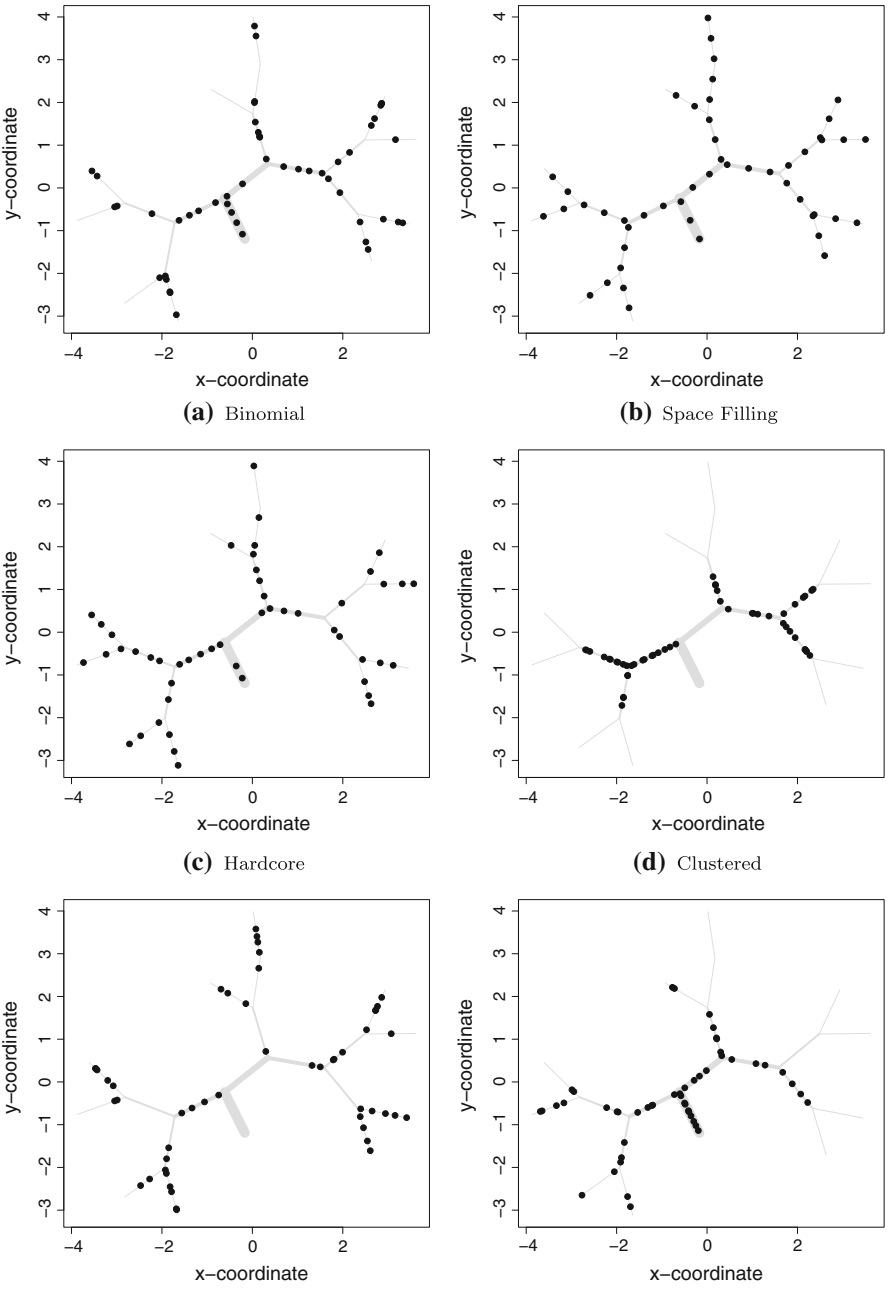
Algorithm 1 is initially used to estimate the utility functions outlined in Sect. 3 for six fixed designs of $N = 50$ sampling locations which are described in Table 1 and presented in Fig. 1. In Fig. 1, the thickness of the stream segments represents stream order (a proxy for flow), with thicker segments indicating higher stream order (higher flow). A mixture of exponential tail-up (Eq. 1) and tail-down (Eq. 3) network covariance (Eq. 5) is considered. Informative independent normal priors (truncated at zero) are placed on covariance parameters (partial sill and range parameters for exponential tail-up and exponential tail down models, $\theta = (\sigma_u^2, \alpha_u, \sigma_d^2, \alpha_d)$) such that the means are (2, 5, 2, 5) and standard deviations (0.25, 1, 0.25, 1), respectively. A continuous response is predicted at $S_p = 500$ evenly spaced locations over the network.

The greedy exchange algorithm is used for retrospective design of synthetic data (Fig. 3a) and the Lake Eacham case study (Fig. 5a, b). Figure 3a shows the possible sampling locations at 0.1, 0.5 and 0.99 units up each segment on the synthetic stream

Table 1 Fixed designs and descriptions

Design	Description
Binomial	Samples allocated to random locations over the entire network
Space filling	Sample locations are evenly spread over the network
Hardcore	As for Binomial except sample locations are removed that are too close to one another[a]
Clustered	Sample locations clustered close to the centre of the network
Upstream	Majority of samples at locations near the source segments (headwaters) of the network
Downstream	Majority of samples on segments close to the outlet

Binomial, space filling and hardcore designs from Ver Hoef et al. (2014)
[a] based on a user entered inhibition region



(a) Binomial **(b)** Space Filling

(c) Hardcore **(d)** Clustered

(e) Mainly Upstream (near headwaters) **(f)** Mainly Downstream (near outlet)

Fig. 1 Six different sampling designs on a simple network, each with 50 locations. The thickness of the stream segments represents stream order (flow), with thicker segments indicating higher stream order (flow)

network. Note the stream segment lines are thicker for higher stream orders (flows). We use the proposed pseudo-Bayesian retrospective selection of design points to find the optimal subset of $n = 20$ from $N = 60$ locations.

Data collected near Lake Eacham (Spatial Reporting Of Ecosystem Health Project 2009) is included in the SSN package version 1.0. Lake Eacham is located on the Atherton Tablelands in Queensland, Australia. The dataset contains $N = 88$ sampling locations which are shown in Fig. 5a and magnified in Fig. 5b showing the sampling locations in more detail. Thicker lines indicate a larger watershed area for that segment which is a proxy for flow. Note there are several sampling sites in very close proximity, particularly in the top right of the network. We use the proposed pseudo-Bayesian retrospective selection of design points to find the optimal subset of $n = 22$ from $N = 88$ locations.

Spatial stream network models are fitted with a mixture of exponential tail-up and tail-down network covariance (as above). Informative independent normal priors centred at the true values (but truncated at zero) are again placed on spatial covariance parameters. For the synthetic network example the true spatial covariance parameters are assumed to be $\theta = (2, 5, 2, 5)$ with corresponding prior standard deviations $(0.25, 1, 0.25, 1)$. For the Lake Eacham dataset the true parameters are assumed to be those calculated (rounded) using data from all the sampling locations, that is $\theta = (1.5, 50{,}000, 3, 20{,}000)$, with corresponding prior standard deviations $(0.25, 5{,}000, 0.5, 2{,}000)$. Distance units are in metres and results are based on $M = 1{,}000$ draws from the prior.

4.3 Adaption for counts and proportions

In determining designs where the response variable is non-continuous we consider two common link functions: the logit for binomial (presence/absence) data and the log for Poisson (count) data. The utility function, u_{GLMM}, is used to find designs for count and binomial data on both the synthetic example and Lake Eacham stream networks introduced earlier. A constant mean random field with the number of trials in a Bernoulli experiment for binomial models, or the length of time that sampling is taking place for Poisson models, denoted by R, is varied to observe the differences in sampling location subset designs ($R = 20, 50, 150$). Additionally, the mean of the random field for the Poisson model (denoted by P) is varied to determine its influence on the design of the subset of sampling locations for count data ($P = 1, 30$). All other details regarding the random field covariance specification are as outlined in Sect. 4.2.

5 Results

We describe the results in the sections below. The models are implemented in the SSN package (Ver Hoef et al. 2014) in R. We firstly evaluate utility functions for a set of fixed designs on a simple network and compare. In the later sections we combine the utilities with the exchange algorithm to find an optimal subset of sample locations from an initial set for a synthetic example and the Lake Eacham networks with a continuous response variable (for example, stream water chemistry). The synthetic

Table 2 Utility function means (standard errors) for sampling designs with a mixture of exponential tail-up and tail-down covariance models

Design	U_{pred}	U_{Epred}	U_{Eparm}	U_{FIM}
Binomial	24.44 (3.84)	21.77 (2.99)	26.82 (1.18)	**25.51** (3.27)
Space filling	**39.76** (3.87)	**42.89** (7.31)	27.92 (1.22)	22.75 (3.45)
Hardcore	33.16 (4.59)	32.28 (5.09)	24.49 (1.07)	11.82 (1.80)
Clustered	10.82 (1.29)	9.16 (1.10)	**42.87** (2.11)	14.11 (2.38)
Upstream	22.32 (3.62)	19.10 (2.47)	25.05 (1.12)	21.74 (3.11)
Downstream	13.68 (2.04)	12.16 (1.25)	32.50 (1.55)	14.80 (2.20)

U_{pred}, U_{Epred}, U_{Eparm} and U_{FIM} correspond to prediction, empirical prediction, empirical parameter estimation and Fisher information utility functions, respectively. Bold values indicate the design with the highest utility

data are presented as a stylised network to aid in visualising the clustering or spread of designs, but this does not influence modelling as this is based on stream network distances. The Lake Eacham case study is presented as a typical dendritic network being faithful to the actual geography. Finally, the optimal subset selection of sample locations is considered for networks of non continuous response variables (for example, fish counts) and these are compared to those designs found for networks with a continuous response variable.

5.1 Pseudo-Bayesian evaluation of sample designs

A number of functions are provided in Ver Hoef et al. (2014) to generate designs on a stream network. We also generate three further designs. These are summarised in Table 1. The most appropriate data collection design is identified according to utility functions in Sect. 3 using Algorithm 1.

The results are presented in Table 2 showing that when the prediction utilities are considered (U_{pred} and U_{Epred}), a space filling design is the most appropriate. However, when the empirical parameter estimation (U_{Eparm}) and Fisher information utilities (U_{FIM}) are considered, the clustered and binomial designs prevail as the best, respectively. It seems sensible that to best predict we need an even coverage of points, but to estimate spatial parameters we need some clustering as observed by Diggle and Lophaven (2006) and Zimmerman (2006) for geostatistical and spatial analyses.

The utilities for exponential tail-up and exponential tail-down covariance models individually are given in Tables 3 and 4, respectively. These utilities for the individual covariance models show the same best designs using the prediction utilities as in Table 2, that is, a space filling design. However, the empirical parameter estimation utility is largest for the binomial design and Fisher information is largest for the clustered design. When considering the empirical parameter estimation utility, this indicates that a binomial design which exhibits some clustering so that a range of stream distances are included to estimate partial sill and range parameters, is best. Note also that the clustered design is the worst performing design for empirical parameter estimation for the individual covariance models which suggests that including only mid-range stream

Table 3 Utility function means (standard errors) for sampling designs with exponential tail-up covariance model

Design	U_{pred}	U_{Epred}	U_{Eparm}	U_{FIM}
Binomial	47.26 (7.48)	43.38 (4.81)	**190.78** (6.91)	73.63 (5.70)
Space filling	**89.13** (11.55)	**90.61** (15.91)	146.89 (4.05)	42.44 (5.33)
Hardcore	66.34 (9.05)	67.68 (9.28)	160.94 (2.69)	70.43 (15.83)
Clustered	17.75 (1.93)	17.51 (2.02)	80.39 (1.79)	**102.37** (13.73)
Upstream	39.50 (6.46)	36.90 (4.75)	140.19 (2.67)	34.93 (2.20)
Downstream	27.49 (4.25)	26.28 (2.62)	156.06 (3.70)	89.94 (10.93)

U_{pred}, U_{Epred}, U_{Eparm} and U_{FIM} correspond to prediction, empirical prediction, empirical parameter estimation and Fisher information utility functions, respectively. Bold values indicate the design with the highest utility

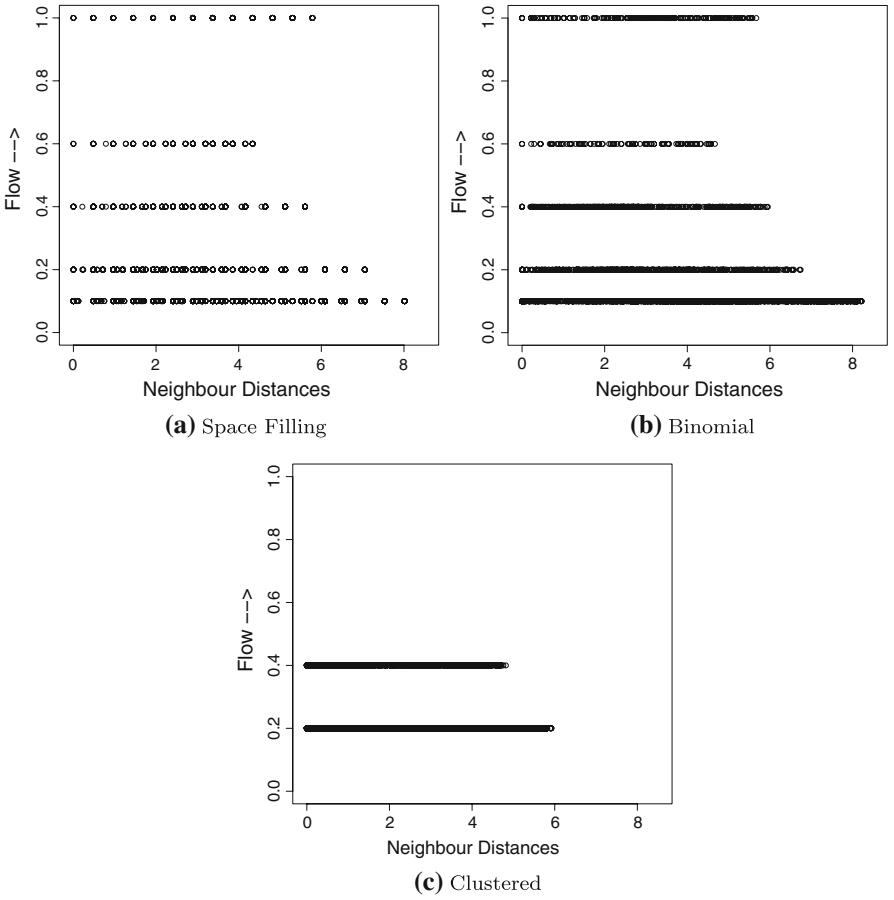
Table 4 Utility function means (standard errors) for sampling designs with exponential tail-down covariance model

Design	U_{pred}	U_{Epred}	U_{Eparm}	U_{FIM}
Binomial	56.82 (8.41)	51.53 (6.56)	**430.97** (4.97)	92.73 (7.92)
Space filling	**95.40** (12.87)	**98.94** (17.38)	216.23 (2.89)	61.48 (6.24)
Hardcore	80.03 (10.99)	80.67 (12.93)	277.16 (3.15)	108.13 (21.41)
Clustered	23.29 (2.36)	21.96 (2.72)	133.94 (3.09)	**206.03** (31.8)
Upstream	53.82 (8.01)	51.99 (7.51)	210.48 (2.52)	70.79 (8.19)
Downstream	32.80 (4.73)	32.83 (3.19)	173.93 (2.95)	185.13 (19.75)

U_{pred}, U_{Epred}, U_{Eparm} and U_{FIM} correspond to prediction, empirical prediction, empirical parameter estimation and Fisher information utility functions, respectively. Bold values indicate the design with the highest utility

sampling locations does not assist in predicting using estimated tail-down (tail-up) parameters. The Fisher information utility indicates the clustered design with sample locations mid way between outlet and source segments conveys the most information about the network.

The differences in designs can be seen further in Fig. 2, which plots stream order (flow, converted to a value over the interval [0, 1]; higher values indicate higher flow) versus neighbour distances for the locations in the design. The best design for prediction is the space filling design, shown in Fig. 2a, which shows an even spread over all distances at all stream orders (flows) in the network. The best designs for parameter estimation and Fisher information are binomial and clustered, respectively, which are shown in Fig. 2b, c. The binomial design shows locations representing all stream orders (flow) in the network, similarly to the space filling design, but more densely placed on the network. The clustered design shows locations densely in mid stream order (flow) locations on the network. Figure 2 contrasts the two main types of designs, space filling and clustered, while later applications of this plot demonstrate how the designs incorporate these two features.



(a) Space Filling **(b)** Binomial

(c) Clustered

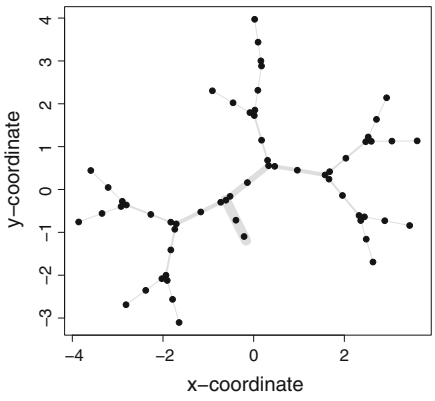
Fig. 2 Space filling, binomial and clustered designs plotted as stream order (flow, converted to a value over the interval [0, 1]) versus neighbour distances. Note there are many non-unique locations for the space filling design and hence there appears to be more locations in the binomial and clustered designs. Higher stream orders (flows) are indicated by a higher value

5.2 Retrospective design for stream network general linear model

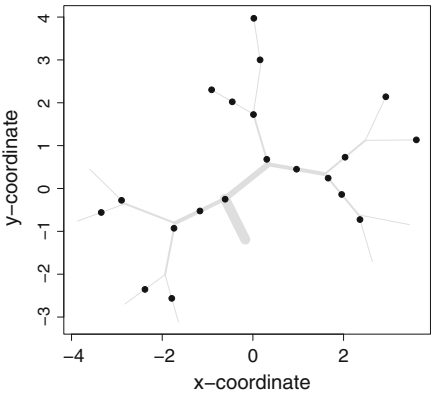
This retrospective design approach is implemented using a greedy exchange algorithm (Algorithm 2) and is applied to a synthetic example and data collected near Lake Eacham in Queensland, Australia.

5.2.1 Synthetic example

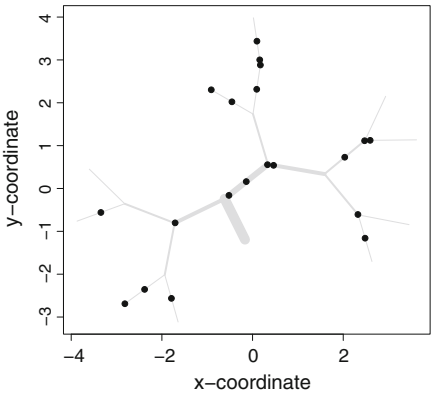
The best subset of sampling locations is shown in Fig. 3b using the prediction variance utility, U_{pred}. Note that the selected subset of locations mostly lie on the side of the network with the most segments, with locations reasonably evenly spread over all network segments.



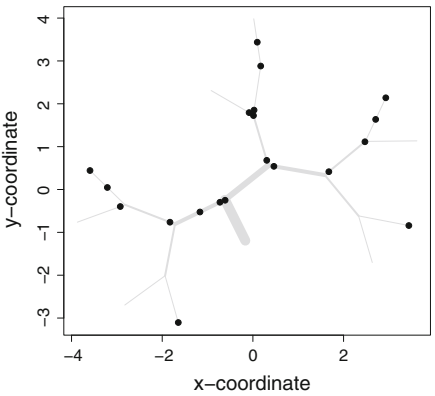
(a) Sampling Locations



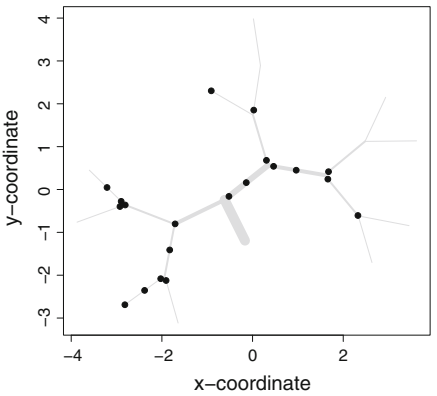
(b) U_{pred}



(c) U_{Eparm}



(d) U_{Epred}



(e) U_{FIM}

Fig. 3 Synthetic network example. **a** 60 possible sampling locations. Best 20 sampling locations for a continuous response using **b** the prediction variance utility, **c** empirical parameter estimation utility, **d** empirical prediction (hybrid) utility and **e** Fisher information utility

Using the empirical parameter estimation utility, U_{Eparm}, the best subset of sampling locations is shown in Fig. 3c. These sampling locations are generally at the more extreme source segment locations, with a few locations on segments near the outlet. There is also some clustering of sampling locations, particularly along the segment at the top of Fig. 3c, which would assist in more accurately estimating the covariance parameters.

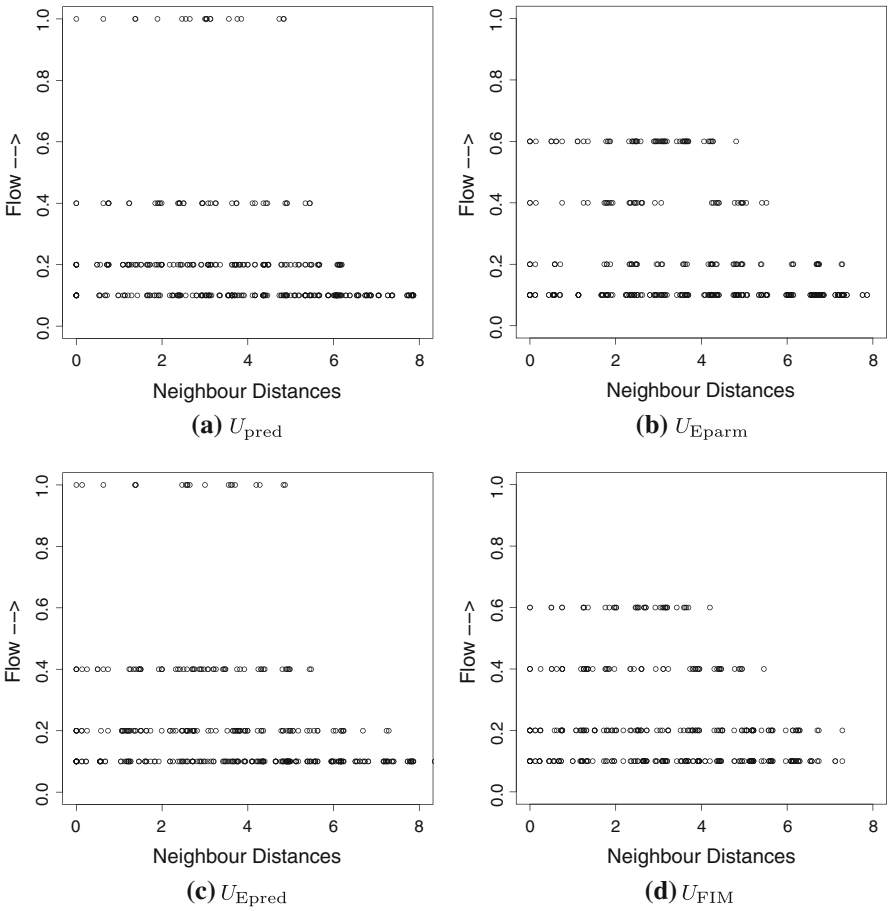
The subset of sampling locations with the highest empirical prediction utility (hybrid, U_{Epred}) is shown in Fig. 3d. This design selects sampling locations with some clustering near two of the junctions, with other locations spread somewhat evenly over the network. This suggests that the design requires a trade off between nearby sample locations for covariance parameter estimation and space filling sample locations to ensure prediction variance is kept at a reasonable level.

Finally, the utility based on Fisher information is presented in Fig. 3e. This design is similar to that found using the empirical parameter estimation utility in that it shows some clustering around junctions and a range of stream distances between locations to estimate parameters.

The designs in Fig. 3 are plotted again in Fig. 4 in terms of stream order (flow, converted to a value over the interval [0, 1]) versus neighbour distances to further examine the selected pseudo-Bayesian designs. Designs for prediction in Fig. 4a, c show sampling locations at stream orders (flows) near the source and outlet segments with perhaps closer neighbours for the empirical parameter estimation, U_{Epred}. The designs for U_{Eparm} and U_{FIM} shown in Fig. 4b, d, respectively, do not include samples at locations near the outlet as information about the stream network here is likely contained in upstream locations.

5.2.2 Lake Eacham

The pseudo-Bayesian designs found using the exchange algorithm with different utilities are presented in Fig. 5, plotting flow (converted to a value over the interval [0, 1], note there are no sampling locations on the most downstream segment of the stream network) versus neighbour distances. The best design using the prediction variance utility, U_{pred}, is shown in Fig. 5c. Again we see that this is a space filling design that covers as much of the network as is possible from the original samples in both upstream and downstream locations over the full range of neighbour distances. The design generated using the empirical parameter estimation utility shown in Fig. 5d again shows some more clustering of sample subset sites for accurate parameter estimation, particularly on source segments. This is similar to the design generated using Fisher information, shown in Fig. 5f. These designs do not have neighbour distances over the full possible range which indicates clustering. Finally, the balance between a space filling and clustering design of locations is again visible in the pseudo-Bayesian design using the empirical prediction utility, shown in Fig. 5e. This design includes locations at low, mid and high stream flow locations, similar to U_{pred}, but also exhibits clustering since some neighbour distances are not over the full range, particularly at upstream locations near the source. Note all designs include at least one location on the most downstream segments of the network near the outlet.



(a) U_{pred} **(b)** U_{Eparm}

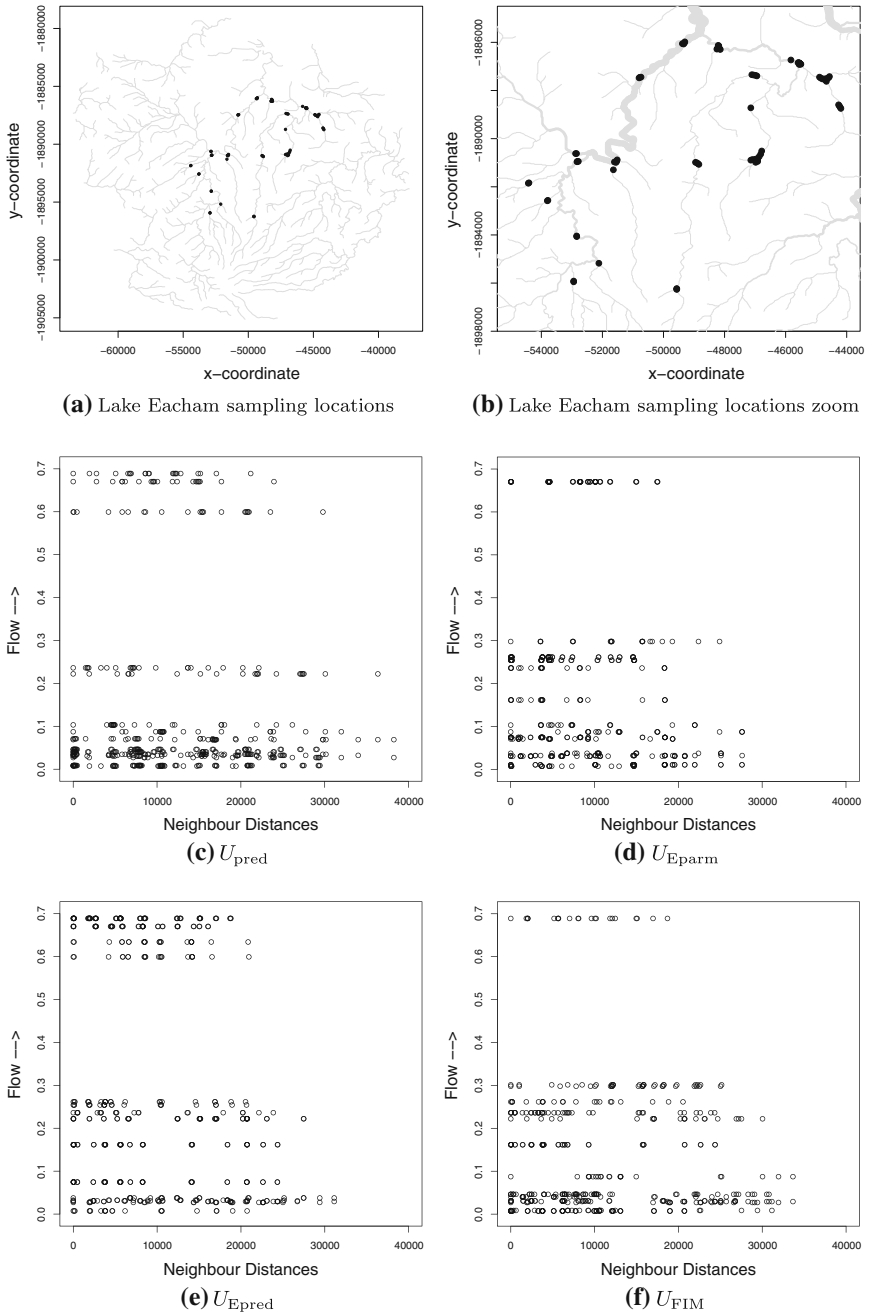
(c) U_{Epred} **(d)** U_{FIM}

Fig. 4 Synthetic network example with designs plotted as flow (larger values indicate higher flow) versus neighbour distances. Best 20 sampling locations for a continuous response using **a** the prediction variance utility, **b** empirical parameter estimation utility, **c** empirical prediction (hybrid) utility and **d** Fisher information utility

This is possibly due to the tail-up model so that including that one location would provide a lot of information since it is flow-connected with every other location in the network.

6 Retrospective design for stream network generalized linear mixed models (GLMM)

This section investigates different response types such as count and binomial data in the GLMM framework on a stream network for the synthetic example and the Lake Eacham case study.



(a) Lake Eacham sampling locations

(b) Lake Eacham sampling locations zoom

(c) U_{pred}

(d) U_{Eparm}

(e) U_{Epred}

(f) U_{FIM}

Fig. 5 **a** Eighty-eight possible sampling locations for the Lake Eacham dataset. **b** Zoomed image of the 88 Lake Eacham sampling sites with thicker lines indicating higher flow and thus closeness to outlet. **c**–**f** present the best subset of 22 sampling locations plotted as flow (larger values indicate higher flow) versus neighbour distances using different utility functions. **c** Prediction variance. **d** Empirical parameter estimation. **e** Empirical prediction (hybrid). **f** Fisher information utility

6.1 Synthetic example

Pseudo-Bayesian retrospective designs on the synthetic example network with non-continuous response variable are shown in Fig. 6. The left hand panels (Fig. 6a, c, e) are designs for the binomial model with different values of R, the number of replicates at each location, (20, 50, 150); the right hand panels (Fig. 6b, d, f) are designs for the Poisson model with the same R values. The binomial model seems to approach the continuous designs for all values presented of R, in that all designs resemble the designs using the prediction variance utility with neighbour distances over the full range of values and locations at low, mid and high stream order (flow) locations. The Poisson model, only with $R = 150$, generates designs representing all stream orders (flows) on the network. Smaller values of R yield more clustered designs by including only low to mid stream order (flow) locations. This appears to indicate the space filling design is appropriate for Poisson and binomial models with large R.

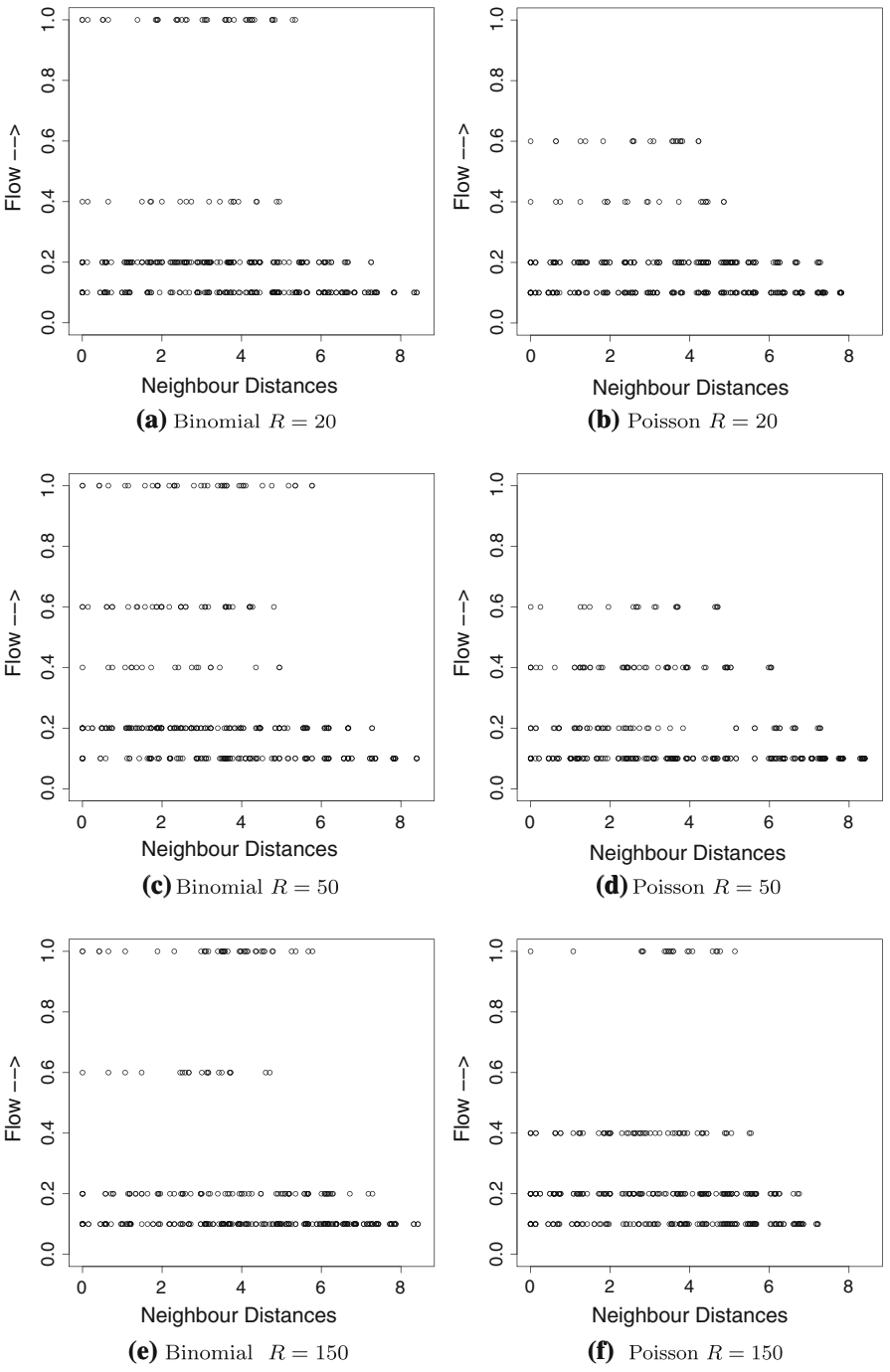
The best subset of design locations for the Poisson model on the synthetic example network with changing values of P are presented in Fig. 7a for $P = 1$ and Fig. 7c for $P = 30$. When P is small it can be seen that the design tends to be more clustered in that sample locations do not cover the full range of stream orders (flows). For $P = 30$, the design appears to be similar to the space filling designs for a continuous response variable.

6.2 Lake Eacham case study

The pseudo-Bayesian retrospective designs for Lake Eacham are also considered for presence/absence and count response variables. The random field has a constant mean and a mixture of exponential tail-up and tail-down covariance structure. Designs again appear to show a trade off between a uniform spread and clustering as R increases, shown in Fig. 8.

The binomial model, for small R, has no sampling locations at high flows near the outlet. As R increases, we see more sampling locations included at locations with high flows near the outlet, thus tending towards a space filling design as for the continuous model. For the Poisson model, as R increases, we see more points in the mid range flows, while all values of R include sampling locations at low, mid and high flows.

The best subset of design locations for the Poisson model on Lake Eacham with changing values of P parameter are presented in Fig. 7b for $P = 1$ and Fig. 7d for $P = 30$. It can be seen that for larger P, the designs are similar to those presented earlier with a continuous response variable, specifically Fig. 4a for the synthetic example and Fig. 5c for Lake Eacham. That is, for larger P these designs approach the space filling designs which are optimal for stream network models using the prediction utility with a continuous response variable. This suggests the space filling design is robust for different types of response variables measured on stream networks.

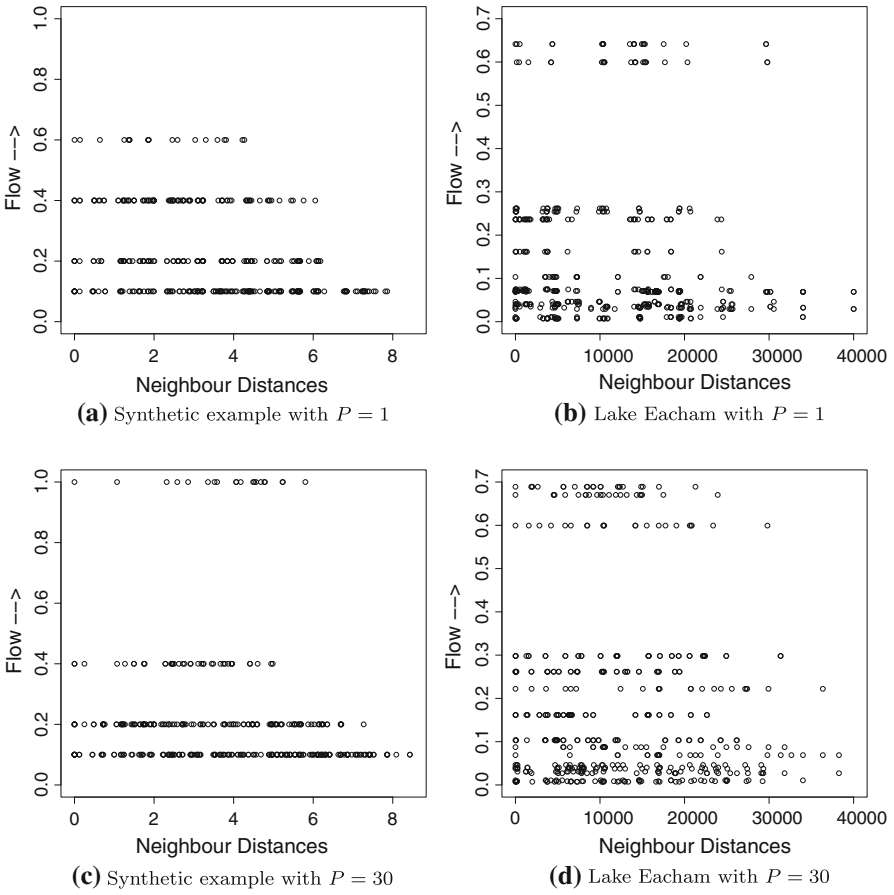


(a) Binomial $R = 20$ **(b)** Poisson $R = 20$

(c) Binomial $R = 50$ **(d)** Poisson $R = 50$

(e) Binomial $R = 150$ **(f)** Poisson $R = 150$

Fig. 6 Best subset of sampling locations plotted as flow (larger values indicate higher flow) versus neighbour distances for the synthetic network with binomial (**a**, **c**, **e**) and count (**b**, **d**, **f**) measurements with varying R values ($R = 20, 50, 150$)



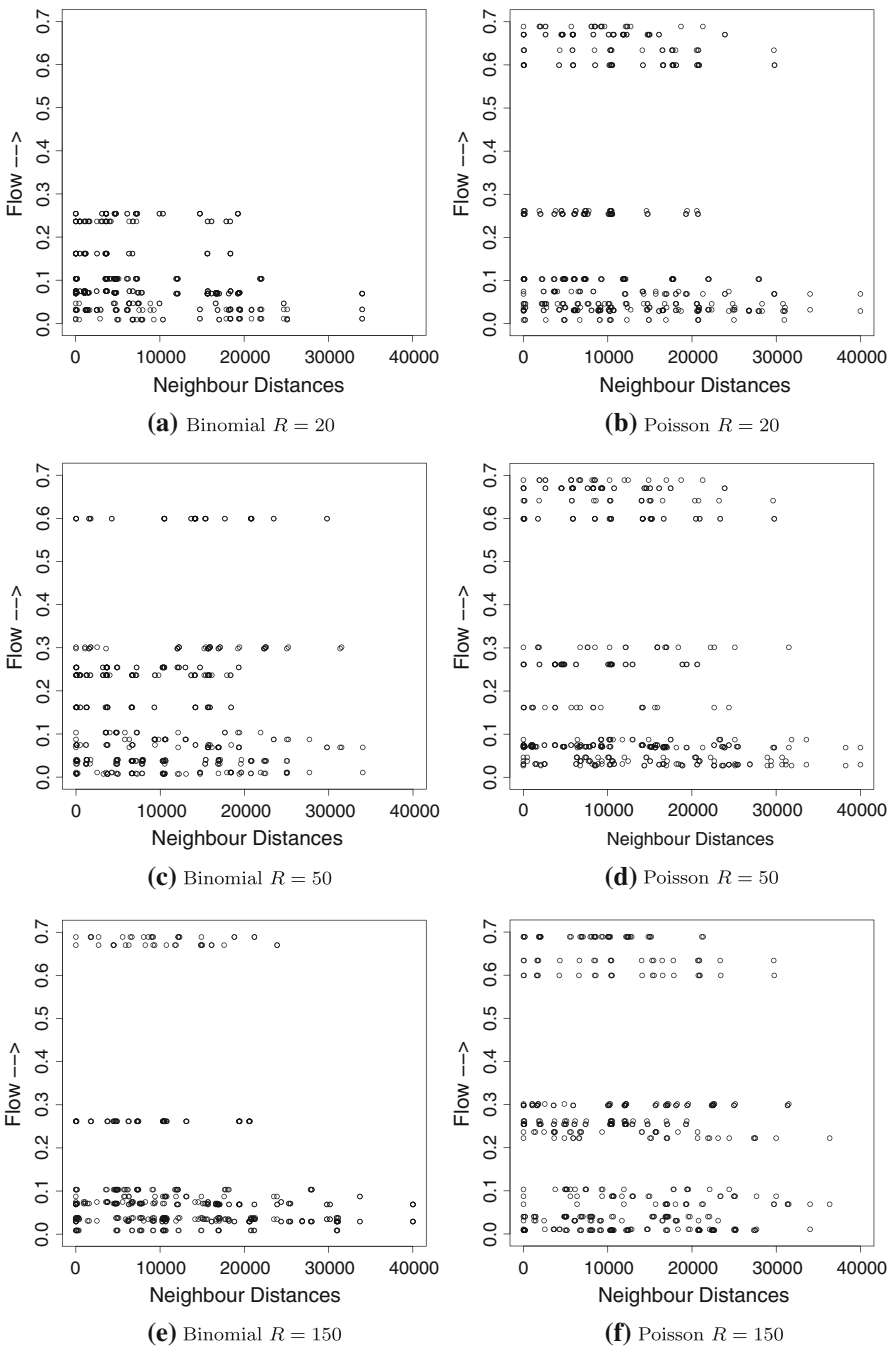
(a) Synthetic example with $P = 1$ **(b)** Lake Eacham with $P = 1$

(c) Synthetic example with $P = 30$ **(d)** Lake Eacham with $P = 30$

Fig. 7 Best subset of sampling locations plotted as flow (larger values indicate higher flow) versus neighbour distances for the synthetic example (**a**, **c**) and Lake Eacham (**b**, **d**) with varying P values ($P = 1, 30$) for the Poisson model

7 Discussion

The aim of this paper was to determine pseudo-Bayesian sample designs and subsets of sampling locations from current designs, acknowledging that accurate prediction of the variable of interest and estimating covariance parameters are two possible goals of the study. For continuous responses, the utility function constructed for prediction yielded sample subsets spread fairly uniformly over the network, while the utilities for parameter estimation found designs with some amount of clustering. Furthermore, the utility for empirical prediction unsurprisingly generated designs with a trade-off between a uniform spread of points and clustering as it sought designs with precise prediction using accurately estimated parameters. These designs were found for a synthetic example and a real dataset collected near Lake Eacham in Queensland, Australia.



(a) Binomial $R = 20$ **(b)** Poisson $R = 20$

(c) Binomial $R = 50$ **(d)** Poisson $R = 50$

(e) Binomial $R = 150$ **(f)** Poisson $R = 150$

Fig. 8 Subset of sampling locations plotted as flow (larger values indicate higher flow) versus neighbour distances for Lake Eacham with binomial (**a**, **c**, **e**) and count (**b**, **d**, **f**) measurements on the network with varying R values ($R = 20, 50, 150$)

The paper also proposed an exchange algorithm to determine these designs. A noted criticism of the exchange algorithm is that it may get stuck on local optima and overlook optimal designs depending on the initial random sample. We have repeated the optimisation several times to attempt to overcome this drawback. A simulated annealing approach is another algorithm that can avoid this problem as the acceptance step allows for swapping between different modes, but comes with additional computational expense especially for large spatial networks.

Finally, we considered designs for models of discrete responses, specifically binomial and Poisson. The utility was considered for different numbers of trials in a Bernoulli experiment or length of time that sampling took place on the network for the Poisson model. It was shown that as this number increases, the designs tend towards the space filling design which is optimal for prediction of a continuous response. Additionally, the mean of the random field was varied to determine the influence on design for the Poisson model. It can be seen that as the Poisson mean increases, the design generally reverts to the space filling design. This appears to indicate that the space filling design is robust for different types of data when the number of trials, length of time samples are collected and/or the Poisson mean is large.

Additional investigation is required to consider the influence on design of including trends in the modelling. Further work is also required to extend these pseudo-Bayesian spatial stream network sampling designs when the response is multivariate. The nature of dependencies in the response would require substantial investigation.

Acknowledgments The authors were supported by an Australian Research Council Discovery Grant. The authors would like to thank Dr. Erin Peterson and Associate Professor Zhengyuan Zhu for very helpful discussions. Additionally, the authors thank two anonymous reviewers for their constructive comments which greatly improved this paper.

References

Caselton WF, Zidek JV (1984) Optimal monitoring network designs. Stat Probab Lett 2(4):223–227

Cressie N (1993) Statistics for spatial data. Wiley, New York

Cressie N, Frey J, Harch B, Smith M (2006) Spatial prediction on a river network. J Agric Biol Environ Stat 11(2):127–150

Diggle PJ, Lophaven S (2006) Bayesian geostatistical design. Scand J Stat 33(1):53–64

Diggle PJ, Ribeiro PJ (2007) Model based geostatistics. Springer, New York

Dobbie MJ, Henderson BL, Stevens DL (2008) Sparse sampling: spatial design for monitoring stream networks. Stat Surv 2(2008):113–153

Evangelou E, Zhu Z (2012) Optimal predictive design augmentation for spatial generalised linear mixed models. J Stat Plan Inf 142(12):3242–3253

Harville DA (1974) Bayesian inference for variance components using only error contrasts. Biometrika 61(2):383–385

Li J (2009) Spatial multivariate design in the plane and on stream networks. PhD thesis, University of Iowa

Muller P (1999) Simulation based optimal design. Bayesian Stat 6:459–474

Müller WG (1998) Collecting spatial data: optimum design of experiments for random fields. Heidelberg: Physica-Verlag

Peterson EE, Ver Hoef JM (2010) A mixed-model moving-average approach to geostatistical modeling in stream networks. Ecology 91(3):644–651

Royle JA (2002) Exchange algorithms for constructing large spatial designs. J Stat Plan Inf 100(2):121–134

Shreve RL (1967) Infinite topographically random channel networks. J Geol 75(1):178–186

Spatial Reporting Of Ecosystem Health Project. (2009) CSIRO and the Queensland Department of Environment and Resource Management. Included in SSN version 1.0 R package http://www.fs.fed.us/rm/boise/AWAE/projects/SpatialStreamNetworks

Ver Hoef JM, Peterson EE (2010) A moving average approach for spatial statistical models of stream networks. J Am Stat Assoc 105(489):6–18

Ver Hoef JM, Peterson EE, Clifford D, Shah R (2014) SSN: an R package for spatial statistical modelling on stream networks. J Stat Softw 56(3):1–45

Ver Hoef JM, Peterson EE, Theobald D (2006) Spatial statistical models that use flow and stream distance. Environ Ecol Stat 13(4):449–464

Wolfinger R, O'Connell M (1993) Generalized linear mixed models: a pseudo-likelihood approach. J Stat Comput Simul 48:233–243

Zhang H (2002) On estimation and prediction for spatial generalized linear mixed models. Biometrics 58(1):129–136

Zhu Z, Stein ML (2006) Spatial sampling design for prediction with estimated parameters. J Agric Biol Environ Stat 11(1):24–44

Zimmerman DL (2006) Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction. Environmetrics 17(6):635–652

Matthew G. Falk is Lecturer in Statistics at Queensland University of Technology. Dr. Falk's interests include the application of Bayesian statistical models to real world phenomena and optimal spatial sampling designs.

James M. McGree is Senior Lecturer in Statistics at Queensland University of Technology. Dr. McGree has a variety of research interests which fall into the general categories of optimal experimental design and nonlinear modelling. He is involved with a project looking at Innovating Optimal Experimental Design through Bayesian Statistics.

Anthony N. Pettitt is Australian Research Council Professorial Fellow at Queensland University of Technology. Professor Pettitt obtained his doctorate on statistical model goodness of fit from the University of Nottingham. His main area of expertise is in applied Bayesian statistics. He has over 100 refered publications on such topics as change point problems, infectious diseases, spatial statistics, motor neuron number estimation and Bayesian computation, and has been a chief investigator on numerous ARC grants.