Bayesian design for minimising prediction uncertainty in bivariate spatial responses with applications to air quality monitoring

Senarathne SGJ∗1, Müller WG2, and McGree JM1

1School of Mathematical Sciences, Science and Engineering Faculty, Queensland University of Technology, Brisbane, Australia
2Department of Applied Statistics, Johannes Kepler University Linz, Austria

ABSTRACT

Model-based geostatistical design involves the selection of locations to collect data to minimise an expected loss function over a set of all possible locations. The loss function is specified to reflect the aim of data collection, which, for geostatistical studies, could be to minimise the prediction uncertainty at unobserved locations. In this paper, we propose a new approach to design such studies via a loss function derived through considering the entropy about the model predictions and the parameters of the model. The approach also includes a multivariate extension to generalised linear spatial models, and thus can be used to design experiments with more than one response. Unfortunately, evaluating our proposed loss function is computationally expensive so we provide an approximation such that our approach can be adopted to design realistically sized geostatistical studies. This is demonstrated through a simulated study and through designing an air quality monitoring program in Queensland, Australia. The results show that our designs remain highly efficient in achieving each experimental objective individually, providing an ideal compromise between the two objectives. Accordingly, we advocate that our approach could be adopted more generally in model-based geostatistical design.

Keywords: Bivariate response models, Copula models, Entropy, Generalised linear spatial models, Spatial dependence.

1 Introduction

The importance of spatial dynamics in natural processes has become a major focus of enquiry in many fields including ecology, agriculture and marine biology (Bloom and Kentwell, 1999; Bruno et al., 2001; Castrignanò et al., 2008; Falk et al., 2014). Often, what can be explored and ultimately inferred about such dynamics depends on how the data were collected, and, in particular, the specific locations in space (Müller, 2007). In this paper, we propose an approach in Bayesian design for selecting sampling locations to efficiently learn about the spatial dynamics underpinning natural processes. In particular, we focus on quantifying and minimising uncertainty about model predictions at unobserved locations which simultaneously provides precise estimates of model parameters.

One common approach to modelling spatial dynamics is via a Gaussian process model (Diggle et al., 2003). Such a model assumes that the observations can be regarded as a realisation from a Gaussian process \( \{S(d); d \in \mathbb{R}^2\} \). Typically, \( S(\cdot) \) is assumed to be stationary and isotropic.

*Corresponding author: jagath.gedara@hdr.qut.edu.au
with mean identically equal to zero and covariance function \( \text{Cov}(h) = \text{Cov}(S(d), S(d')) \) where \( h = ||d - d'|| \) is the Euclidean distance between locations \( d \) and \( d' \). For ease of notation, the \( i \)th observation \( S(d_i) \), where \( d_i \) is the location of the \( i \)th observation, will be abbreviated to \( s_i \) throughout this article. Through such a model, one can then leverage information about the spatial variability between locations to yield predictions at unobserved locations.

Bayesian inference provides a rigorous framework to quantify and handle uncertainty in spatial predictions, and a number of authors have considered this framework to design geostatistical studies. The work of Diggle and Lophaven (2006) compared the prediction performance of different classes of designs such as the ‘lattice plus close pairs’ and ‘lattice plus in-fill’ designs under parameter uncertainty. As such, no optimisation of the design was undertaken. This is presumably because of the large computational time involved in evaluating their loss function, and this is typical of designs for prediction. A pseudo-Bayesian design approach was proposed by Falk et al. (2014) for sampling on stream networks. Optimal designs were found under various loss functions but their work was limited to assuming some or all parameter values were known \textit{a priori}. Bayesian approaches to design monitoring networks were proposed by Müller et al. (2004) and Fuentes et al. (2007) such that accurate predictions could be obtained while minimising the cost of monitoring. However, both approaches do not take into account parameter uncertainty. Further, to form a dual-objective loss function, both approaches considered a linear combination of loss functions which requires specifying the relative importance of each objective via pre-defined weights. Unfortunately, such an approach has been shown to be difficult to apply in practice (Hill et al., 1968; Cook and Wong, 1994; McGree et al., 2008). For a similar purpose, Le et al. (2003) proposed an entropy-based criterion for designing networks for monitoring multivariate environmental fields. However, their approach can only be applied to network problems where the responses are approximately normally distributed, and hence has limited applicability.

Across all of the design methods proposed in the above cited papers, only a limited number of approaches have taken into account both parameter uncertainty and the uncertainty in the predicted outcomes when assessing a design. This is potentially a major limitation as both sources of uncertainty could be significant. Further, most of these approaches were limited to consider spatially dependent univariate responses or multivariate responses that are marginally Gaussian. This is potentially because of the difficulty in constructing a multivariate distribution that appropriately describes the dependencies between each response. As each response may be of a different type (i.e. continuous, count, binary, etc), this can lead to a rather complex model, rendering many approaches in Bayesian design computationally infeasible. Such a limitation seems rather restrictive as multiple responses are often observed in geostatistical studies e.g. Bohorquez et al. (2017); Musafer and Thompson (2017).

To address the limitations of previous research, we consider a generalised linear spatial modelling (GLSM) framework where the dependence between responses is described by a Copula model. Such functions are defined based on the cumulative distribution function (CDF) of the univariate responses, meaning that the joint distribution can be defined for a variety of different data types. Then, given such a model, we propose to quantify the uncertainty in spatial predictions via an entropy-based loss function. The benefit of this is that one can then exploit the additivity property of entropy, avoiding the need to pre-specify weights on different objectives like estimation and prediction. To demonstrate our approach, we design a simulated study and also an air quality monitoring program in Queensland, Australia, and assess the performance of the resulting designs.

This paper is outlined as follows. In Section 2, the GLSM framework for modelling spatial outcomes is defined, and the Copula representation for describing multivariate spatial data is introduced. Our Bayesian design framework is described in Section 3, along with our approach.
to quantify and minimise uncertainty in model predictions. In doing so, we show how such an approach also minimises uncertainty about parameter estimates, and provide an approximation to efficiently evaluate this loss function. To illustrate our methodologies, Section 4 focuses on finding designs in two examples where bivariate mixed spatial outcomes are observed. The paper concludes with a discussion of key findings and suggestions for future research.

2 Motivating example

Exposure to high levels of air pollution can have a variety of adverse effects on human health, including respiratory infections, heart disease, and lung cancer (Watson, 2006; Jiang et al., 2016; Zhang et al., 2016). This has led to an initiative by the Queensland government in Australia to monitor air quality across the state to provide information for managing and ultimately reducing air pollution. Key indicators of air quality that are measured across the state include $NO_2$ and $PM_{2.5}$ concentrations which are known to have significant adverse effects on human health (Roberts et al., 2019; Huang et al., 2019). However, such monitoring is hugely expensive, particularly in terms of setting up and maintaining the monitoring stations. Consequently, there is recent interest in exploring the impact of reduced sampling across the network, and how this might affect regulator’s ability to predict air pollution at unsampled locations of interest. This provides motivation for the research proposed in this paper where a retrospective sampling design is sought across the current network of air quality monitoring stations in Queensland.

2.1 Available data

The Queensland air quality monitoring network consists of 22 monitoring stations; seven of which measure both $NO_2$ and $PM_{2.5}$ concentrations. Figure 1 shows the spatial configuration of these 22 stations where we will refer to stations that measure both $NO_2$ and $PM_{2.5}$ concentrations as “sampled” locations and stations that do not measure both of these concentrations as “unsampled” locations.

For the sampled locations, the annual mean $NO_2$ concentrations ($Y_1$), the number of days (per annum) which exceed the daily $PM_{2.5}$ limit of $10\mu g m^{-3}$ ($Y_2$), and meteorological data from year 2013 to 2016 were sourced from the Queensland government website (https://data.qld.gov.au/dataset). These data were then used to develop a generalised linear spatial model to describe the observed variability in both pollutants. This model was then used to inform the selection of proposed monitoring stations within the retrospective design.
3 Modelling bivariate responses in a geostatistical study

In this section, we describe our modelling framework for bivariate responses collected in a geostatistical study. We start by first describing how a univariate response could be modelled, then extend to bivariate responses.

3.1 Spatial model for a univariate response

To model a univariate response collected in a geostatistical study, we consider a GLSM which has the following form:

\[ \mu_i = X_i^T \beta + s_i \quad \text{and} \quad E(y_i|s_i) = g^{-1}(\mu_i), \text{ for } i = 1, 2, \ldots, n, \]

where \( X_i = (1, X_{i1}, \ldots, X_{ip-1})^T \) is the covariate vector associated with the location \( d_i \), \( \beta \in \mathbb{R}^p \) are the regression coefficients (fixed effects), and \( s_i \) is the value of the \( i \)th random effect. Here, the link function \( g(.) \) defines the relationship between the linear predictor \( \mu_i \) and the expected outcome given \( s_i \).

In a GLSM, conditional on the random effects, the dependent variable \( Y \) follows a distribution from the exponential family. The joint distribution of the random effects \( s_1, s_2, \ldots, s_n \) is assumed to be Multivariate Normal with mean vector zero and covariance matrix \( \Sigma \). When formulating this covariance matrix, it is convenient to consider a specific parametric family of covariance functions (Albert and McShane, 1995; Diggle et al., 2003). In this paper, the Matérn covariance function is used, and has the following form:
\[
\text{Cov}(h; \gamma, \nu) = \gamma_1 \rho(h, \gamma_2, \nu) \quad \text{and} \quad \rho(h; \gamma_2, \nu) = \frac{1}{2^{\nu-1} \Gamma(\nu)} \left( \frac{h}{\gamma_2} \right)^\nu K_\nu(h/\gamma_2),
\]

where \( \nu > 0 \), \( \gamma = (\gamma_1, \gamma_2) \) and \( h \) is the Euclidean distance between two locations. \( \Gamma(\cdot) \) is the gamma function and \( K_\nu(\cdot) \) is the modified Bessel function of order \( \nu \). The parameters \( \gamma_1 \) and \( \gamma_2 \) are the partial sill and the spatial range, respectively. Here, we define the covariance function at the linear predictor level. Additional variability will be assumed at the data level, the form of which will depend on the data type.

To describe a Bayesian framework for the above model, let \( y_i = (y_{i1}, \ldots, y_{im})^T \) denote the observed data at location \( d_i \), where \( n_i \) is the number of observations collected at location \( d_i \), for \( i = 1, \ldots, n \).

In spatial design, the design \( d = (d_1, \ldots, d_n)^T \in \mathcal{D} \) represents the locations where the outcomes \( y = (y_1, \ldots, y_n)^T \) are measured, and \( \mathcal{D} \) represents the design space for the study. Let \( p(\theta) \) denote the prior distribution about the parameters \( \theta \). Then, within a Bayesian framework, all inferences are based on the posterior distribution defined as follows:

\[
p(\theta | y, d) \propto \int \prod_{i=1}^n \prod_{j=1}^{n_i} p(y_{ij} | d_i, s_i, \beta) p(s | \gamma, \nu) p(\theta) \, ds,
\]

where \( s = (s_1, \ldots, s_n)^T \) is vector of random effects, and the parameter vector \( \theta \) includes both model parameters \( \beta \) and covariance parameters \( \gamma \) and \( \nu \). \( p(y_{ij} | d_i, s_i, \beta) \) is the conditional likelihood of observing \( y_{ij} \) at location \( d_i \) given the model parameters \( \beta \) and random effects \( s_i \) for \( i = 1, \ldots, n \) and \( j = 1, \ldots, n_i \).

To extend the above model to handle bivariate responses, we consider Copula models to describe the dependence between responses (Nelsen, 2006; Bárdossy, 2006; Kazianka and Pilz, 2011; Gräler and Pebesma, 2011) (see next section). Adopting such an approach provides a framework to flexibly model dependence between bivariate responses.

### 3.2 Spatial model for bivariate responses

Here, we define a modelling approach for bivariate spatial responses by combining the univariate spatial models (defined above) via a suitable Copula model. For this, assume that two response variables will be observed with one being continuous \( Y_1 \) and the other being discrete \( Y_2 \); thus marginally they can be modelled via two GLSMs as follows:

- \[
\mu_{1i} = X_i^T \beta_1 + s_{1i}, \quad E(y_{1i} | s_{1i}) = g_1^{-1}(\mu_{1i}),
\]
- \[
\mu_{2i} = X_i^T \beta_2 + s_{2i}, \quad E(y_{2i} | s_{2i}) = g_2^{-1}(\mu_{2i}) \quad \text{for } i = 1, 2, \ldots, n.
\]

We assume that \( Y_1 \) and \( Y_2 \) have marginal probability distributions \( f_{Y_1 | s_{1i}} \) and \( f_{Y_2 | s_{2i}} \), given the random effects \( s_{1i} \) and \( s_{2i} \), respectively. Further, denote the marginal cumulative distribution function of \( Y_1 | s_{1i} \) and \( Y_2 | s_{2i} \) as \( F_{Y_1 | s_{1i}} \) and \( F_{Y_2 | s_{2i}} \), respectively. Then, the Copula representation of the joint distribution \( G_{Y_1 | s_{1i}, Y_2 | s_{2i}} \) is given by,

\[
G_{Y_1 | s_{1i}, Y_2 | s_{2i}}(y_{1ij}, y_{2ij}) = C(F_{Y_1 | s_{1i}}(y_{1ij}), F_{Y_2 | s_{2i}}(y_{2ij}); \alpha)
\]

\[
= C(u_{1ij}, u_{2ij}; \alpha) \quad \text{for } i = 1, 2, \ldots, n \text{ and for } j = 1, 2, \ldots, n_i,
\]

where \( C \) and \( \alpha \) denote the Copula function and the Copula parameter, respectively. The Copula parameter \( \alpha \) defines the strength of dependence within the bivariate distribution (Nelsen, 2006).

From this, the Copula representation of the joint distribution of \( Y_1 | s_{1i} \) and \( Y_2 | s_{2i} \) is given by,
\[ f_{Y_i|s_1,s_2}(y_{1ij}, y_{2ij}) = f_{Y_i|s_1}(y_{1ij})(c_{1ij} - c_{1ij}^*), \]  

(1)

where \( c_{1ij} = \frac{\partial C(u_{1ij}, u_{2ij}; \alpha)}{\partial u_{1ij}} \) and \( c_{1ij}^* = \frac{\partial C(u_{1ij}, u_{2ij}; \alpha)}{\partial u_{1ij}} \) and \( u_{2ij} \) is the left hand limit of \( u_{2ij} \), see Joe (2014) and Tao et al. (2013) for further details.

Then, using Equation (1), the conditional likelihood of a bivariate mixed outcome \( \mathbf{y} = (y_{1ij}, y_{2ij}) \) for \( i = 1, 2, \ldots, n \) and \( j = 1, 2, \ldots, n_i \) can be expressed as follows:

\[
p(\mathbf{y}|d, \theta, s_1, s_2) = \prod_{i=1}^{n} \prod_{j=1}^{n_i} \left[ \left( f_{Y_i|s_1}(y_{1ij}) \right) \left( c_{1ij} - c_{1ij}^* \right) \right],
\]

where \( \theta \) includes all the model parameters \( (\beta_1, \beta_2) \), the covariance parameters \( (\gamma_1 = (\gamma_{11}, \gamma_{21}), \gamma_2 = (\gamma_{21}, \gamma_{22}), \nu_1, \nu_2) \), and the Copula parameter \( \alpha \).

The bivariate Copula function describes the dependence structure between two random variables. A large number of bivariate Copulas and their dependence properties have been discussed in the literature (Durante and Sempi, 2010; Genest and MacKay, 1986; Nelsen, 2006). Among them, the bivariate Archimedean Copula models such as Clayton, Gumbel and Frank Copulas have been widely used due to their flexibility in describing different tail behaviours and different dependence structures.

The Clayton Copula is considered in this paper to model the joint distribution of the responses with the reader referred to Clayton (1978) and Cook and Johnson (1981) for further details. Let \( Y_1 \) and \( Y_2 \) be two responses with CDFs \( u_1 = F_{Y_1}(y_1) \) and \( u_2 = F_{Y_2}(y_2) \), respectively, then the joint CDF of \( Y_1 \) and \( Y_2 \) can be expressed using the Clayton Copula as follows:

\[
C(u_1, u_2; \alpha) = (u_1^{-\alpha} + u_2^{-\alpha} - 1)^{-1/\alpha}; \quad \alpha > 0,
\]

where \( \alpha \) is the Copula parameter.

The choice of this Copula model was motivated by the tail dependence structure of the data observed in the motivating example and desirable properties of the Archimedean Copula family to which this Copula belongs. For example, as provided by Genest and MacKay (1986), there is a closed form relationship between the bivariate Archimedean Copula parameter and Kendall’s tau \( (\tau) \). Therefore, once the Copula parameter is estimated, it is straightforward to define the dependence between the two responses within the intuitive scale of \(-1\) to \(+1\). For example, the relationship between the Clayton Copula parameter \( \alpha \) and Kendall’s tau can be defined as follows:

\[
\tau = 4E[C(u_1, u_2; \alpha)] - 1 = \frac{\alpha}{\alpha + 2}.
\]

4 Bayesian design framework for geostatistical studies

Our aim of minimising the uncertainty in spatial predictions can be quantified within a Bayesian design framework by a loss function which we will denote as \( \lambda(d, \theta, \mathbf{y}) \). Such a loss function compares a summary of the posterior distribution for \( \theta \) (based on observing \( \mathbf{y} \) from \( d \)) with its true value. However, as this function depends on \( \theta \) and \( \mathbf{y} \), which are unknown \textit{a priori}, it cannot be used to select designs. Accordingly, the expectation of the loss function is used to locate designs, and this expectation can be defined as follows:

\[
\tau = 4E[C(u_1, u_2; \alpha)] - 1 = \frac{\alpha}{\alpha + 2}.
\]
\[ L(d) = E_{y, \theta}[\lambda(d, \theta, y)] = \int \int \lambda(d, \theta, y) p(y|\theta, d) p(\theta) d\theta dy, \tag{2} \]

where \( p(y|\theta, d) = \int_{s \in S} p(y|\beta, s, d) p(s|\gamma, \nu) ds \).

When the loss function does not depend on the model parameters \( \theta \), Equation (2) can alternatively be expressed as:

\[ L(d) = E_y[\lambda(d, y)] = \int \lambda(d, y) p(y|d) dy. \tag{3} \]

A Bayesian design is then found by minimising the expected loss over the space of all possible locations. However, in general, the expected loss function does not have a closed-form solution, and hence, needs to be approximated. Monte Carlo (MC) integration is the most commonly used approach for this approximation. This is achieved by simulating a large number of prior predictive data sets, evaluating the loss function for each data set, and then taking the average as the approximation to the expected loss. Formally:

\[ \hat{L}(d) = \frac{1}{K} \sum_{k=1}^{K} \lambda(d, \theta_k, y_k), \tag{4} \]

where \( \theta_k \) and \( y_k \) are generated from the distributions \( p(\theta) \) and \( p(y|d, \theta_k) \), respectively.

The current approach to Bayesian design for spatial prediction is based on a loss function proposed in Diggle and Lophaven (2006). This loss function quantifies the spatially averaged prediction variance of the unobserved random field \( S(\cdot) \) over the predicted region \( A \) as follows:

\[ \lambda_{\text{pred}}(d, y) = \int_{\xi \in A} \text{Var}\{S(\xi)|y, d\} d\xi. \tag{5} \]

When the prediction region \( A \) consists a discrete set of locations \( \xi_1, \xi_2, \ldots, \xi_T \), the above loss function can be expressed as follows:

\[ \lambda_{\text{pred}}(d, y) = \frac{1}{T} \sum_{t=1}^{T} \text{Var}\{S(\xi_t)|y, d\}. \]

As discussed in Diggle and Lophaven (2006), this loss function provides some provision to also address the objective of parameter estimation. In the next section, we describe an alternative approach for addressing these dual objectives.

### 4.1 An entropy-based loss function for spatial prediction

Here, we derive a loss function to quantify uncertainty in a spatial process by considering the entropy about the predictions, following the principles outlined in the seminal book by Le and Zidek (2006). For this, we note that, for a given model, there are two sources of uncertainty about the predictions: (1) Uncertainty in the parameter values; and (2) Uncertainty in the outcome \( Z \) conditional on the parameter values. Thus, to derive this loss function, we start by considering the entropy in these two random variables \emph{a priori}. Initially we consider a univariate outcome \( Z \) where the entropy about this random variable and the parameter values is:
Following this, we define the loss function in terms of the change in entropy about \( Z \) and the parameters upon observing data \( y \) at design \( d \) as follows:

\[
\lambda_D(d, y) = H(Z, \theta|y, d, \xi) - H(Z, \theta|\xi),
\]

where \( H(Z, \theta|y, d, \xi) \) is the entropy about \( Z \) and the parameters upon observing data \( y \) at design \( d \).

Using the chain rule of entropy, it is straightforward to show that the joint entropy of \( Z \) and the parameters is equal to the conditional entropy of \( Z \) given the parameters plus the entropy of the parameters. That is:

\[
H(Z, \theta|y, d, \xi) = H(Z|\theta, y, d, \xi) + H(\theta|y, d), \quad \text{and} \quad H(Z, \theta|\xi) = H(Z|\theta, \xi) + H(\theta),
\]

where, when \( Z \) is a discrete random variable, \( H(Z|\theta, \xi) = -\sum_{z \in \mathcal{Z}} p(z|\theta, \xi) \log p(z|\theta, \xi) \) and \( H(\theta) = -\int p(\theta) \log p(\theta) d\theta \).

Then, by substituting the above expressions into Equation (6), the loss function \( \lambda_D(d, y) \) can be expressed as follows:

\[
\lambda_D(d, y) = \{ H(Z|\theta, y, d, \xi) + H(\theta|y, d) \} - \{ H(Z|\theta, \xi) + H(\theta) \}
\]

\[
= \{ H(Z|\theta, y, d, \xi) - H(Z|\theta, \xi) \} + \{ H(\theta|y, d) - H(\theta) \}
\]

\[
= \lambda_P(d, y) + \lambda_E(d, y).
\]

As shown in Equation (8), the loss function \( \lambda_D(d, y) \) can be expressed as a sum of two loss functions in which the first (\( \lambda_P(d, y) \)) quantifies the change in entropy about \( Z \) given the parameters \( \theta \) while the second (\( \lambda_E(d, y) \)) quantifies the change in entropy about the parameter values. Thus, the loss function \( \lambda_D(d, y) \) is termed as a dual-purpose loss function for parameter estimation and prediction. The derivation of Equation (8) is essentially the same as the fundamental identity given in Le and Zidek (2006).

For parameter estimation, the loss function \( \lambda_E(d, y) \) is equivalent to the Kullback-Leibler divergence (KLD) between the prior and the posterior distributions of parameters (Kullback and Leibler, 1951), and can be expressed as follows:

\[
\lambda_E(d, y) = -\int_\Theta p(\theta|y, d) \log \left( \frac{p(\theta|y, d)}{p(\theta)} \right) d\theta.
\]

When comparing designs via the loss function \( \lambda_P(d, y) \), the term \( H(Z|\theta, \xi) \) is independent of the design \( d \). As such, for simplicity, we consider \( H(Z|\theta, y, d, \xi) \) as our loss function for prediction given the parameters as follows:
However, when assessing the performance of designs under this expected loss, the term \( H(Z|\theta, \xi) \) is subtracted so that the resulting values are interpretable.

When a bivariate mixed outcome (i.e. one outcome is continuous and the other is discrete) is observed \( (z = (z_1, z_2)) \), the above loss function can be expressed as follows:

\[
\lambda_P(d, y) = -\int_{\Theta} p(\theta|y, d) \sum_{z \in Z} p(z|\theta, y, d, \xi) \log p(z|\theta, y, d, \xi) dz d\theta.
\]

(10)

where \( y = (y_1, y_2) \) contains data collected for the two responses at the locations \( \xi \) based on design \( d \).

The loss function \( \lambda_{E}(d, y) \) for the bivariate mixed outcome is defined as above, where \( \theta \) contains all model parameters for both responses.

5 Efficient approximations to loss functions

Evaluating the MC approximation to the expected loss in Equation (11) requires approximating or sampling from a large number of posterior distributions. Unfortunately, this renders algorithms like Markov chain Monte Carlo computationally infeasible to use in locating designs. Accordingly, in this section, we describe computationally efficient methods for approximating the above loss functions.

5.1 Approximating the posterior distribution

To efficiently evaluate the MC approximation to the expected loss function in Equation (11), fast methods for approximate inference are required. For this purpose, we consider the Laplace approximation to the posterior distribution, see Long et al. (2013); Overstall, McGree and Drovandi (2018). For this, we assume, approximately, that:

\[
(\theta|y, d) \sim MVN(\theta^*, A(\theta^*)^{-1}),
\]

where \( \theta^* = \arg \max_{\theta} \{ \log p(y|d, \theta) + \log p(\theta) \} \) and \( A(\theta^*) \) is the negative Hessian matrix:

\[
A(\theta^*) = -\frac{\partial^2 \{ \log p(y|d, \theta) + \log p(\theta) \}}{\partial \theta \partial \theta'} \bigg|_{\theta = \theta^*}.
\]

In our spatial model, the linear predictor contains random effects, and these need to be integrated out in order to evaluate the likelihood. That is:

\[
p(y|d, \theta) = \int_{s_1} \cdots \int_{s_2} \cdots \int_{s_{n}} \prod_{i=1}^{n_1} \prod_{j=1}^{n_2} \left[ (f_{y_i|s_{1i}(y_{1ij})}) (c_{1ij} - c_{1ij}^*) \right] p(s_1|\gamma_1, \nu_1) \times \]

\[
p(s_2|\gamma_2, \nu_2) ds_1 \cdots ds_{n_1} ds_{s_2} \cdots ds_{n_2}.
\]

Unfortunately, given the form of our model, there will typically be no closed-form solution to the above integral, and therefore the likelihood. To handle this, we again employ MC integration.
to approximate this integral by simulating random effects as follows \( s_{1b} \sim p(s_1|\gamma_1, \nu_1), \ s_{2b} \sim p(s_2|\gamma_2, \nu_2) \) for \( i = 1, 2, \ldots, n \) and \( b = 1, 2, \ldots, B \), and we note that a variety of approximations could be used, see for example Gómez-Rubio et al. (2019). Using MC integration, the likelihood can be approximated as follows:

\[
p(y|d, \theta) \approx \frac{1}{B} \sum_{b=1}^{B} \prod_{i=1}^{n} \prod_{j=1}^{m} p(y_{ij}|d_i, \theta, s_{1b}, s_{2b}).
\] (12)

It is this approximation to the likelihood that will be used to form a Laplace approximation to the posterior distribution (as shown above).

### 5.2 Approximating the loss function

When both the prior and posterior distributions follow Multivariate Normal distributions with mean vectors \((\theta_0^*, \theta^*)\) and covariance matrices \((\Omega_0, \Omega)\), respectively, the loss function \(\lambda_E(d, y)\) can be evaluated as follows:

\[
\hat{\lambda}_E(d, y) = -\frac{1}{2} \left( tr(\Omega_0^{-1}\Omega) + (\theta^* - \theta_0)^T \Omega_0^{-1}(\theta^* - \theta_0) - k + \log \left( \frac{\det \Omega_0}{\det \Omega} \right) \right),
\] (13)

where \(k\) is the dimension of the two Multivariate Normal distributions.

In terms of approximating \(\lambda_P(d, y)\), this proves to be a little more complicated as summations need to be taken across simulated values of \(\theta, y\) and \(z\). That is:

\[
\hat{\lambda}_P(d, y) = -\frac{1}{K} \sum_{k=1}^{K} \frac{1}{R} \sum_{r=1}^{R} \log \left( \frac{1}{B} \sum_{b=1}^{B} p(z_r|s_b, \theta_k, y, d, \xi) \right),
\] (14)

where \(\theta_k \sim p(\theta|y, d), s_b \sim p(s|\theta_k, y, d)\) and \(z_r \sim p(z|s_b, \theta_k, y, d, \xi)\), and the posterior distribution is approximated via a Laplace approximation.

In using this approach, in order to precisely estimate this loss function, a sufficiently large number of samples should be taken from the posterior distributions of the parameters, random effects, and the posterior predictive distribution (i.e. all \(B, K\) and \(R\) should be sufficiently large). Thus, evaluating this loss function is hugely expensive computationally. Indeed, if each integral requires \(N\) MC samples, the computation time is of order \(N^3 (\mathcal{O}(N^3))\). Unfortunately, this renders such an approach computationally infeasible for use in locating designs (even when adopting a Laplace approximation to the posterior distribution). To overcome this, we consider forming a Multivariate Normal approximation to the joint distribution of simulated data as the entropy for such a distribution has a closed form. We propose forming this approximation via moment matching. That is, in the two examples considered in this paper, a bivariate mixed outcome is observed where a Normal response \((Y_1)\) and a Poisson response \((Y_2)\) are considered. Given the mean value for the Poisson response is reasonably large, the joint distribution of \(Y_1\) and \(Y_2\) given \(d\) and \(\xi\) can be approximated by a bivariate Normal distribution. Given this, the entropy of bivariate Normal response was considered to approximate the loss function given in Equation (11), and therefore the first term of the total loss function given in Equation (8) was approximated as follows:

\[
\tilde{\lambda}_P(d, y) = \frac{1}{2} \sum_{\xi \in \xi} \log \det (2\pi e \hat{\Sigma}_\xi) = T(1 + \log(2\pi)) + \sum_{\xi \in \xi} \log \det (\hat{\Sigma}_\xi),
\] (15)
where $\Sigma_{\xi_t}$ is the approximated variance-covariance matrix of the bivariate Normal distribution at location $\xi_t$, for $t = 1, 2, \ldots, T$. Here, $\Sigma_{\xi_t}$ was approximated by taking a sufficiently large sample from the posterior predictive distribution $p(z | s, \theta, y, d, \xi_t)$. Thus, this approximation has complexity of order $N$ i.e. $O(N)$, making it substantially more efficient to compute than the approximation in Equation (14). We explore the accuracy and computational benefits of this approximation in Section 5.

Given the above two approximations, the dual-purpose loss function can be approximated as follows:

$$\hat{\lambda}_D(d, y) = \hat{\lambda}_P(d, y) + \hat{\lambda}_E(d, y).$$  \hspace{1cm} (16)

All the loss functions discussed above depend on either a set of discrete choices for $\xi$ or the space of $\xi$. Throughout this paper, we consider a set of discrete choices, and thus we average the utility across these when approximating the expected utility. If we were to consider a space for $\xi$, then the utility would need to be integrated over this space. In cases where there is no analytic solution to this integration, then one could approximate the solution via averaging over a random selection of locations.

### 5.3 Design algorithm

With the use of the Laplace approximation for approximating the posterior distribution and the above approximations to the expected loss, we propose Algorithm 1 to derive optimal designs for spatially dependent bivariate outcomes described by the Copula-based GLSM.

**Algorithm 1** Approximating the expected loss function for the location of geostatistical designs

1: Initialise the prior information $p(\theta)$ and the prediction set $\mathcal{S}$.
2: for $k = 1$ to $K$ do
3: \hspace{1cm} Draw $\theta_k = (\beta_{1k}, \beta_{2k}, \gamma_{1k}, \gamma_{2k}, \nu_{1k}, \nu_{2k}, \alpha_k) \sim p(\theta)$
4: \hspace{1cm} Draw $s_k = (s_{1k}, s_{2k}) \sim p(s | \gamma_{1k}, \gamma_{2k}, \nu_{1k}, \nu_{2k})$
5: \hspace{1cm} Simulate data $y_k$ at the design $d$ from the assumed Copula model $y_k \sim p(y | s_k, \theta_k, d)$
6: \hspace{1cm} Form Laplace approximation via $\theta^*_k = \arg \max_{\theta} \{ \log p(y_k | \theta, d) + \log p(\theta) \}$ such that the posterior distribution can be approximated by $MVN(\theta^*_k, A_k^{-1})$, where $A_k$ is the Hessian matrix defined as:
7: \hspace{2cm} $A_k = \left. -\partial^2 \{ \log p(y_k | \theta, d) + \log p(\theta) \} \right|_{\theta=\theta^*}$.
8: \hspace{1cm} Evaluate $\lambda_E(d, y_k)$ via Equation (13) where $\Omega = A_k^{-1}$
9: \hspace{1cm} Simulate data $z_b$ at $\xi$ via $z_b \sim p(z | s_b, \theta_b, \xi)$, where $\theta_b = (\beta_{1b}, \beta_{2b}, \gamma_{1b}, \gamma_{2b}, \nu_{1b}, \nu_{2b}, \alpha_b) \sim p(\theta | y_k, d)$ and $s_b \sim p(s | \gamma_{1b}, \gamma_{2b}, \nu_{1b}, \nu_{2b})$, for $b = 1, \ldots, B$
10: \hspace{1cm} Approximate $\lambda_P(d, y_k)$ via Equation (15)
11: \hspace{1cm} Approximate $\lambda_D(d, y_k)$ via Equation (16)
12: end for
13: Approximate the expected loss $\hat{L}(d) = \frac{1}{K} \sum_{k=1}^{K} \lambda_D(d, y_k)$
14: Find optimal design $d^* = \arg \min_{d \in \mathcal{S}} \hat{L}(d)$

Implementing this algorithm requires defining a model for analysis, the prior information about
the parameters and the prediction region (line 1). Once defined, at each iteration, a parameter \( \theta_k = (\beta_{1k}, \beta_{2k}, \gamma_{1k}, \gamma_{2k}, \nu_{1k}, \nu_{2k}, \alpha_k) \) is drawn from the prior distribution (line 3). Then, \((\gamma_{1k}, \gamma_{2k}, \nu_{1k}, \nu_{2k})\), a random effect \( s_k \) is simulated (line 4). Next, data \( y_k \) are simulated from the assumed Copula model (line 5). The posterior distribution of the parameters is then approximated via the Laplace approximation (line 6). This requires finding the posterior mode and evaluating the Hessian matrix at this mode. Here, we found the mode via a gradient-based optimisation algorithm, and the Hessian matrix was evaluated numerically. When the dual-purpose loss function is considered for design selection, \( \lambda_E(d, y_k) \) and \( \lambda_P(d, y_k) \) in lines 7 and 9 of Algorithm 1 can be approximated via Equations (13) and (15), respectively. The approximation of \( \lambda_P(d, y_k) \) requires generating data from the posterior predictive distribution as shown in line 8 of Algorithm 1. Following this, the dual-purpose loss function can be approximated via Equation (16) (line 10). Then, once \( K \) iterations have been completed, the MC approximation to the expected loss is evaluated in line 12. To find the optimal set of locations, a suitable optimisation algorithm is implemented to minimise the approximate expected loss function (line 13).

6 Case studies

A simulated example and the motivating example described in Section 2 were considered to demonstrate the performance of Algorithm 1 and the dual-purpose loss function derived in Section 4 to design geostatistical studies. In both examples, bivariate mixed spatial outcomes are observed. As such, we first model the univariate spatial outcomes via a GLSM, and then combine these via the Clayton Copula model as described in Section 3.2.

In Example 1, we simulate data within a unit square with a Matérn covariance function under parameter uncertainty. As designs can vary depending upon the strength of spatial dependence, three different settings (weak, moderate and strong) were considered. In Example 2, we apply the proposed approach to find a retrospective design for the Queensland air quality monitoring network.

In both of these examples, we consider the case where a practitioner needs to select the most appropriate set of locations to collect data from within a region or finite set of locations. To explore this, we evaluated the effectiveness of selecting dual-purpose designs compared to designs that only address a single aim. For this, we first obtained prediction only and estimation only designs using the loss functions given in Equation (5) and Equation (9), respectively. Then, the dual-purpose designs were evaluated against these single-purpose designs to assess performance. As Example 1 considers a continuous design space, the approximate coordinate exchange (ACE) algorithm (Overstall and Woods, 2017) was used to find optimal designs. Here, default settings as detailed in Overstall, Woods and Adamou (2018) were used. As Example 2 was undertaken across a restricted design space with only a fixed number of locations being available, the standard coordinate exchange algorithm (Meyer and Nachtsheim, 1995) was used to obtain optimal designs. All simulations were run using R 3.5.2, and R code to reproduce the results in this paper is available via the following GitHub repository, https://github.com/SenarathneSGJ/Model-based_geostatistical_design.

6.1 Example 1: Exploring dual-purpose design for spatial processes

In this example, consider collecting bivariate outcomes across a spatial process. Here, \( Y_1 \) was assumed to follow a Normal distribution given the random effects \( s_1 \) as follows:

\[
Y_1|s_1 \sim N(\mu_1, \sigma_1^2) \quad \text{and} \quad \mu_1 = \beta_{10} + \beta_{11}X_1 + \beta_{12}X_2 + s_1, \quad (17)
\]
where \((\beta_{10}, \beta_{11}, \beta_{12})\) are the model parameters, \(X_1\) and \(X_2\) are the two covariates in the model and \(\sigma^2_i\) is the residual variance of \(Y_i\). The random effects \(s_i\) were assumed to follow a Multivariate Normal distribution with mean 0 and covariance matrix \(\Sigma_1\), where the elements of \(\Sigma_1\) were obtained from a Matérn covariance function with parameters \((\gamma_1, \nu)\) as described in Section 3.

\(Y_2\) was assumed to follow a Poisson distribution given the random effects \(s_2\) as follows:

\[
Y_2|s_2 \sim \text{Pois}(\exp(\mu_2)) \quad \text{and} \quad \mu_2 = \beta_{20} + \beta_{21}X_1 + \beta_{22}X_2 + s_2,
\]

(18)

where \((\beta_{20}, \beta_{21}, \beta_{22})\) are the model parameters, and \(X_1, X_2\) are two covariates of the model. Again, the random effects \(s_2\) were assumed to follow a Multivariate Normal distribution with mean 0 and covariance matrix \(\Sigma_2\). The elements of the covariance matrix \(\Sigma_2\) were obtained from a Matérn covariance function with parameters \((\gamma_2, \nu_2)\).

Next, the model for the bivariate responses \(Y_1\) and \(Y_2\) was obtained using the Clayton Copula function as detailed in Section 3. It was assumed that accurate predictions were required at 25 locations \((\xi)\) defined as follows:

\[
\xi_{vw} = (0.25v, 0.25w) \text{ for } v, w = 0, 1, \ldots, 4.
\]

To investigate the performance of our design algorithm and the dual-purpose loss function for obtaining designs under various spatial conditions, three design scenarios were considered. These design scenarios differ in terms of the strength of the spatial covariance, constructed via three different values for the prior mean of the spatial range parameter as given in Table 1.

| Scenario | Strength of dependence | Prior mean of range parameter (a) |
|----------|------------------------|----------------------------------|
| 1        | Weak                   | 0.2                              |
| 2        | Moderate               | 0.5                              |
| 3        | Strong                 | 0.8                              |

The prior distributions of the remaining parameters were same across all the scenarios. For all the model and covariance parameters, Normal priors were considered as defined in Table 2. Further, we assumed that there is a positive dependence between the two responses, and therefore, a Normal prior was considered for the parameter \(
\logit(\tau)\), which we also estimated within our framework.

| Parameter | Prior distribution | Parameter | Prior distribution |
|-----------|--------------------|-----------|--------------------|
| \(\beta_{10}\) | \(N(5, 4)\) | \(\beta_{20}\) | \(N(3.8, 0.125)\) |
| \(\beta_{11}\) | \(N(-2.8, 4)\) | \(\beta_{21}\) | \(N(-0.5, 0.125)\) |
| \(\beta_{12}\) | \(N(8, 4)\) | \(\beta_{22}\) | \(N(-0.7, 0.125)\) |
| \(\log(\sigma_1)\) | \(N(\log(1.2), 0.25)\) | \(\log(\gamma_{21}/\gamma_{22})\) | \(N(\log(0.6/a), 0.125)\) |
| \(\log(\gamma_{11}/\gamma_{12})\) | \(N(\log(0.7/a), 0.25)\) | \(\log(\gamma_{22})\) | \(N(\log(a), 0.125)\) |
| \(\log(\gamma_{12})\) | \(N(\log(a), 0.25)\) | \(\log(\nu_2)\) | \(N(\log(0.25), 0.25)\) |
| \(\log(\nu_1)\) | \(N(\log(1.5), 0.25)\) | \(\logit(\tau)\) | \(N(0.85, 0.25)\) |

*The parameter \(a \in \{0.2, 0.5, 0.8\}.*
**Results:** We first explore the accuracy of the proposed approximation to the loss function $\lambda_P(d, y)$ in Equation (15). To do this, we evaluated $\tilde{\lambda}_P(d, y)$ and $\hat{\lambda}_P(d, y)$ for 100 randomly generated designs each with 10 design points under moderate spatial dependence. The results of this are shown in Figure 2 where a clear but noisy relationship between the two loss functions can be observed. Despite the noise, there is generally agreement between the two approximations, particularly in terms of designs that minimise each loss function. Similar results were also observed under weak and strong spatial dependence. As such, we propose that our approximation can be used to select designs for this example.

Figures 3 and 4 compare the optimal designs selected under the three scenarios with 5 and 10 locations, respectively. As can be seen, when there is moderate or strong spatial covariance, the designs selected under the three loss functions have different spatial configurations. A detailed comparison of these designs with those that might be suggested in the literature for estimation and prediction is given in Appendix B. Furthermore, as shown in Figure 4, this example yielded some replicate points which may provide additional information on variance parameters. However, in real spatial settings, it may not be realistic to have such replicate points so we note that it would be straightforward to extend the adopted optimisation algorithm to avoid these.

Once the optimal designs had been found for each scenario, they were evaluated for the goals of parameter estimation and spatial prediction. For this evaluation, 100 independent estimates of the approximate loss were used. In each simulation, for a given optimal design, the expected values of the loss functions $\lambda_E(d, y)$ and $\lambda_P(d, y)$ were estimated using the MC integration as shown in Equation (4). This yielded a distribution of the expected loss values for each optimal design (see Figure 5 and Figure 6). Further, to quantitatively compare designs, a design efficiency was calculated. For this purpose, the average of the 100 expected loss values was used as follows:

$$\text{Eff}(d, d^*_\phi) = \frac{\sum_{k=1}^{100} E_{y_k}[\lambda_\phi(d, y_k)]}{\sum_{k=1}^{100} E_{y_k}[\lambda_\phi(d^*_\phi, y_k)]}, \text{ for } \phi \in \{E, P\},$$

where $d^*_E$ and $d^*_P$ are the optimal designs selected from the loss functions $\lambda_E(d, y)$ and $\lambda_P(d, y)$, respectively.

Figure 5 shows the parameter estimation performance of designs selected under each loss function. As expected, the estimation designs have the lowest value of expected loss within each scenario. Of note is the fact that the dual-purpose designs are generally highly efficient when $n = 10$, and this efficiency appears to increase with the spatial dependency. The prediction only designs (selected from $\tilde{\lambda}_P(d, y)$) did not perform well for the goal of parameter estimation, and this appears to become worse as the strength of spatial dependence increases.

Figure 6 shows the distribution of the expected loss values for different designs when the goal was prediction. As can be seen, the prediction only designs have the lowest expected loss when compared to the other designs. The dual-purpose designs are highly efficient with respect to the prediction (only) designs. Of note, this high efficiency is obtained despite the estimation designs on occasion being relatively inefficient.

Table 3 compares the efficiency of designs found under each experimental goal. These results reflect those seen in the above figures in that the dual-purpose designs are efficient under each objective, and there are occasional failures when the estimation designs perform relatively poorly for prediction. In contrast, the prediction designs maintain reasonable efficiency for estimation but not to the same extent as the dual-purpose designs.
Figure 2: The relationship between the loss values obtained from $\hat{\lambda}_P(d, y)$ and $\tilde{\lambda}_P(d, y)$ for 100 randomly generated designs under moderate spatial dependence from Example 1.

Figure 3: The optimal designs selected from each loss function ($n=5$).
Figure 4: The optimal designs selected from each loss function (n=10)

Figure 5: The boxplot of the distribution of the expected values for the loss function that focuses on parameter estimation for different designs over 100 simulations in Example 1
Figure 6: The boxplot of the distribution of the expected values for the loss function that focuses on prediction for different designs over 100 simulations in Example 1

| Number of design points | Spatial correlation | Loss function   | Estimation efficiency (%) | Prediction efficiency (%) |
|-------------------------|---------------------|-----------------|---------------------------|--------------------------|
|                         |                     | Estimation      |                          |                          |
|                         | 5                   | Weak            | 100.00                    | 39.16                    |
|                         |                     | Dual-purpose    | 78.33                     | 81.15                    |
|                         |                     | Prediction      | 70.76                     | 100.00                   |
|                         | 5                   | Moderate        | 100.00                    | 66.75                    |
|                         |                     | Estimation      | 77.92                     | 81.27                    |
|                         |                     | Dual-purpose    | 72.96                     | 100.00                   |
|                         | 10                  | Strong          | 100.00                    | 49.62                    |
|                         |                     | Estimation      | 91.62                     | 93.47                    |
|                         |                     | Dual-purpose    | 76.58                     | 100.00                   |
|                         | 10                  | Weak            | 100.00                    | 77.72                    |
|                         |                     | Estimation      | 94.27                     | 97.42                    |
|                         |                     | Dual-purpose    | 70.21                     | 100.00                   |
|                         |                     | Prediction      | 78.88                     | 100.00                   |
|                         | 10                  | Moderate        | 100.00                    | 75.61                    |
|                         |                     | Estimation      | 93.43                     | 98.01                    |
|                         |                     | Dual-purpose    | 78.88                     | 100.00                   |
|                         |                     | Prediction      | 90.33                     | 92.95                    |
|                         | 10                  | Strong          | 74.65                     | 100.00                   |

6.2 Example 2: Motivating example revisited

Here, we assess the performance of Algorithm 1 and the dual-purpose loss function in finding a retrospective design for Queensland’s air quality monitoring network based on two pollutants $NO_2$ and $PM_{2.5}$. Similar to Example 1, $Y_1 (NO_2)$ was assumed to follow a Normal distribution while
\( Y_2 \ (PM_{2.5}) \) was assumed to follow a Poisson distribution as follows:

\[
\begin{align*}
Y_1|s_1 \sim & \ N(\mu_1, \sigma^2) \quad \text{and} \quad \mu_1 = \beta_{10} + \beta_{11}X_1 + \beta_{12}X_2 + s_1, \\
Y_2|s_2 \sim & \ Pois(\exp(\mu_2)) \quad \text{and} \quad \mu_2 = \beta_{20} + \beta_{21}X_1 + \beta_{22}X_3 + s_2,
\end{align*}
\]

where \( X_1, X_2 \) and \( X_3 \) are the Y-coordinate (in UTM system), annual mean humidity level, and the annual mean wind speed measured at the given location, respectively. Again, the Matérn covariance functions with parameters \((\gamma_1, \nu_1)\) and \((\gamma_2, \nu_2)\) were used to form each variance-covariance matrix to capture the spatial variability in each response.

To find the prior distributions for the parameters, we considered the data collected from the sampled locations over the period from 2013 to 2016. First, for each parameter, vague prior distributions were assumed. Then, the observed data were used to update this prior information to obtain a posterior distribution. It is this posterior distribution that was used for design selection as detailed in Algorithm 1. Here, the purpose of this experiment is to select an optimal set of monitoring stations to collect data so that they can be used to predict the annual \( NO_2 \) and \( PM_{2.5} \) concentrations for all the monitoring stations in the network. As such, the optimal design is selected from all available locations across the network (i.e. both sampled and unsampled locations).

**Results:** Similar to the first example, we compared values of loss functions \( \hat{\lambda}_{P}(d, y) \) and \( \tilde{\lambda}_{P}(d, y) \) for 100 randomly generated designs each with 10 design points. Again, similar loss function values were obtained from each approximation, see Figure 7. As such, adopting our approximation to form efficient designs seems reasonable in this example.

![Figure 7: The relationship between the loss values obtained from \( \hat{\lambda}_{P}(d, y) \) and \( \tilde{\lambda}_{P}(d, y) \) for 100 randomly generated designs from Example 2.](image)

In Table 4, we have summarised the designs found for different values of \( n \) under each loss function. The numbers shown in this table correspond to the monitoring stations in Figure 1, with the bold numbers representing the sampled locations. Figure 8 shows the locations of the selected monitoring stations under each loss function. It can be seen that the designs selected under each loss function are different but have common stations. Further, when comparing the designs with
5 or 7 design points, it can be seen that the designs points selected from $\lambda_E(\mathbf{d}, \mathbf{y})$ were spread over the three clusters (C1, C2 and C3) while the majority of design points selected from $\lambda_P(\mathbf{d}, \mathbf{y})$ and $\lambda_D(\mathbf{d}, \mathbf{y})$ were clustered in C3. Since a large number of predicted locations belonged to C3, to minimise the prediction uncertainty, it seems reasonable to select more design points from C3 when $\lambda_P(\mathbf{d}, \mathbf{y})$ and $\lambda_D(\mathbf{d}, \mathbf{y})$ were responsible for design selection. For the designs with 10 or 15 designs points, the prediction only design points were also spread over the three clusters while the majority of dual-purpose design points were clustered around sampled locations in C2 and C3.

After locating the designs under each loss function, parameter estimation and prediction performance was assessed. Similar to Example 1, we considered 100 independent evaluations of the expected loss under each design objective. Figure 9 shows the distribution of values of the expected loss in terms of parameter estimation under different designs. Similar to Example 1, all estimation designs have lower expected loss when compared to other designs. As in the simulated example, the dual-purpose designs appear to be highly efficient in terms of parameter estimation, and this efficiency appears to increase with $n$. The results also show that the relative performance of the prediction designs decreases as $n$ increases. Figure 10 shows the distribution of expected loss values under the prediction loss function for different designs. As can be seen, the designs selected under the dual-purpose loss function perform well compared to the prediction only designs. Further, the relative performance of the estimation designs for prediction appears to decrease with $n$.

Table 5 shows the design efficiencies in terms of parameter estimation and prediction. Of note, the dual-purpose designs generally maintain high efficiency under both objectives. For designs found under the other two objectives, there are occasions each perform relatively poorly, highlighted by efficiencies of less than 70%.

Table 4: The optimal monitoring stations selected under each loss function

| $n$ | Estimation design | Dual-purpose design | Prediction design |
|-----|-------------------|---------------------|------------------|
| 5   | 15,21,27,31,33    | 8,11,14,31,33       | 2,11,21,25,31    |
| 7   | 2,12,25,27,28,29,34 | 5,8,11,21,27,33,35 | 5,11,14,21,26,29,34 |
| 10  | 5,6,8,9,11,14,15,28,29,31 | 2,5,8,11,21,24,25,26,27,33 | 5,6,8,11,21,24,27,33,34,35 |
| 15  | 1,2,3,5,6,8,9,11,14,15,28,29,31,34,35 | 2,5,8,9,11,14,15,21,27,28,29,31,33,34,35 | 2,3,6,14,15,21,24,25,27,28,29,31,33,34,35 |
Figure 8: The optimal monitoring stations selected from each loss function

Figure 9: The boxplot of the distribution of the expected values for the loss function that focuses on parameter estimation for different designs over 50 simulations in Example 2
To further investigate future designs for monitoring air quality in Queensland, we re-ran this example by only considering the monitoring stations in C3. Optimal designs with 5 and 7 design points were selected using the three loss functions with the same prior distribution as detailed at the beginning of this example. The selected optimal designs are shown in Figure 11 in the Appendix A. Similar to the above results, the dual-purpose designs remain highly efficient for both design objectives (see Table 6). Further, the relative performance of the single purpose designs appear to decrease as $n$ increases, see Figures 12 and 13.

![Figure 10: The boxplot of the distribution of the expected values for the loss function that focuses on prediction for different designs over 50 simulations in Example 2](image)

**Table 5: Efficiencies of designs selected under each loss function from Example 2**

| Number of design points | Loss function | Estimation efficiency (%) | Prediction efficiency (%) |
|------------------------|---------------|---------------------------|---------------------------|
| 5                      | Estimation    | 100.00                    | 34.69                     |
|                        | Dual-purpose  | 96.56                     | 74.91                     |
|                        | Prediction    | 85.11                     | 100.00                    |
| 7                      | Estimation    | 100.00                    | 61.57                     |
|                        | Dual-purpose  | 95.16                     | 89.47                     |
|                        | Prediction    | 75.99                     | 100.00                    |
| 10                     | Estimation    | 100.00                    | 55.58                     |
|                        | Dual-purpose  | 89.1                      | 85.44                     |
|                        | Prediction    | 76.01                     | 100.00                    |
| 15                     | Estimation    | 100.00                    | 79.61                     |
|                        | Dual-purpose  | 91.94                     | 89.76                     |
|                        | Prediction    | 82.46                     | 100.00                    |
7 Discussion

In this paper, we have developed a Bayesian design approach to minimise uncertainty in spatial predictions. The approach is based on the derivation of a dual-purpose loss function which quantifies the uncertainty in a given spatial process. Given the substantial computational time needed to evaluate this loss function, a number of approximations were proposed, and these were shown to work well for locating an optimal design. In addition, we extended our methodology to handle bivariate responses in spatial settings, and this was motivated by what has been observed in real-world studies.

In the first example, three design scenarios were considered to test our proposed algorithm and the dual-purpose loss function under various levels of spatial dependence. Overall, the results showed that the dual-purpose designs remained highly efficient under each objective, and this was despite some single objective designs being inefficient under the other objective. Similar results were observed when designing Queensland’s air quality monitoring network based on two air quality measurements $NO_2$ and $PM_{2.5}$. The clustered nature of the stations (in terms of spatial locations) was accounted for via the range parameter in our covariance function which resulted in the three clusters being independent. Alternative approaches and/or covariance functions to account for this clustering could be considered in the future, and this may have implications in terms of the design.

Despite the theoretical underpinnings of the loss functions developed in this study, the computational complexity of the loss function $\lambda_P(d, y)$ is a drawback of this design approach. Indeed, this difficulty led to the development of an efficient approximation based on the joint distribution of the parameters and the data. We evaluated the proposed approximation to the loss functions, in terms of the accuracy and the computational time required to obtain the approximate loss values. A comparison of the computational times required to evaluate each approximation is shown in Figures 14 and 15. This was based on evaluating 500 independent utilities on a computer with an Intel core-i7 processor running at 2.2 Gigahertz. For the models considered in this paper, this was shown to yield a reasonable approximation. However, such an approach may not be appropriate in general. For example, if bivariate continuous and binary data were observed, it is unlikely that a Multivariate Normal approximation will be appropriate for estimating the entropy of the distribution of the data. As such, alternative methods are needed, and this is an area we plan to explore into the future.

Another possible extension to this work could include quantifying the uncertainty in different components of the spatial process. This can be achieved by quantifying the uncertainty in the mean model and the covariance function, potentially following the work of Borth (1975); McGree (2017). The proposed loss function could then consider a set of all plausible mean and covariance functions, which is flexible enough to represent the uncertainty in the spatial process (Pilz et al., 1997). In studies where high-dimensional multivariate data are observed, it may be necessary to adopt particular Vine-Copulas (Brechmann et al., 2013; Panagiotelis et al., 2017). Other areas we hope to pursue into the future include extensions to quantify uncertainty in temporal (and thus, spatio-temporal) processes (Liu and Vanhatalo, 2019).

Acknowledgement

SGJS was supported by QUTPRA scholarship from the Queensland University of Technology. WGM was partially supported by Austrian Science Fund (FWF): I 3903-N32. JMM was supported by an Australian Research Council Discovery Project (DP200101263). Computational resources
and services used in this work were provided by the HPC and Research Support Group, Queensland University of Technology, Brisbane, Australia.

References

Albert, P. S. and McShane, L. M. (1995), ‘A generalized estimating equations approach for spatially correlated binary data: Applications to the analysis of neuroimaging data’, Biometrics 51(2), 627–638.

Bárdossy, A. (2006), ‘Copula-based geostatistical models for groundwater quality parameters’, Water Resources Research 42(11).

Bloom, L. M. and Kentwell, D. J. (1999), A geostatistical analysis of cropped and uncropped soil from the jimperding brook catchment of western australia, in J. Gómez-Hernández, A. Soares and R. Froidevaux, eds, ‘geoENV II — Geostatistics for Environmental Applications’, Springer Netherlands, Dordrecht, pp. 369–379.

Bohorquez, M., Giraldo, R. and Mateu, J. (2017), ‘Multivariate functional random fields: prediction and optimal sampling’, Stochastic Environmental Research and Risk Assessment 31(1), 53–70.

Borth, D. M. (1975), ‘A total entropy criterion for the dual problem of model discrimination and parameter estimation’, Journal of the Royal Statistical Society. Series B (Methodological) pp. 77–87.

Brechmann, E. C., Schepsmeier, U. et al. (2013), ‘Modeling dependence with C-and D-vine Copulas: The R-package CDVine’, Journal of Statistical Software 52(3), 1–27.

Bruno, R., Palumbo, V. and Bonduà, S. (2001), Identification of regional variability component by geostatistical analysis of stream sediments, in P. Monestiez, D. Allard and R. Froidevaux, eds, ‘geoENV III — Geostatistics for Environmental Applications’, Springer Netherlands, Dordrecht, pp. 113–123.

Castrignanò, A., Buttafuco, G. and Giasi, C. (2008), Assessment of groundwater salinisation risk using multivariate geostatistics, in A. Soares, M. J. Pereira and R. Dimitrakopoulos, eds, ‘geoENV VI – Geostatistics for Environmental Applications: Proceedings of the Sixth European Conference on Geostatistics for Environmental Applications’, Springer Netherlands, Dordrecht, pp. 191–202.

Clayton, D. G. (1978), ‘A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence’, Biometrika 65(1), 141–151.

Cook, R. D. and Johnson, M. E. (1981), ‘A family of distributions for modelling non-elliptically symmetric multivariate data’, Journal of the Royal Statistical Society. Series B (Methodological) 43(2), 210–218.

Cook, R. D. and Wong, W. K. (1994), ‘On the equivalence of constrained and compound optimal designs’, Journal of the American Statistical Association 89(426), 687–692.

Diggle, P. and Lophaven, S. (2006), ‘Bayesian geostatistical design’, Scandinavian Journal of Statistics 33(1), 53–64.

Diggle, P. and Ribeiro, P. J. (2007), Model-based Geostatistics, Springer Series in Statistics, Springer, New York.

Diggle, P., Ribeiro, P. J. and Christensen, O. F. (2003), An introduction to model-based geostatistics, in J. Møller, ed., ‘Spatial Statistics and Computational Methods’, Springer New York, pp. 43–86.

Durante, F. and Sempi, C. (2010), Copula theory: an introduction, in ‘Copula theory and its applications’, Springer, pp. 3–31.

Falk, M. G., McGree, J. M. and Pettitt, A. N. (2014), ‘Sampling designs on stream networks using the pseudo-Bayesian approach’, Environmental and Ecological Statistics 21(4), 751–773.

Fuentes, M., Chaudhuri, A. and Holland, D. M. (2007), ‘Bayesian entropy for spatial sampling design of environmental data’, Environmental and Ecological Statistics 14(3), 323–340.

Genest, C. and MacKay, J. (1986), ‘The joy of Copulas: bivariate distributions with uniform marginals’, The American Statistician 40(4), 280–283.

Gómez-Rubio, V., Bivand, R. S. and Rue, H. (2019), Spatial models using laplace approximation methods, in M. M. Fischer and P. Nijkamp, eds, ‘Handbook of Regional Science’, Springer, Berlin, Heidelberg, pp. 1–16.
Gräler, B. and Pebesma, E. (2011), ‘The pair-Copula construction for spatial data: a new approach to model spatial dependency’, *Procedia Environmental Sciences* **7**, 206 – 211.

Hill, W. J., Hunter, W. G. and Wichern, D. W. (1968), ‘A joint design criterion for the dual problem of model discrimination and parameter estimation’, *Technometrics* **10**(1), 145–160.

Huang, Y., Hanneke, R. and Jones, R. (2019), ‘Bibliometric analysis of cardiometabolic disorders studies involving \( NO\_2 \), \( PM\_2.5 \) and noise exposure’, *BMC Public Health* **19**(1), 1–14.

Jiang, X. Q., Mei, X. D. and Feng, D. (2016), ‘Air pollution and chronic airway diseases: what should people know and do?’, *Journal of thoracic disease* **8**(1), 31–40.

Joe, H. (2014), *Dependence Modeling with Copulas*, 1 edn, New York: Chapman and Hall/CRC.

Kazianka, H. and Pilz, J. (2011), ‘Bayesian spatial modeling and interpolation using Copulas’, *Computers & Geosciences* **37**(3), 310 – 319.

Kullback, S. and Leibler, R. A. (1951), ‘On information and sufficiency’, *The Annals of Mathematical Statistics* **22**(1), 79–86.

Le, N. D., Sun, L. and Zidek, J. V. (2003), Designing networks for monitoring multivariate environmental fields using data with monotone pattern, Technical report, TR 2003-5, Statistical and Applied Mathematical Sciences Institute, RTP, NC.

Le, N. D. and Zidek, J. V. (2006), Statistical analysis of environmental space-time processes, in ‘Springer Series in Statistics’, Springer, New York, NY.

Liu, J. and Vanhatalo, J. (2019), ‘Bayesian model based spatiotemporal survey designs and partially observed log Gaussian cox process’, *Spatial Statistics* **35**, 1–27.

Long, Q., Scavino, M., Tempone, R. and Wang, S. (2013), ‘Fast estimation of expected information gains for Bayesian experimental designs based on Laplace approximations’, *Computer Methods in Applied Mechanics and Engineering* **259**, 24 – 39.

McGree, J. M. (2017), ‘Developments of the total entropy utility function for the dual purpose of model discrimination and parameter estimation in Bayesian design’, *Computational Statistics and Data Analysis* **113**, 207–225.

McGree, J. M., Eccleston, J. A. and Duffull, S. B. (2008), ‘Compound optimal design criteria for nonlinear models’, *Journal of Biopharmaceutical Statistics* **18**(4), 646–661.

Meyer, R. K. and Nachtsheim, C. J. (1995), ‘The coordinate-exchange algorithm for constructing exact optimal experimental designs’, *Technometrics* **37**(1), 60–69.

Müller, P., Sansó, B. and Iorio, M. D. (2004), ‘Optimal bayesian design by inhomogeneous Markov chain simulation’, *Journal of the American Statistical Association* **99**(467), 788–798.

Müller, W. (2007), *Collecting Spatial Data, 3rd revised and extended edition*, Springer Verlag, Heidelberg.

Musafir, G. N. and Thompson, M. H. (2017), ‘Non-linear optimal multivariate spatial design using spatial vine Copulas’, *Stochastic Environmental Research and Risk Assessment* **31**(2), 551–570.

Nelsen, R. (2006), ‘An introduction to Copulas, 2nd’, *New York: SpringerScience Business Media*.

Overstall, A. M., McGree, J. M. and Drovandi, C. C. (2018), ‘An approach for finding fully Bayesian optimal designs using normal-based approximations to loss functions’, *Statistics and Computing* **28**(2), 343–358.

Overstall, A. M. and Woods, D. C. (2017), ‘Bayesian design of experiments using approximate coordinate exchange’, *Technometrics* **59**(4), 458–470.

Overstall, A. M., Woods, D. C. and Adamou, M. (2018), ‘acebayes: An R package for Bayesian optimal design of experiments via approximate coordinate exchange’, *arXiv:1705.08096*.

Panagiotelis, A., Czado, C., Joe, H. and Stöber, J. (2017), ‘Model selection for discrete regular vine Copulas’, *Computational Statistics & Data Analysis* **106**, 138–152.

Pilz, J., Schimek, M. and Spöck, G. (1997), Taking account of uncertainty in spatial covariance estimation, in ‘Geostatistics Wollongong 96’, Vol. 8, Kluwer Academic, pp. 302–313.

Roberts, S., Arsenaeult, L., Barratt, B., Beevers, S., Danese, A., Odgers, C. L., Moffitt, T. E., Reuben, A., Kelly, F. J. and Fisher, H. L. (2019), ‘Exploration of \( NO\_2 \) and \( PM\_2.5 \) air pollution and mental health problems using high-resolution data in london-based children from a uk longitudinal cohort...
study’, *Psychiatry Research* **272**, 8–17.
Tao, Y., Liu, J., Li, Z., Lin, J., Lu, T. and Yan, F. (2013), ‘Dose-finding based on bivariate efficacy-toxicity outcome using archimedean Copula’, *PLoS ONE* **8**(11), 1–6.
Watson, K. E. (2006), ‘Air pollution and heart disease’, *Reviews in cardiovascular medicine* **7**(1), 44–44.
Zhang, S., Li, G., Tian, L., Guo, Q. and Pan, X. (2016), ‘Short-term exposure to air pollution and morbidity of copd and asthma in east asian area: A systematic review and meta-analysis’, *Environmental research* **148**, 15–23.

**Appendix A**

![Figure 11](image)

**Figure 11:** The optimal monitoring stations selected from C3 in Example 2 using each loss function

![Figure 12](image)

**Figure 12:** The boxplot of the distribution of the expected values for the loss function that focuses on parameter estimation for different designs over 50 simulations from C3 in Example 2
Figure 13: The boxplot of the distribution of the expected values for the loss function that focuses on prediction for different designs over 50 simulations from C3 in Example 2

Table 6: Efficiencies of designs selected under each loss function from C3 in Example 2

| Number of design points | Loss function | Estimation efficiency (%) | Prediction efficiency (%) |
|-------------------------|---------------|---------------------------|--------------------------|
| 5                       | Estimation    | 100.00                    | 80.62                    |
|                         | Dual-purpose  | 93.56                     | 82.88                    |
|                         | Prediction    | 71.73                     | 100.00                   |
| 7                       | Estimation    | 100.00                    | 82.03                    |
|                         | Dual-purpose  | 83.13                     | 87.34                    |
|                         | Prediction    | 72.13                     | 100.00                   |

Figure 14: The comparison of computational time required to estimate the two loss functions for Example 1
Appendix B

In spatial settings, the literature suggests that parameter estimation designs often feature clustered sampling points (Diggle and Ribeiro, 2007). Here the ratio of the range and nugget to the partial sill influence the configuration of clusters. In contrast, the prediction designs often resemble a regularly spaced design. It is presumed that designs for both of these goals will combine features from both.

In Example 1, we potentially observe some of the above features across the three different designs with some apparent difference. To investigate about these differences, a new simulation study has been conducted where the estimation and prediction performance of a selected set of designs were evaluated. These designs were: an equally spaced triangular design, a design with all the points on the boundary, and a design with all points close to the prediction locations, see Figure 16 below which shows these three designs. The performance of these designs for estimation and prediction were compared to the designs found in this work.

Figure 16: A selected set of theoretical designs
To evaluate the prediction performance of the designs shown in Figure 16, we evaluated the expected prior and posterior prediction variance at each of the prediction locations. This was achieved by simulating 500 data sets from each design and approximating the posterior distribution for each. For each of these posterior distributions, posterior predictive data were simulated at each prediction location, and then the variance of these at each location was evaluated. The average of this variance was then evaluated across the 500 simulations. The results for this for both responses are shown in Figures 17 and 18.

From Figure 17, it is clear that the prediction design found in our paper yields the smallest average prediction variance across all locations. The next best design appears to be the ‘close.pred’ design suggesting that a design that is more regularly spaced across the entire region performs well here for prediction, and this is what would be expected based on results from the literature. Otherwise, there appears to be very little difference in the performance of the remaining designs. Given one of these designs in the estimation design found in our work, then potentially these designs are focussing more on estimation.

From Figure 18, similar conclusions can be drawn.

Figure 17: Posterior prediction variances vs prior prediction variances of response 1 (Y1) for prediction locations
Figure 18: Posterior prediction variances vs prior prediction variances of response 2 (Y2) for prediction locations

To evaluate the estimation performance of the designs, a similar approach to evaluate the prediction performance was followed. That is, 500 prior predictive data sets were generated based on each design, and a posterior distribution was found for each. The average posterior variance of each parameter was then evaluated based on each design. These averages are shown in Figure 19. Across all designs, there does not appear to be an appreciable difference in terms of estimation performance. However, our estimation design is performing better but only marginally. This may be because the prior variance for most parameters is quite low, so attention may be given to estimating $\beta_{11}$, $\beta_{12}$ and $\beta_{10}$ more precisely.

To do so, at least for the slope terms, a contrast between data collected at extreme values of X1 and also X2 would seem useful, and this is a feature of all designs considered.

The above highlights the potentially influence the prior distribution can have in Bayesian design, and provide a potential reason for differences observed between our designs and those that might be considered as being more standard for estimation and prediction in spatial settings. Given this, such standard designs may not necessarily be optimal under a variety of prior configurations. This was observed here but we also note that these standard designs did not perform poorly.
Figure 19: The prior variance vs the posterior variance for each parameter for different designs