Received: 3 February 2020 | Revised: 29 March 2021 | Accepted: 30 June 2021

DOI: 10.1111/sjos.12554

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

Efficient spatial designs using Hausdorff distances and Bayesian optimization

Jacopo Paglia[1] | **Jo Eidsvik**[1] | **Juha Karvanen**[2]

[1]Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway

[2]Department of Mathematics and Statistics, University of Jyvaskyla, Jyväskylä, Finland

Correspondence
Jacopo Paglia, Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim 7034, Norway.
Email: jacopo.paglia91@gmail.com

Funding information
Jacopo Paglia's and Jo Eidsvik's work are supported by the KPN project 255418/E30: "Reduced uncertainty in overpressures and drilling window prediction ahead of the bit (PressureAhead)", of the Norwegian Research Council and the DrillWell Centre (AkerBP, Wintershall, ConocoPhillips and Equinor). Juha Karvanen's work is supported by Grant number 311877 "Decision analytics utilizing causal models and multiobjective optimisation" (DEMO), of the Academy of Finland.

Abstract

An iterative Bayesian optimization technique is presented to find spatial designs of data that carry much information. We use the decision theoretic notion of value of information as the design criterion. Gaussian process surrogate models enable fast calculations of expected improvement for a large number of designs, while the full-scale value of information evaluations are only done for the most promising designs. The Hausdorff distance is used to model the similarity between designs in the surrogate Gaussian process covariance representation, and this allows the suggested algorithm to learn across different designs. We study properties of the Bayesian optimization design algorithm in a synthetic example and real-world examples from forest conservation and petroleum drilling operations. In the synthetic example we consider a model where the exact solution is available and we run the algorithm under different versions of this example and compare it with existing approaches such as sequential selection and an exchange algorithm.

KEYWORDS

Bayesian optimization, decision-making, Hausdorff distance, value of information

Abbreviations: EI, expected improvement; GP surrogate, Gaussian process; PoV posterior value; PV prior value; VOI, value of information

1 | INTRODUCTION

This paper is inspired by challenging decision situations in the earth and environmental sciences. In these situations, data are gathered to support decisions about resource management. Data acquisition and processing is often costly, and it is then important to choose the sampling design wisely. There exist several common design or information criteria, see for example, Ryan et al. (2016) for a recent review. For decision-makers, value of information (VOI) analysis is useful in this context (Abbas & Howard, 2015; Eidsvik et al., 2015), as it is directly connected with the information gain associated with the decision situation and it provides a bound on the expected monetary amount one should be willing to pay for data to aid in resolving this decision situation.

We focus on designing experiments in spatial domains. Here, Kriging interpolation (Stein, 2012) is often used to propagate the effects of observations based on spatial correlations. When choosing the criterion for the selection of the sampling design it is important to keep in mind the scope of the experiment. One could, for example, be interested in choosing a design that spread well across the domain, also called spatially balanced designs (Grafström et al., 2012; Stevens Jr & Olsen, 2004). In the study presented here, however, the focus is on finding a design that maximizes the VOI, and where the spatial balance may arise as a consequence of the spatial modelling of the variables of interest. Using VOI analysis, we aim to provide the decision-maker with efficient survey designs including the optimal number of measurement locations and their spatial configuration. We assume that the spatial domain is discretized to a grid so that there is a finite set of possible observation locations. Moreover, we limit scope to static designs (Diggle & Lophaven, 2006; Dobbie et al., 2008; Huan & Marzouk, 2013), where the experimental configuration is selected once, at the onset of data gathering. The alternative is sequential data gathering, where the design can be adapted based on the observations made in the first (batches of) measurements (Binois et al., 2019; Drovandi et al., 2013; Eidsvik et al., 2018), but this is not always possible in practical experimental planning, which must comply with project management and budgetary limitations.

As pointed out by several others, this design problem is not trivial as the number of possible designs grows combinatorially fast. Royle (2002) proposed a random exchange algorithm to search for the optimal design. García-Ródenas et al. (2020) presented an interesting overview of some of the main algorithms for finding efficient designs. Weaver et al. (2016) and Overstall and Woods (2017) applied Bayesian optimization to focus the search for good designs. We use a Gaussian process (GP) surrogate model enabling fast computation of the expected improvement (EI) in Bayesian optimization. This is combined with techniques from search algorithms, to find efficient spatial designs. As was also regarded as a possibility in Ginsbourger et al. (2016) in the context of computer experiments, the current paper presents an approach for using the Hausdorff distance between various designs. The contribution of our work is using this to correlate outcomes of similar site configurations, within a realistic statistical model, and combining this in an algorithm for quickly locating valuable sampling designs. Even though our focus is on spatial decision situations and design, we believe that this approach can also be applicable to other big-data challenges (Drovandi et al., 2017) and active learning approaches (Bouneffouf, 2016; Settles, 2012), where the challenge is more related to which data to process for learning and improved classifications.

In Section 2 we describe the spatial design problem in mathematical detail and define the VOI criterion which we use as a practically relevant information measure. In Section 3 we outline the Bayesian optimization approach using Hausdorff distances to borrow information among

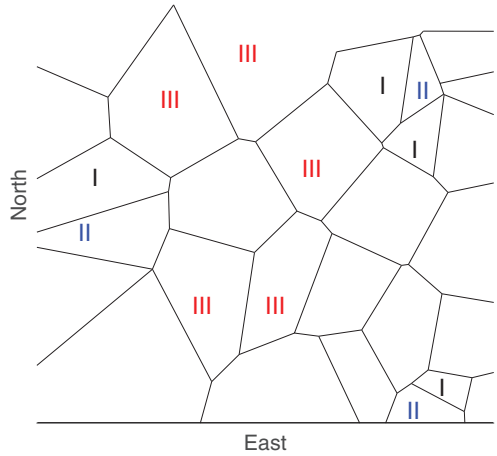

FIGURE 1 Illustration of a spatial domain split in 40 regional units of varying size. Three different designs are indicated (Design I, II, and III) of different cardinality and spatial allocation [Colour figure can be viewed at wileyonlinelibrary.com]

similar designs. In Section 4 we study the properties of the methodology via simulations. In Section 5 we show results on forestry and petroleum examples to demonstrate possible applications of the methods. Section 6 has closing remarks on the methodological contributions presented here, including viable opportunities future work.

2 | SPATIAL DESIGN OF EXPERIMENTS

2.1 | Spatial survey designs

We consider a situation as illustrated in Figure 1, with a spatial phenomenon allocated to a two-dimensional domain divided in grid cells or sites. The distances between the sites are defined as the Euclidean distances between the centers of the cells. The approach presented in the paper can be extended to higher dimensions with minor changes.

The spatial variables of interest are represented at n sites, denoted $\boldsymbol{s}_1, \dots, \boldsymbol{s}_n$ with $\boldsymbol{s}_i = (\text{north}_i, \text{east}_i)$, $i = 1, \dots, n$. In our applications, these sites have a particular interest to the decision-maker. For instance, in the forestry example, the governmental institute must choose at each of the n sites whether this forest unit should be harvested or left for conservation. Because there is much at stake and uncertain outcomes, the decision-maker is likely to benefit from doing surveys at (a subset of) the sites.

Data can be gathered at any of the n sites in our description, and a design defines a subset of these n sites where the data collection will be conducted (other cases can be constructed similarly, see e.g. Section 5.2). The possible spatial designs then include no sites, single sites, couples, triplets, and so on, up to all n sites in the design. We denote these by $\mathscr{D} = \bigcup_{i=0}^{n} \mathscr{D}_i$, defined by;

$$\mathscr{D}_0 = \emptyset, \qquad \text{no sites in design,}$$
$$\mathscr{D}_1 = \{(\boldsymbol{s}_1), (\boldsymbol{s}_2), \dots, (\boldsymbol{s}_n)\}, \qquad \text{one site in design,}$$

$$\begin{aligned}
\mathscr{D}_2 &= \{(\boldsymbol{s}_1, \boldsymbol{s}_2), (\boldsymbol{s}_1, \boldsymbol{s}_3), \ \dots\ , (\boldsymbol{s}_{n-1}, \boldsymbol{s}_n)\}\,, && \text{two sites in design,}\\
&\ \ \vdots && \ \ \vdots\\
\mathscr{D}_n &= \{(\boldsymbol{s}_1, \ \dots\ , \boldsymbol{s}_n)\}\,, && \text{all sites in design.}
\end{aligned}$$

There are n possible designs of cardinality one, $\binom{n}{2}$ possible designs of cardinality two, etc. This means that there are 2^n possible designs in $\mathscr{D}$. We will further denote a general design by $D \in \mathscr{D}$ and its cardinality by $|D|$. The sites in this design are then $\boldsymbol{s}_{D,1}, \ \dots\ , \boldsymbol{s}_{D,|D|}$. The number of sites shared by designs C and D is $|C \cap D|$, while the number of sites in at least one of the designs is $|C \cup D|$.

In our setting we compare the information gain obtained by different designs, and it makes sense that similar spatial designs contain almost the same information. In Figure 1 three different designs are indicated (I, II, and III). Designs I and II appear very similar in the spatial allocation of survey sites even though they have different cardinalities (three and four). Most likely, Design I will not have much to offer over Design II, unless there is much noise in the data or large gain in capturing additional covariate information which could be important for predictive purposes. Say, in the forestry example, a biologist would spend time doing one more experiments in Design I, at an extra cost. But unless she learns substantially more about the model, there is not much additional spatial information in Design I compared with doing just the three measurements in Design II. The last survey plan, Design III, is spatially very different from the others because it allocates the measurements in the central parts of the domain. The value of this design could be very different from that of Design I and II.

To find the optimal design one must evaluate the information gain and cost for all possible design sets, but in practice one can only evaluate it for a fraction of all possible designs. We suggest a statistical approach for this optimization problem, where we utilize the similarity of spatial designs to estimate the information gain.

2.2 | Value of information

The goal of spatial design of experiments is to choose a valuable survey plan for information gathering. This choice must balance expected information gain with the cost of data acquisition and processing. To evaluate the expected information gain associated with designs, one must formulate a value or utility function. In the applications that we consider here, it is relatively straightforward to relate the question about information gain to an underlying decision situation, meaning that data are only valuable when their outcome can materialize in different decisions. For instance, in the forestry example the underlying decision is to conserve forest units or not, and data can help the decision-maker to decide one or the other, depending on what the information reveals. Managers are further often willing to phrase these decision situation in terms of monetary units, and then the VOI which gives the expected gain in information is directly comparable to the cost of data gathering. If the VOI exceeds this cost, the experiment is worthwhile and the decision-maker should commit to gather the information, if the budget permits the cost. We next define the VOI formally through a model for the random variables of interest, the decision alternatives and the information gathered by a chosen design.

The variables of interest are denoted by $\boldsymbol{x} = (x_1, \ \dots\ , x_n)$, where $x_i = x(\boldsymbol{s}_i)$, $i = 1, \ \dots\ , n$. In our context these are directly tied to a decision situation and connected to economic values. For

instance, they can be random profits allocated to forest units or loss associated with a drilling operation. Note that other parameters will be important in the statistical modeling of the phenomenon of interest, such as regression parameters and covariance function parameters, but in our setting they are only used in the construction of a realistic statistical model for the phenomenon that is studied, and in particular for the variables of interest $\boldsymbol{x}$. Assuming a continuous sample space for the variables of interest, we denote its probability density function by $p(\boldsymbol{x})$, with marginal density $p(x_i)$ for each sites $\boldsymbol{s}_i$.

The decision alternatives are generally denoted by $\boldsymbol{a} \in \mathcal{A}$, where $\mathcal{A}$ is the set of all possible alternatives. In some situations, the alternatives decouple (Eidsvik et al., 2015), involving for instance local decisions about harvesting units in our forest conservation example. In general, the prior value (PV), without any additional information, is defined as the value from doing the optimal decisions. Assuming a risk-neutral decision maker (Abbas & Howard, 2015), the PV is calculated from expected values as follows;

$$\text{PV} = \max_{\boldsymbol{a}\in\mathcal{A}} \{\mathbb{E}(v(\boldsymbol{x}, \boldsymbol{a}))\}, \quad \mathbb{E}(v(\boldsymbol{x}, \boldsymbol{a})) = \int v(\boldsymbol{x}, \boldsymbol{a}) p(\boldsymbol{x}) d\boldsymbol{x}. \tag{1}$$

Here, $v(\boldsymbol{x}, \boldsymbol{a})$ represents the value function, which could be quite general, but in our application it is the monetary profits associated with choice $\boldsymbol{a} \in \mathcal{A}$ when the variable outcome is $\boldsymbol{x}$. In the forestry example, the decision-maker will choose to conserve the sites that have high preservation value, while the others are harvested.

It is difficult to make decisions under uncertainty, and one can choose to purchase information that facilitate decision-making. We here let $\boldsymbol{y}_D$ denote the data gathered by design $D \in \mathscr{D}$. This data is relevant to the decision situation in the sense that it will provide information about the variable of interest $\boldsymbol{x}$. In the applications below, the model for data is given as a conditional probability density or mass function $p(\boldsymbol{y}_D|\boldsymbol{x})$, and the marginal model for data is then $p(\boldsymbol{y}_D) = \int p(\boldsymbol{y}_D|\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}$.

When the data are available, the conditional value (CV) is

$$\text{CV}(\boldsymbol{y}_D) = \max_{\boldsymbol{a}\in\mathcal{A}} \{\mathbb{E}(v(\boldsymbol{x}, \boldsymbol{a})|\boldsymbol{y}_D)\}, \tag{2}$$

and the expected posterior value (PoV) before the data gathering is obtained by taking the expectation of expression (2) over the possible data outcomes:

$$\text{PoV}(D) = \mathbb{E}_{\boldsymbol{y}_D}\left[CV(\boldsymbol{y}_D)\right] = \mathbb{E}_{\boldsymbol{y}_D}\left[\max_{\boldsymbol{a}\in\mathcal{A}} \{\mathbb{E}(v(\boldsymbol{x}, \boldsymbol{a})|\boldsymbol{y}_D)\}\right]. \tag{3}$$

The VOI is defined as the difference between the expected PoV in (3) and the PV in (1):

$$\text{VOI}(D) = \text{PoV}(D) - \text{PV}. \tag{4}$$

The goal is to choose a valuable design D. Keeping in mind that data comes with a cost, we should compare the VOI with the cost $C(D)$ of design D. This means that the objective is to optimize

$$D^* = \text{argmax}_{D\in\mathscr{D}} I(D), \quad I(D) = \text{VOI}(D) - C(D). \tag{5}$$

Other objectives are of course possible. For instance, a decision-maker might have a fixed budget for the design, and the goal would then be to maximize the VOI among all designs that have a cost less than the budget.

With large opportunities for data gathering, it is extremely difficult to find the optimal design. First, the complexity grows extremely fast with the number of sites. Second, in common settings, the calculation of the information design criterion in (5) for a fixed design typically requires quite a bit of computational effort as is emphasized by the complexity of the integral maximum expectation expressions required in (3). In practice one must often turn to heuristic approaches to such design problems (García-Ródenas et al., 2020. We suggest to use a statistical approximation strategy that evaluates $I(D)$ in (5) only for a few promising designs which are extracted by a fast Bayesian optimization approach building on GP surrogate models and EI.

3 | BAYESIAN OPTIMIZATION FOR DESIGNS

We develop a Bayesian optimization approach to guide the search for the maximum of $I(D)$ in (5). We combine computational search algorithms with the EI acquisition criterion to select which designs to evaluate in an iterative optimization workflow. In doing so, we suggest to model the information measure $I(D)$ using a GP surrogate model. This is in line with the common approaches for Bayesian optimization (Brochu et al., 2010; Frazier, 2018). The benefits of using a GP surrogate for the information measure is that it enables:

- efficient model updating based on evaluations (Section 3.1),
- learning across different but similar designs (Section 3.2),
- computing EI in closed form, to focus on evaluating promising designs (Section 3.3),
- framing a useful algorithmic description of the overall procedure (Section 3.4).

3.1 | GP surrogate

The information gain $I(D)$ is represented by a GP surrogate model. This relies on mean and variance–covariance specifications of the information gain for input designs. In the current setting with Bayesian optimization, the GP surrogate model is updated sequentially when more evaluations become available.

When m designs $D_{(1)}, \dots, D_{(m)}$ have been evaluated, the knowledge is denoted $\mathcal{F} = \left\{(I_{(j)}, D_{(j)}); j = 1, \dots, m\right\}$. By standard multivariate Gaussian theory, the conditional distribution for the information measure at design D is then Gaussian with mean and variance

$$\begin{aligned} \mu(D;\mathcal{F}) &= \mu + \boldsymbol{k}_{D,\mathcal{F}}^t \boldsymbol{K}_{\mathcal{F}}^{-1}(\boldsymbol{I}(\mathcal{F}) - \mu\mathbf{1}), \\ \sigma^2(D;\mathcal{F}) &= \sigma^2\left(1 - \boldsymbol{k}_{D,\mathcal{F}}^t \boldsymbol{K}_{\mathcal{F}}^{-1}\boldsymbol{k}_{D,\mathcal{F}}\right). \end{aligned} \tag{6}$$

Here, $\boldsymbol{I}(\mathcal{F}) = (I_{(1)}, \dots, I_{(m)})^t$ is the length m vector of information gain evaluations, $\boldsymbol{K}_{\mathcal{F}}$ the $m \times m$ correlation matrix between evaluations of designs, $\boldsymbol{k}_{D,\mathcal{F}}$ the length m vector of correlations between the evaluations and the information gain for design D, and $\mathbf{1}$ is a length m vector of 1 entries. The representation requires specification of the mean μ and variance σ^2, which are

assumed constant for all designs. It further needs a valid correlation function specification $K(C, D)$ between two different designs C and D (see Section 3.2).

3.2 | Distance between designs

The correlation function gauges the similarity between designs, as defined via $\boldsymbol{k}_{D,\mathcal{F}}$ and $\boldsymbol{K}_{\mathcal{F}}$ in (6). This specification of a correlation function is a common task in spatial statistics and Bayesian optimization over a regular input space. In our setting with spatial designs, it is not obvious how to assign this correlation function, and a main contribution of this paper is to formulate a distance measure between designs which is useful in the context of Bayesian optimization. Our proposed distance measure for this task is the Hausdorff distance which is presented next, but we also outline other distance measures below to discuss this topic in a more general context. Throughout this description, we consider two general designs $D = (\boldsymbol{s}_{D,1}, \ldots, \boldsymbol{s}_{D,|D|})$ and $C = (\boldsymbol{s}_{C,1}, \ldots, \boldsymbol{s}_{C,|C|})$. For two sites $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$, we let $\|\boldsymbol{s}_i - \boldsymbol{s}_j\|$ be the Euclidean distance between the two sites.

The Hausdorff distance is commonly used to measure the distance between curves, images, or point sets (Huttenlocher et al., 1992). In our context it represents the maximum of the minimal distances from sites in one set to sites in the other set, and it hence measures similarity of designs:

$$h = \text{dist}_H(D, C) = \max\{h_H(D, C), h_H(C, D)\}, \tag{7}$$

$$h_H(D, C) = \max_{i=1:|D|} \left\{ \min_{j=1:|C|} \|\boldsymbol{s}_{D,i} - \boldsymbol{s}_{C,j}\| \right\}. \tag{8}$$

Figure 2 illustrates several designs of size 1, … 4. For each subplot the maximum distances from sites in one set to the other is calculated and shown. One design D is marked as circle, the other design C is marked as cross. The Hausdorff distance in (7) is printed in the displays, and $h_H(D, C)$ and $h_H(C, D)$ are indicated.

We note that in some cases the maximum distances from one set to the other are identical (upper right display and bottom middle display), but for most of these site configurations this symmetry is not present. For instance, in the upper left display, the circle is relatively close to the southernmost site in the cross set, but the northernmost site in the cross set is quite far from the circle. Similarly, in the centre display, both sites in the circles set are close to a site in the cross set, but one site in the cross set is far from the closest site in the circle set. In the bottom-middle display the designs are rather similar, and the distance is small ($h = 0.113$). In all the right displays, the designs are very different, and the distances are large. Based on Hausdorff distances for sets like that displayed in Figure 2, it seems to be a useful way to measure the difference between designs.

We next present some alternative distances (Fujita, 2013; Min et al., 2007) that could be useful for our purpose, and discuss their pros and cons. One alternative distance is defined by the minimum of the distances between design sites of the two sets. The main problem with this distance is that designs with sites in common are not separated because the distance in this case will be zero. In addition, this distance is not a proper metric because the triangular inequality does not hold. We will hence discard this distance—it is not suitable for our purpose. Similarly it is possible to define another distance considering the maximum of the distances between designs sites instead. Again, this is not a proper distance metric, and it is not convenient for our measure of similarity because the distance between two equal sets is greater than 0. Yet another candidate is the Jaccard distance (Levandowsky & Winter, 1971) which is defined via the relative counts of

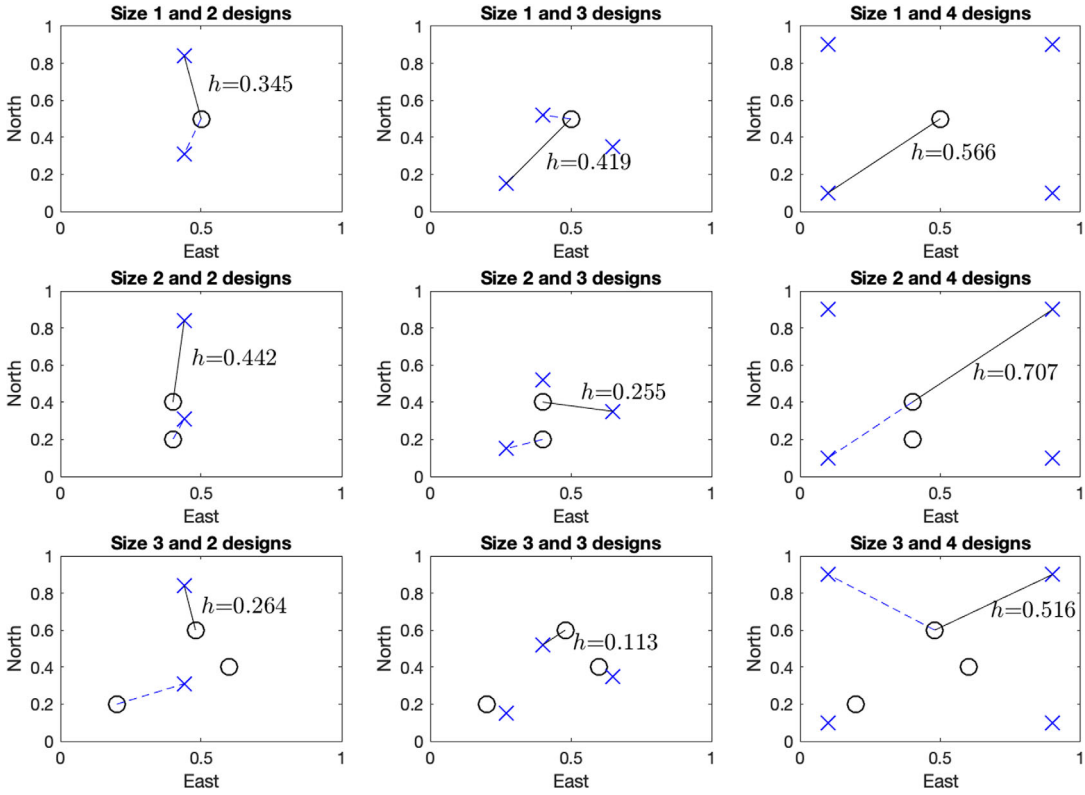

FIGURE 2 Hausdorff distance between various designs, h is the Hausdorff distance between the two sets, marked as circle for D and cross for C. The solid lines represent the maximum of the minimal distances between C and D, while the dashed line represents the maximum of the minimal distances between D and C [Colour figure can be viewed at wileyonlinelibrary.com]

sites that are not shared in the designs. We believe this could be a very sensible way to consider dissimilarity between designs for features or class covariates, but not for our type of applications because it does not account for the spatial distance between sites in the sets.

Fujita (2013) describes another metric based on the average distances between the two designs:

$$\text{dist}_F(D, C) = \frac{1}{|D \cup C||D|} \sum_{\boldsymbol{s}_{D,i} \in D} \sum_{\boldsymbol{s}_{C,j} \in C \setminus D} \left\| \boldsymbol{s}_{D,i} - \boldsymbol{s}_{C,j} \right\| + \frac{1}{|D \cup C||C|} \sum_{\boldsymbol{s}_{D,i} \in D \setminus C} \sum_{\boldsymbol{s}_{C,j} \in C} \left\| \boldsymbol{s}_{D,i} - \boldsymbol{s}_{C,j} \right\|. \quad (9)$$

This metric seems to work sensibly for our purpose, and we study the possibility of using dist_F in Section 4. Yet another possibility is the modified Hausdorff distances. A variant is defined as the average of minimum (or alternatively the minimum squares) distances. Dubuisson and Jain (1994) show that the modified Hausdorff distance is a valid tool for object matching. The problem with this measure in our application is that it smooths the effect of outlier sites, whereas we believe that even a single outlier site could add valuable information to the design, giving knowledge of a larger area, and that should then have an important impact on the distance.

In summary, we use the regular Hausdorff distance h in (7) to model design dissimilarities. When building the covariance matrix in the GP surrogate formulation (6) it is important to

guarantee that the matrix is positive definite. This is more complex to prove when dealing with non-Euclidean distances (Bachoc et al., 2020), such as the Hausdorff distance. Schabenberger and Gotway (2017), chapter 4.8, suggest to first use multidimensional scaling (Borg et al., 2018) for projecting the Hausdorff distances in Euclidean space, and then build the covariance matrix. We follow the same approach for $h(D, C)$. We tested various parametric correlation functions and ended up with an exponential type $K(D, C) = \exp(-h(D, C)/\theta_H)$, which seems to give the best fit to our cases. In a situation with spatial covariates $\boldsymbol{z}(\boldsymbol{s}_i) = \boldsymbol{z}_i$, $i = 1, \ldots, n$, the expression can be modified to include the distance between covariates as well. The mean and variance–covariance model parameters are specified by empirical mean and a robust weighted least squares matching implementation to the empirical variogram, see e.g. chapter 2.6.2 in Chiles and Delfiner (2012). This approach is faster and seems more computationally reliable than maximum likelihood estimation of θ_H in our setting. In the optimization it is more robust to avoid the far distances, prone to hold elements not captured by the GP surrogate. Based on our experiments, outliers in the evaluation $I(D)$ give a difficult likelihood surface for the correlation parameters. This tends to be more of a problem for the suggested design distances than for ordinary spatial statistics applications with distances between spatial sites alone.

3.3 | Expected improvement

The number of possible designs is huge, and in most situations the evaluation of $I(D)$ requires substantial computational resources. Only in special cases, such as two-action decision situations and Gaussian distributions (Section 4), can one write down closed form solutions for the VOI expression (Equation (4)). In most real-world situations, the required integrals are solved by sophisticated analytical or numerical approximation methods or Monte Carlo sampling. Hence, it is not feasible to evaluate $I(D)$ for all designs D. The goal is then to find the optimal design in as few evaluation of $I(D)$ as possible. We use EI as an acquisition function (Frazier, 2018; Gramacy, 2020) to guide the evaluation of designs. The role of the acquisition function is then to find promising designs D for evaluating $I(D)$. The chosen designs are the ones that maximize the acquisition function. This procedure is run iteratively, relying on the updated distribution for $I(D)$, given $\mathcal{F}$. Assuming m_0 initial evaluations, after t iterations with m evaluations of $I(D)$ each time, the EI is defined by

$$\mathrm{EI}(D; \mathcal{F}) = \mathbb{E}\left(I(D) - I^+ | \mathcal{F}\right), \quad I^+ = \max\left\{I_{D_{(1)}}, \ldots, I_{D_{(m_0+mt)}}\right\}. \tag{10}$$

In the case of a GP surrogate model for $I(D)$, there is a closed form solution for EI, see for example, Brochu et al. (2010). We have

$$\mathrm{EI}(D; \mathcal{F}) = \left(\mu(D; \mathcal{F}) - I^+\right)\Phi(z) + \sigma(D; \mathcal{F})\phi(z), \quad z = \frac{\mu(D; \mathcal{F}) - I^+}{\sigma(D; \mathcal{F})}, \tag{11}$$

where Φ and ϕ are, respectively, the cdf and pdf of a standard Gaussian distribution; $\mu(D; \mathcal{F})$ and $\sigma^2(D; \mathcal{F})$ are the conditional mean and variance defined in (6).

By having the GP surrogate model, and accepting that EI is a useful acquisition function, the problem of maximizing $I(D)$ is now transformed to the problem of maximizing EI. This is relatively fast to compute for several designs, and the ones with large EI are selected for further evaluation.

3.4 | Algorithm

The iterative algorithm is summarized in Algorithm 1 where we describe the methodology in pseudocode. We let $\mathcal{F}_t$ denote all design evaluations done up to and including iteration t, while $D_{t,(1)}, \ldots, D_{t,(m)}$ and $I_{t,(1)}, \ldots, I_{t,(m)}$ denote the design and information gain evaluations at iteration t. For the size m_0 initial batch, the designs are randomly selected from all possible designs. At each iteration t, we augment the $\mathcal{F}_t$ set with m new evaluations. We keep track of the current best design D^+ and the associated information gain I^+. Via the iterative procedure, the current maximum in information gain will not decrease and eventually reach the global maximum and return the optimal design. In practice, the algorithm terminates when the maximum value for information gain has not increased over a trailing buffer of iterations (ΔI^+) or if a maximum number of iterations ($T_{\max}$) is reached.

Algorithm 1. Search for designs by Bayesian optimization

Result: Design D^+ with the largest information gain I^+.
Iteration $t = 0$;
$\Delta I^+ = 1$;
Evaluate $I(D)$ for m_0 randomly selected designs $D_{t,(1)}, \ldots, D_{t,(m_0)}$ to get $I_{t,(1)}, \ldots, I_{t,(m_0)}$;
$I^+ = \max\left\{I_{t,(1)}, \ldots, I_{t,(m_0)}\right\}$;
$\mathcal{F}_t = \left\{(I_{t,(j)}, D_{t,(j)}); j = 1, \ldots, m_0\right\}$;
while $t \leq T_{\max}$ *or* $\Delta I^+ = 0$ **do**
  $t = t + 1$;
  Mix existing design sites and random sites to suggest M designs;
  Compute the Hausdorff distances for the suggested and available designs ; ▷ expression (7)
  Estimate parameters and fit a GP surrogate model for $I(D)$ given all evaluations ; ▷ expression (6)
  Compute EI over I^+ for each of the M design ; ▷ expression (11)
  Find the m designs with largest EI to obtain $D_{t,(1)}, \ldots, D_{t,(m)}$;
  Evaluate $I(D_{t,(1)}), \ldots, I(D_{t,(m)})$; ▷ expression (5)
  $I^+ = \max\left\{I^+, I_{t,(1)}, \ldots, I_{t,(m)}\right\}$, $D^+ = \{D; I(D) = I^+\}$;
  $\mathcal{F}_t = \mathcal{F}_{t-1} \cup \{(I_{t,(1)}, D_{t,(1)})\} \cup \ldots \cup \{(I_{t,(m)}, D_{t,(m)})\}$;
  Compute an average increase over the last buffer of iterations ΔI^+;
end

After each batch iteration, the Hausdorff distances and the GP surrogate model are updated. The GP surrogate model mean and covariance parameters are also re-estimated after every batch. This is as recommended by Gramacy (2020), even though the computational costs of parameter specification grow over batch iterations. When we use moment matching to the empirical variogram, the computational costs increase more slowly than with maximum likelihood estimation. At each iteration, the EI is computed for M designs while only the m ($m << M$) designs with largest EI, given the current evaluations, are selected for the batch $I(D)$ evaluation. The M proposed new designs are obtained using a technique not dissimilar from a classical genetic algorithm (Goldberg, 1989), where the proposal set are mixed to create new designs, while allowing new

sited to enter a design randomly. In this step a large part of the proposed designs come from the set of all possible designs, while the remaining parts comes from a set obtained by mixing the best sites of previous steps. When we process a new design we check if the information gain has already been evaluated for that design, and if so we replace that design with a new one. In this way we do not recompute designs that have already been evaluated. In selecting new designs we adopt a weighted random selection of the design size to cover all cardinalities.

We next demonstrate the performance of the algorithm via simulation studies and real-world examples. Simulation examples allow users to gain insight in the statistical properties of the methods, which are often difficult to attain via theoretical calculations. There have been limited theoretical studies of the heuristic search algorithms for design problems. It is difficult to provide generally applicable convergence derivations and asymptotic results. There are a few theoretical design results for greedy algorithms in the setting with submodular functions in the design criteria (Krause & Golovin, 2014). For the determinant design criteria (D-design), Madan et al. (2019) present worst-case bounds for the local search exchange algorithm and the greedy algorithm. In that paper, they further show that similar tight asymptotic guarantees cannot be proven for the sum of variance criteria (A-design). This means that a locally optimal design solution can be arbitrarily poor compared with the global optimum. Harman et al. (2020) show that a batch randomized exchange version of the vertex exchange method converges for the D-design criteria and under special modeling assumptions such as independent measurement errors. Their numerical experiments indicate that elements of randomness help convergence. In our setting, we also use a mix of promising directions from previous evaluations as well as random designs to find sets that are computed in the GP step. As seen in the other papers, the goal of this heuristic approach is that this will balance the exploitation and exploration of high-value design points.

4 | SIMULATION STUDY

We study the properties of the design algorithm in a situation with $n = 30$ spatial sites of interest (Figure 1). The case is largely inspired by the forestry application (Section 5.1). The profits $x_i = x(\boldsymbol{s}_i), x_i \in \mathbb{R}, i = 1, \dots, n$, of the spatial sites are the quantity of interest. The decision-maker can choose, at each site, not to take any action or to exploit that site. We are hence in a situation where there is high decision flexibility and decoupled value, meaning that the decision-maker is free to choose the best alternative in a given site without accounting for the other parts. The set of alternatives is thus defined by $\mathcal{A} = \{a_i; i = 1, \dots, n\}$, where $a_i = \{\text{not exploit}, \text{exploit}\} = \{0, 1\}$. At each site, the two-action value function is then

$$v(x_i, a_i) = \begin{cases} 0 & a_i = 0, \\ x_i & a_i = 1, \end{cases} \tag{12}$$

and the PV in (1) involves a separate maximization for each site, that is, PV$= \sum_{i=1}^{n} \max\{0, E(x_i)\}$.

The profits $\boldsymbol{x} = (x_1, \dots, x_n)$ are represented by a hierarchical Gaussian model with mean $\mathbb{E}(x_i|\boldsymbol{\beta}) = \beta_0 + \beta_1 z_i + \beta_2 z_i^2$, where explanatory variable $z_i = z(\boldsymbol{s}_i)$ classify the site to one of four (age) categories, and the regression coefficients have a trivariate Gaussian distribution $\boldsymbol{\beta} \sim N(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$. Defining length n vectors $\boldsymbol{z} = (z_1, \dots, z_n)^t$ and $\boldsymbol{z}^2 = (z_1^2, \dots, z_n^2)^t$, the profits $\boldsymbol{x}$ are multivariate Gaussian with mean and covariance $\boldsymbol{\mu}_{\boldsymbol{x}} = \mu_{\beta_0} + \mu_{\beta_1}\mathbf{z} + \mu_{\beta_1}\mathbf{z}^2$, and covariance

matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{\Sigma} + \begin{bmatrix}1 & \boldsymbol{z} & \boldsymbol{z}^2\end{bmatrix} \boldsymbol{\Sigma}_{\beta} \begin{bmatrix}1 & \boldsymbol{z} & \boldsymbol{z}^2\end{bmatrix}^t$. The matrix $\boldsymbol{\Sigma}$ holds the structural spatial variability with entries defined by a stationary variance term and a Matern correlation function: $\Sigma_{i,j} = \sigma_x^2 \left(1 + \eta \left\|\boldsymbol{s}_i - \boldsymbol{s}_j\right\|\right) \exp\left(-\eta \left\|\boldsymbol{s}_i - \boldsymbol{s}_j\right\|\right)$, where η is the spatial correlation decay parameter. The regression uncertainty is set to realistic inputs based on forestry data similar to that presented in Section 5.1.

Data $\boldsymbol{y}_D$ can be gathered at any subset of $|D| \leq n$ sites. These data will carry profit information at the design sites, and at other sites through the spatial dependence in the profits $\boldsymbol{x}$ and via learning the regression effect. We assume that gathered data will be directly indicative of the profits, but measured with Gaussian additive noise. The conditional model for the data, given the profits, is then defined by

$$\boldsymbol{y}_D = \boldsymbol{G}_D \boldsymbol{x} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(0, \boldsymbol{T}_D), \tag{13}$$

where the size $|D| \times n$ matrix $\boldsymbol{G}_D$ picks the design sites by having one 1 entry in each row at the index of the sampled site and otherwise 0 entries. Moreover, $N(0, \boldsymbol{T}_D)$ denotes a random Gaussian vector with zero-mean and covariance matrix $\boldsymbol{T}_D = \tau_D^2 \boldsymbol{I}_{|D|}$.

The optimal design size and configuration are chosen by the decision-maker to maximize the VOI compared with the costs of data gathering. When the value function is in the form of (12) and the profits are Gaussian and measured with Gaussian additive noise, it is possible to compute the VOI in a closed form for each design (Bhattacharjya et al., 2013). The closed form calculation builds on the distribution of the conditional mean $\boldsymbol{\mu}_{x|y_D}$ with respect to the random data $\boldsymbol{y}_D$, which is Gaussian with $\mathbb{E}(\boldsymbol{\mu}_{x|y_D}) = \boldsymbol{\mu}_x$ and $\mathrm{Var}(\boldsymbol{\mu}_{x|y_D}) = \boldsymbol{\Sigma}_x \boldsymbol{G}_D^t (\boldsymbol{G}_D \boldsymbol{\Sigma}_x \boldsymbol{G}_D^t + \boldsymbol{T}_D)^{-1} \boldsymbol{G}_D \boldsymbol{\Sigma}_x = \boldsymbol{R}$. The PoV in (3) is then

$$\mathrm{PoV}(D) = \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{y}_D} \left[\max\{0, \mathbb{E}(v(x_i, a_i)|\boldsymbol{y}_D)\}\right] = \sum_{i=1}^{n} \left(\mu_{x_i} \Phi\left(\frac{\mu_{x_i}}{r_i}\right) + r_i \phi\left(\frac{\mu_{x_i}}{r_i}\right)\right), \tag{14}$$

where μ_{x_i} is element i in the mean vector for $\boldsymbol{x}$, and $r_i = \sqrt{R_{i,i}}$ is available from the ith diagonal entry of $\boldsymbol{R}$.

In this simulation study the costs $C(D)$ increase with the size of the design. We write $C(D)$ as a weighted sum ($C(D) = \sum_i^{|D|} \omega_i C(s_{D,i})$) of the costs of acquiring data in each site ($s_{D,i}$) of the design. The cost $C(s_{D,i})$ depends on the covariates. The weights ω_i are greater than 1, and they increase as the size of the design grows. If we consider for example a design with $|D| = 5$, and assign weights $\omega = (1, 1.1, 1.2, 1.3, 1.4)$, the costs rapidly increase with the size of the design. In this case for designs with $|D| > 5$, the VOI never exceeds the cost. This means that we can focus on all sites combinations up to size 5, and there are then about $1.7 \cdot 10^5$ possible designs. It is feasible to compute the exact VOI for all designs, and compare the optimal designs with the results obtained by Algorithm 1. The Bayesian optimization approach is run for 15 iterations and for a number of replicate restarts. Each batch iteration consists of $m_0 = m = 50$ evaluations.

Via simulation results we first study convergence over the restarts and as the number of iterations increases. After five iterations none of the restarts reached the maximum value, and overall they are quite far from the optimal value. After 10 iterations we see improvement; although none reached the maximum, all the iterations end with values rather close to the optimal one. After 15 iterations, one restart has reached the optimal value and the others are relatively high-ranked. After 30 iterations two restart have reached the optimal value and all restart show very good performance. Here the number of possible sites is large so several iteration would be necessary to finally see most of the restarts converging. We also run an experiment with a smaller design set,

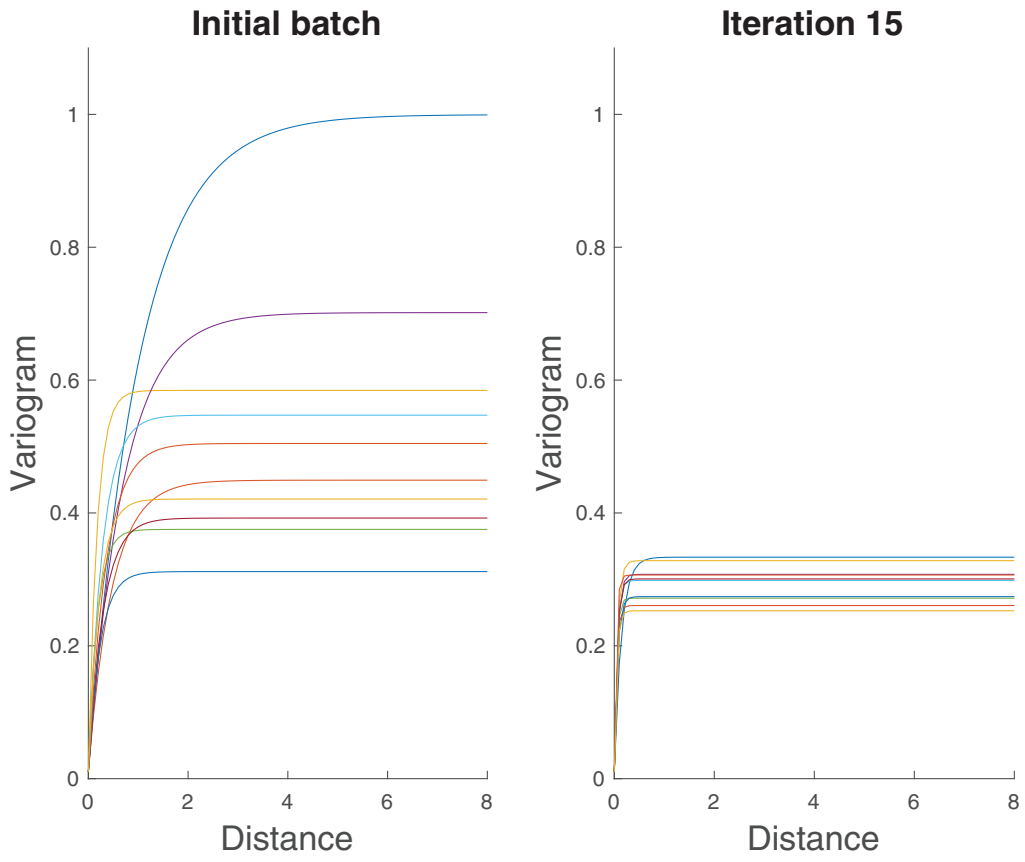

FIGURE 3 Variograms estimated from 10 replicate re-starts of Algorithm 1. The left display is after the initial batch, while the right display is after 15 iterations of batch evaluations [Colour figure can be viewed at wileyonlinelibrary.com]

having only a subset of 10 among the 30 sites as feasible locations. When it is never profitable to select more than five sites, there are only 637 possible designs for this situation. The results show that already after three iterations (200 evaluations) 4 out of 10 restarts reach the optimal value. After seven iterations (400 evaluations), six restarts have reaching the maximum value, two restarts get the second largest value, and one third largest value. Overall, our simulations show that the suggested algorithm tends to converge quickly to very high-ranked designs, but it might take many iterations to reach the single best one.

We next study the parameter specification in the GP surrogate model. This is done at every batch iteration and for each of the ten replicate restarts of the algorithm. Figure 3 shows the fitted variograms after the initial batch (50 evaluations) and after 15 iterations (800 evaluations). Clearly, the variability in the variograms is reduced with more evaluations, because there is more data to infer the model parameters. Also, the variance of the GP surrogate model, which is indicated by the sill of the variogram, appears to be a bit lower after 15 iterations. This likely occurs because of the preferential evaluations, where the algorithm tries to find the highest values of $I(D)$.

We now compare the results obtained by Algorithm 1 with that of other methods: (i) sequential selection algorithm, (ii) modified exchange algorithm (Mitchell, 1974; Royle, 2002), and (iii) using dist_F (expression (9)) instead of the Hausdorff distance.

Method (i) sequentially chooses the best site in a forward selection: it first evaluates each of the single sites and selects the one that maximizes information gain $I(D)$. Next, it looks at all the couples that contain the selected site, and finds which of these couples that has the largest VOI gain. It proceeds in this way until $I(D)$ does no longer increase. The computational cost of the sequential algorithm is relatively small as the total number of VOI evaluations is $\sum_{i=1}^{|D|+1}(|D|+1-i) = 140$. For our reference model, the sequential selection algorithm ends up with design $(1, 9, 23)$. This is quite far from the global maximum and also significantly below most of the replicate results achieved using the Bayesian optimization approach.

Method (ii) iteratively exchanges, adds or removes random sites to an existing design. When a random site is added to the design, we have cardinality $|D| \to |D|+1$, while the cardinality $|D| \to |D|-1$ when a random site is removed from the design. For a random exchange the cardinality remains the same. For each suggested design the information gain $I(D)$ is evaluated, and one keeps track of the best design so far. The solution paths of the exchange algorithm will change every time because of the random selection of moves. The exchange algorithm is often able to find the best design in less the 10^4 iterations, and could probably get there faster with some kind of weighted resampling. Still, it seems to require more evaluations than the Bayesian optimization approach, and we believe it is difficult to tune this method for larger-size problems where the number of required evaluations will also increase dramatically.

Method (iii) using dist_F has performance very similar to that of using the Hausdorff distance, but does not reach the optimum as often. The computational cost is about the same.

In our context multidimensional scaling helps to visualize the Hausdorff distances between designs in a two-dimensional space that largely maintains the distances. Figure 4 shows the best 3000 designs (in grey) projected in this two-dimensional space. The pink star represents the best design while the green diamond is the design obtained with the sequential selection method. In this display we indicate typical paths that Algorithm 1 (red) and the exchange algorithm (blue) take to get their final best results. We clearly see the randomness of the path taken by the exchange algorithm to reach the maximum. We do not show all the sites explored to get to the maximum, only the ones of local maximum. Algorithm 1 reaches the maximum following a much more efficient path.

For further comparison with the exchange algorithm, we study the performances over 100 restarts, and observe the maximum score after 250 (Figure 5a), 500 (Figure 5b), and 800 (Figure 5c) evaluations of $I(D)$. Figure 5 (red, solid) shows results of Algorithm 1, while dashed blue colors are used to represent the ones from the exchange algorithm. Here, the spikes at the right end indicate the fraction of restart replicates that has reached the optimum. The runs yet to reach this optimum are plotted in a kernel density display. We observe that the Bayesian optimization method gets larger values of $I(D)$ after relatively few iterations because the exchange algorithm struggles with its random structure.

We perform sensitivity analysis to gain insight in the effect of having different model and algorithm input parameters. We modify the prior model variance σ^2 of the profits to high and low values. This represent different models, and we check how the prior variability influences the optimization results. We further study the effect of the algorithm's iterations and re-estimation strategy for GP surrogate model parameter specification. We compare algorithm performance metrics for combinations of these inputs, in completely independent runs. Metrics include the highest value for $I(D)$ and how many times we get a score among the best 100 $I(D)$ values over 10 replicate restarts.

Table 1 shows the results of this sensitivity analysis.

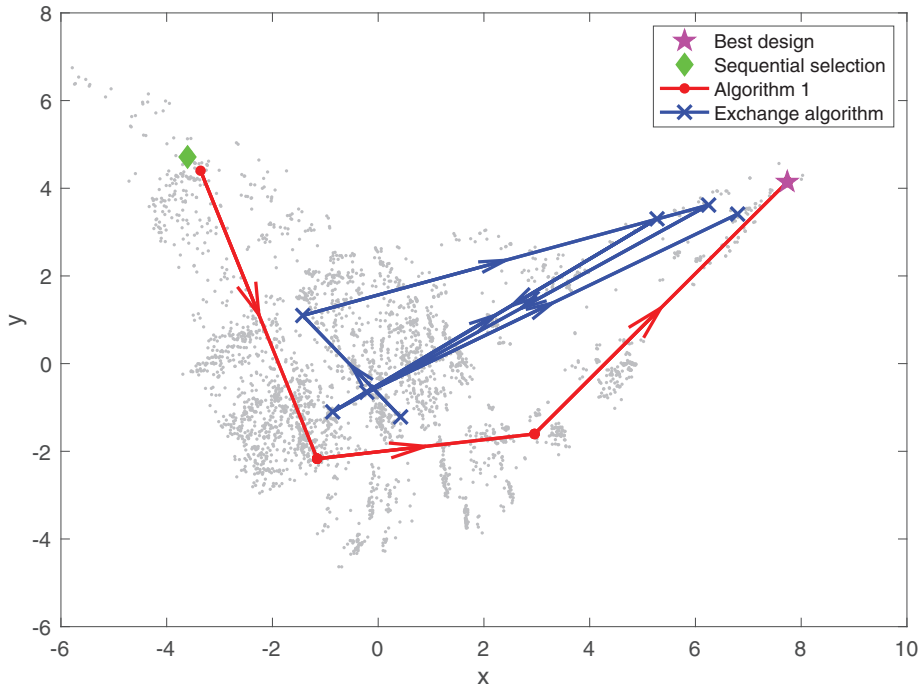

FIGURE 4 Representation of the spatial Hausdorff distance for the best 3000 designs (grey) in a two-dimensional Euclidean space using multidimensional scaling technique. The pink star represents the best design and the green diamond the best design from the sequential selection. The red dots and lines represents the path of Algorithm 1 while the blue crosses and lines represents the results from the exchange algorithm [Colour figure can be viewed at wileyonlinelibrary.com]

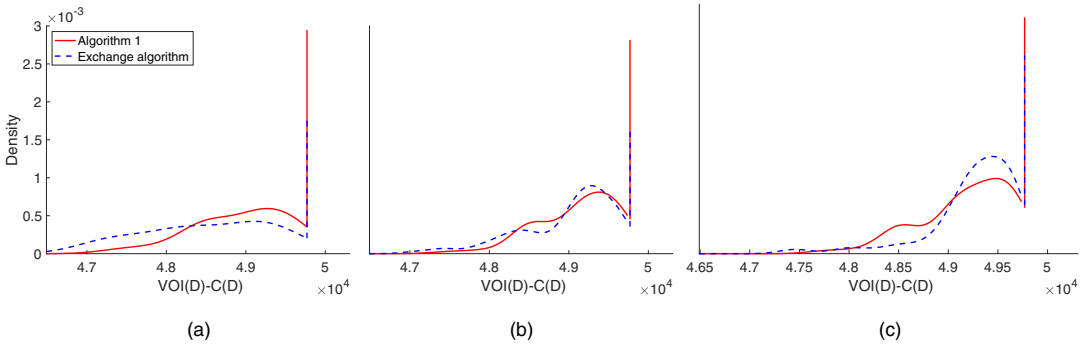

FIGURE 5 Comparison of the performances of Algorithm 1 (red) and exchange algorithm (dashed blue) over 100 replicate restarts. The Bayesian optimization approach is able to get large values of $I(D)$ after few iterations. When the number of evaluations grows the exchange algorithm starts to perform well. (a) 250 evaluations of $I(D)$ (in €); (b) 500 evaluations of $I(D)$ (in €); (c) 800 evaluations of $I(D)$ (in €) [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 1 Sensitivity analysis of different inputs on the algorithm performance. The column "Algorithm 1" represents the highest $I(D)$ among the 10 replicates, with its rank in parenthesis, the column "% best 100" indicates the fraction of times we get a score among the best 100 information gain values over the replicates. Note that low and high σ_x represent different models with varying value of information and hence also different values of $I(D)$

σ_x^2	Parameter re-estimation	Iterations ($T_{\max}$)	Algorithm 1	% best 100
Low	off	5	€33,011 (3)	100%
High	off	5	€85,740 (1)	20%
Low	on	5	€33,012 (1)	100%
High	on	5	€84,486 (18)	60%
Low	off	15	€33,011 (3)	100%
High	off	15	€85,365 (3)	100%
Low	on	15	€33,012 (1)	100%
High	on	15	€85,740 (1)	100%

Overall, higher prior variability appears to be more difficult for the algorithm, especially with few iterations, where only 20 and 60% of the runs resulted in the top 100 ranking designs. Still, one lucky restart run without any re-estimation ended up with the highest rank. The performance clearly increases with iterations since all restarts are in the top 100 ranking after 15 iterations, no matter prior variance high/low or re-estimation on/off. Re-estimation of parameters gives improved performance, especially when there is much prior uncertainty in the profits model. The running time is a bit more than three times larger for 15 iterations compared with 5, because of the growing matrix expressions in the GP surrogate calculations. Doing parameter re-estimation after every batch also takes some additional time, but the empirical variogram calculation and least squares matching is very fast. The basic assumption of our work is that for real-world settings, most of the computer time is spent on evaluating $I(D)$, and the number of such evaluations is considered to be the main computational restriction.

5 | EXAMPLES

5.1 | Forestry

This example regards forest management and conservation (Eyvindson et al., 2017; Kangas et al., 2008). In this application the decision-maker must choose to conserve forest stands or not. The decision-maker is here a governmental institute that has a budget for conservation. The forest stands are owned by private owners who may harvest the timber unless the forest is conserved. In order to conserve a forest stand, the institute must pay a compensation (b_i) to the forest owner. When a forest stand is conserved, the ecological benefit (r_i) is proportional to a biodiversity indicator.

The study is inspired by data analyzed by Eyvindson et al. (2019). The data consist of 70 forest stands (sites) of various size from the Satakunta region in southwest Finland (Figure 6a). Each stand is classified according to the age class (1: <= 80, 2: 81 − 95, 3: 96 − 110, 4: > 110 years). Figure 6b sketches forest stands, with colors identifying the different age groups.

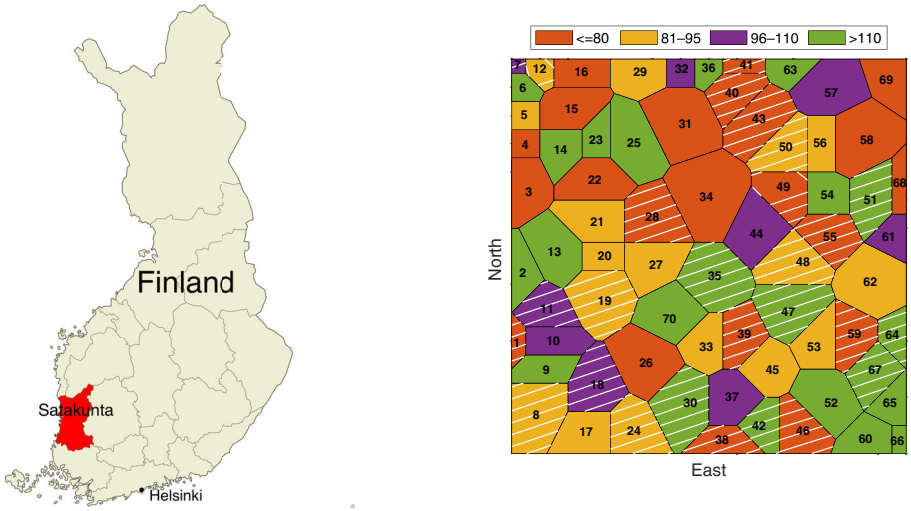

FIGURE 6 Study area for forest conservation. (a) The region of interest is located in the southwest of Finland. The red identify the geographic location of the Satakunta region (Wikimedia Commons, 2010). (b) Forest stands of different size and age numbered from 1 to 70. Each color identify a different age group, orange: <= 80, yellow: 81–95, violet: 96–110, green: >110 years. The hatched regions correspond to the best design [Colour figure can be viewed at wileyonlinelibrary.com]

The possible alternatives for the decision-maker are $\mathcal{A} = \{a_i; i = 1, \dots, n\}$, where $a_i = \{\text{not conserve, conserve}\} = \{0, 1\}$ at stand i (Eyvindson et al., 2019). We let x_i be the log-intensity of the number of wood inhabiting fungi at stand $i = 1, \dots, n$. This number is a commonly used biodiversity indicator. The value function is then

$$v(x_i, a_i) = \begin{cases} 0 & a_i = 0, \\ re^{x_i} - b_i & a_i = 1. \end{cases} \tag{15}$$

The log-intensities are here modeled with a multivariate Gaussian distribution over the stands. In doing so, the age of the forest stand is treated as a covariate $\boldsymbol{z}$ in the simulation study. The mean and the covariance matrix of log intensities vector variable $\boldsymbol{x}$ are computed by double mean and variance over the regression uncertainty (Section 4).

Designs are constructed to gather information that can assist the decision maker. There are age-dependent inventory costs, so it is important to plan wisely and obtain effective designs at a low overall cost. The measurements of species richness in fungi are defined with a Poisson likelihood function

$$y_i | x_i \sim \text{Poisson}\left(e^{x_i}\right), \tag{16}$$

assuming conditional independence between the stands and constant area for each inventory.

The VOI is here defined by

$$\text{VOI}(D) = \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{y}_D}\left[\max\{0, \mathbb{E}(re^{x_i} - b_i | \boldsymbol{y}_D)\}\right] - \sum_{i=1}^{n} \max\{0, \mathbb{E}(re^{x_i} - b_i)\}, \tag{17}$$

and the information gain $I(D)$ is obtained as the difference $\text{VOI}(D) - C(D)$ where $C(D)$ denotes the inventory costs, obtained by accumulating costs over all design sites. The inventory cost depends on the age of the forest (1: €5100; 2: €5800; 3: €6200; 4: €5600). We use the method developed by Evangelou and Eidsvik (2017) to compute the VOI in (17). This is an approximation where the evaluation of $\text{VOI}(D)$ relies on iterative matrix linearizations and refitting Gaussian approximations going into the Laplace approximation. Even though this evaluation can be done in reasonable time for a candidate design, it is relatively time-demanding, and as there is a total of $1.18 \cdot 10^{21}$ possible designs, it is not feasible to calculate the VOI for all of them to find the optimal design. Instead, we use the suggested Bayesian optimization method to find efficient designs.

We initiate the algorithm by evaluating $m_0 = 50$ random designs of various cardinalities. The GP surrogate model parameters (σ, θ_H) are specified for each of 10 such re-starts. Based on this, the approximate 20 percentiles of these parameters are (€76,750, 0.25) and 80 percentiles are (€136,140, 5.4). At each batch of size $m = 50$, new evaluations are selected using the EI acquisition function. The parameters are re-estimated at each batch, and after 15 iteration the approximate 20 percentiles are (€42,430, 0.07) and 80 percentiles are (€135,500, 1.27). The results show that one gets lower SDs and correlation range by re-estimation at each batch. In the actual optimization, the re-estimation tends to give slightly faster improvements for I^+ in the algorithm. For the average value μ of the GP surrogate model for $I(D)$, the initial approximate 20 percentile is €325,320 and 80 percentile is €354,270. After 15 iterations the approximate 20 percentile is €301,520 and 80 percentile is €352,430. It is perhaps surprising that the 20 percentile decreases when the goal is to maximize the function. However, the exploration elements of the algorithm can also lead to the evaluation of very poor designs, and possibly outliers with very small $I(D)$.

The results from the Bayesian optimization are shown in Figure 7.

Here, we plot the optimum evaluation so far, for each restart (different lines). Even though we do not know the optimal solution in this case, the results improve over batches and this indicates that the algorithm finds efficient designs within a few batch iterations.

Table 2 provides a list of the top five largest values of $I(D)$ obtained by running the algorithm, together with the associated design. The best designs have many sites in common, even though the designs have different cardinalities. In this way the algorithm spots the sites that carry more information. The highest value of $I(D)$ corresponds to a rather large set of $|D| = 26$. This design is illustrated using the hatched areas in Figure 6b, where we observe that the stands of the design tend to spread and cover both the geographical region and also the various age levels.

Similar to what was done in the simulation study, we also run both the exchange algorithm and the sequential selection method. The exchange algorithm gets a largest replicate information gain of only $I(D) =$ €491,760 after 800 evaluations, and is not doing so well in this case. The sequential selection algorithm gives $I(D) =$ €552,930 with 2485 evaluations. The associated design is $D = (1, 2, 3, 4, 5, 8, 15, 16, 19, 24, 26, 32, 40, 41, 44, 47, 49, 57, 59, 60, 66, 69)$. In this example the sequential selection algorithm hence performs better than the iterative Bayesian optimization in Table 2, at a cost of extra VOI evaluations to find the sequential solution. With this in mind, we added the sequential solution to the evaluations of the Bayesian optimization method, and continued to run that algorithm. We then achieved slightly larger information gain for designs very similar to the one detected with the sequential search, but no significant improvement. We hence suspect that the sequential method gives a near optimal solution for this example.

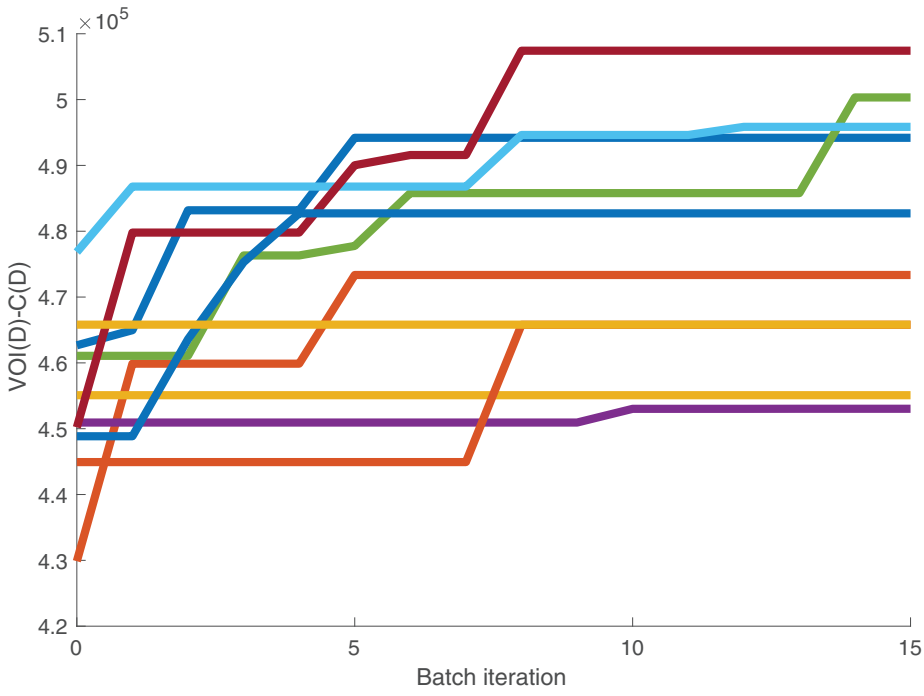

FIGURE 7 Performance of the Bayesian optimization algorithm for the application in forest conservation. The algorithm gives improved results over batches for the replicate restarts [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 2 The best five designs obtained over 10 re-start replicates of the algorithm in the forestry example, listed in descending order

Design D	$I(D)$
1, 7, 8, 11, 12, 18, 19, 24, 28, 30, 35, 38, 40, 41, 42, 43, 46, 47, 48, 49, 50, 51, 55, 59, 64, 67	€507,420
3, 11, 12, 13, 15, 16, 17, 18, 21, 22, 26, 28, 30, 32, 39, 44, 47, 48, 49, 50, 52, 56, 57, 59, 62	€500,320
1, 4, 6, 8, 9, 11, 13, 15, 16, 17, 20, 24, 25, 26, 27, 28, 31, 32, 34, 44, 45, 48, 49, 50, 51, 53, 57, 62, 63, 66, 68	€495,860
2, 5, 6, 8, 9, 13, 14, 17, 19, 20, 23, 24, 30, 32, 35, 37, 40, 42, 44, 46, 49, 51, 53, 54, 56, 57, 61, 62, 66, 69	€494,190
8, 9, 12, 13, 18, 19, 22, 23, 26, 27, 32, 33, 36, 38, 40, 42, 51, 52, 53, 56, 60, 61, 62, 65, 68	€482,710

5.2 | Petroleum drilling risks

This example regards decision making during drilling operations in the petroleum industry (Lothe et al., 2019; Paglia et al., 2020). We study a drilling situation in the Alvheim oil field located in the central part of the North Sea, on the Norwegian continental shelf (Figure 8a). The field is divided in 68 compartments (Figure 8b) separated by faults. The circles in Figure 8b represent wells, and Figure 8c highlights these to indicate the decision site (black star) and the potential data

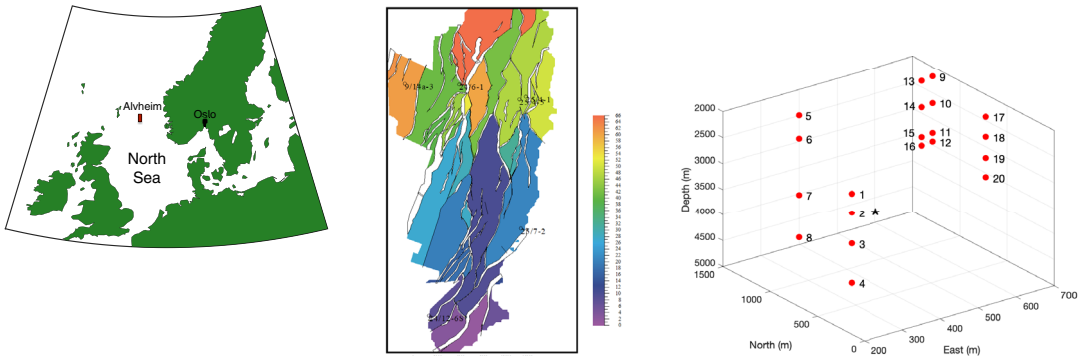

FIGURE 8 The study area is an offshore oil and gas field in the central part of the North Sea. (a) Geographical location of Alvheim. The rectangle indicates the position of the field. (b) Map view of the oil field, circles indicate the locations of wells for data gathering. Different colours are used to identify different geological compartments. (c) Three-dimensional view of the location of the measurements in red and the decision site in black star [Colour figure can be viewed at wileyonlinelibrary.com]

gathering site (red dots) in a three-dimensional plot. In the following we describe this decision situation and the opportunities for data gathering to make improved decisions.

Drilling operations at the Alvheim field are characterized by the risk of overpressure, which occurs when the pore pressure in the rock exceeds the hydrostatic pressure. To prevent big hazards, the drilling mud pressure must be calibrated. We study a specific layer, located at about 3700 m depth, as marked in black in Figure 8c. This layer is composed of mainly shale rocks and believed to be at drilling risk. The decision maker, which is the petroleum company in this case, must decide if it is safe enough to just keep drilling, or if they should set casing to strengthen the well because of a high risk of blowout. The alternatives are $\mathcal{A} = \{0, 1\} = \{\text{keep drilling}, \text{ set casing}\}$. To set casing is an expensive operation and it will reduce the borehole diameter, so the decision-maker is interested in trying to postpone this operation, if not necessary because of very high risk. The value function of the decision is $v(x, a = 0) = -c_0(\text{LB}(x) - x)$ and $v(x, a = 1) = -c_1(\text{LB}(x) - x)$, where x is the unknown pore pressure variable and LB is the lower bound of the mud weight drilling window. Here, c_0 is the cost when one keeps drilling, while c_1 is the additional cost of casing. We note that the costs stretch the value functions so that it becomes more valuable to set casing instead of continued drilling for some values of pore pressure. Critically, the mud weight is used during the drilling of a well to exert a pressure on the borehole wall and avoid well collapse, and it is difficult to make decisions when the pore pressure is not known. Please keep in mind that there are a number of other parameters that would also affect LB, but pore pressure is an important parameter that is always taken into consideration during drilling operation (Moos et al., 2004).

Figure 8c (red) shows the possible measurement sites. The design will entail any combination of these sites, and the measurements gathered at the design sites will be informative of the pore pressure where they are made and at other sites via the statistical model formulation. There are five wells where accurate measurements of pore pressure can be gathered in four different layers. The cost of data acquisition is assumed to be the same for each well. However it will be cheaper to obtain more information from the same well, since the tools for gathering the measurements have been already been placed into the well.

TABLE 3 The largest five designs obtained over 10 restarts of the algorithm in the petroleum drilling risk example, listed in descending order

Design (D)	$I(D)$
5, 6, 8, 14	€9,115,100
6, 7, 8, 15	€9,115,100
5, 7, 8,11	€9,115,100
3, 5, 6, 7, 8, 19, 20	€9,114,900
3, 5, 6, 7, 8, 9, 10, 12	€9,114,600

Based on seismic data from the region along with geological simulations of pressure build-up and release (Borge, 2000; Lothe, 2004; Paglia et al., 2019) we fit an initial multivariate Gaussian model for the pore pressure at the well location and at neighboring wells. The pore pressure measurements $\boldsymbol{y}_D$ in neighboring wells will then be indicative of the pore pressure in the target well via correlations. These modeling assumptions simplify the computation of the conditional expectation of pore pressure in the well of interest, given data obtained in the other wells. The main challenge of the VOI evaluation in the current setting is then to compute the expectation of the nonlinear value function and to find its integral over all possible data $\boldsymbol{y}_D$. These expectations required for the PV (expression (1)) and the PoV (expression (3)) are here calculated with numerical approximations of the integrals. This entails rather time demanding computations for the value function v over discretized levels of pore pressure, where the LB is then finally obtained with a spline interpolation. We let Δx_j and $\Delta_{\boldsymbol{y}_D}$ denote the distances between two consecutive discretized levels of x and $\boldsymbol{y}_D$, respectively. The VOI approximation can then be computed from

$$\begin{aligned}
\mathrm{PV} &= \max_{a\in\mathcal{A}}\{\mathbb{E}\left(v(\boldsymbol{x},a)\right)\} \approx \max_{a\in\mathcal{A}}\left\{\sum_j v(x_j,a)p(x_j)\Delta x_j\right\},\\
\mathrm{PoV}(D) &= \mathbb{E}_{\boldsymbol{y}_D}\left[\max_{a_\in\mathcal{A}}\{\mathbb{E}(v(\boldsymbol{x},a)|\boldsymbol{y}_D)\}\right] \approx \sum_{\boldsymbol{y}_d}\max_{a\in\mathcal{A}}\left\{\sum_j v(x_j,a)p(x_j|\boldsymbol{y}_D)\Delta x_j\right\}p(\boldsymbol{y}_D)\Delta_{\boldsymbol{y}_D},\\
\mathrm{VOI}(D) &= \mathrm{PoV}(D) - \mathrm{PV}.
\end{aligned} \tag{18}$$

Figure 9 shows the performance of Algorithm 1 for this application. As in the first example we do not know the design with the largest information gain value, but we observe a convergence of the method toward larger information gain as more batches are evaluated.

Table 3 shows the five largest values of $I(D)$ in descending order, together with the associated design. Once one starts gathering data at a specific well, acquiring more data at other depth is relatively cheap. This implies that the cost effective designs suggest to explore more then one depth for a single well. We notice from the table that there are designs with the same large value of $I(D)$, but with different sites configurations. All the best designs have four sites and have some common sites. We believe that these sites are highly informative.

In this case the sequential selection method gives a solution of €9,114,900 with 210 iterations, which is rather good but not as high as the best one achieved with the Bayesian optimization method. The exchange algorithm obtains €9,114,600 with 800 iterations. That small monetary

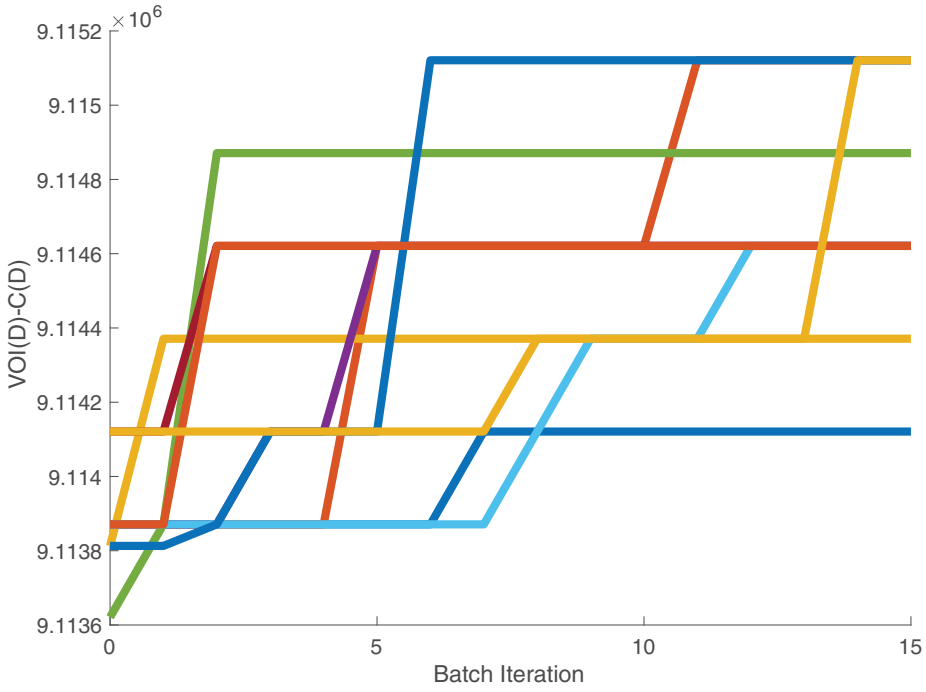

FIGURE 9 Performance of the Bayesian optimization algorithm for a drilling application. Over the 10 restarts, the algorithm converges to a large value of VOI-C [Colour figure can be viewed at wileyonlinelibrary.com]

differences for the three methods is due to the symmetry assumption of cost being the same for sites at the same depth in different wells.

6 | CLOSING REMARKS

The main purpose of this study is to develop an algorithm that can assist a decision-maker in choosing a spatial design for collecting information. The methodology has its applications in earth sciences, where data are often distributed over a spatial domain. We illustrated the approach by presenting one example from the forestry and one from petroleum.

We have adopted the Hausdorff distance to model dissimilarities between designs, and incorporate this in the kernel of a GP surrogate model. We demonstrated its use in examples. We believe that, depending on the field where the methodology is applied, other metrics can work as well. The Jaccard distance and the metrics introduced by Fujita (2013), can be valid alternatives to the Hausdorff distance. For instance, the Jaccard distance could work in situations where we are not too interested in spatial distance between designs, but in a more machine learning oriented context where one must select the appropriate number of sets for training, and because data may come from different sources it is important to guide the active learning wisely to cover appropriate features (Settles, 2012). The developed methodology could also be applied in subset selection problems such as the choice of individuals in epidemiological follow-up studies (Reinikainen et al., 2016) or in genotyping (Karvanen et al., 2009). It would be very useful to gain insight in

theoretical properties of the algorithm, at least in special situations, or potentially to connect the mix of design combinations in the algorithm to some useful theoretical properties.

The value function could be extended to include a parametric form. For instance, a possibility is to focus on learning the regression parameters $\boldsymbol{\beta}$ in the simulation study, or other parameters. This will go beyond only expected economic outputs, and rather contain additional (hybrid) terms in the prior and PoV related with the standard deviation in profits. The common space-covering geometric designs are sometimes considered robust because they do not use target specific prediction purposes. It is possible to blend our approach with other design constructions to balance multiple criteria.

It is possible to extend the study considering more challenging probability distributions for data, where the computation of VOI becomes more difficult. The combination of the VOI analysis with Bayesian optimization techniques gives us an efficient way to find satisfactory data gathering scheme. With the Bayesian optimization we move the problem of evaluating VOI to that of computing EI, which requires less computational effort. The total number of evaluation of VOI is considerably reduced. In situations where computing the information gain is computationally demanding or the number of alternatives to explore is too large, the developed methodology reduces the time of computation.

ORCID

Jacopo Paglia https://orcid.org/0000-0003-0314-9284
Juha Karvanen https://orcid.org/0000-0001-5530-769X

REFERENCES

Abbas, A. E., & Howard, R. A. (2015). *Foundations of decision analysis*. Pearson Higher Education.

Bachoc, F., Suvorikova, A., Ginsbourger, D., Loubes, J.-M., & Spokoiny, V. (2020). Gaussian processes with multidimensional distribution inputs via optimal transport and Hilbertian embedding. *Electronic Journal of Statistics*, *14*, 2742–2772.

Bhattacharjya, D., Eidsvik, J., & Mukerji, T. (2013). The value of information in portfolio problems with dependent projects. *Decision Analysis*, *10*, 341–351.

Binois, M., Huang, J., Gramacy, R. B., & Ludkovski, M. (2019). Replication or exploration? Sequential design for stochastic simulation experiments. *Technometrics*, *61*, 7–23.

Borg, I., Groenen, P. J., & Mair, P. (2018). *Applied multidimensional scaling and unfolding*. Springer.

Borge, H. (2000) *Fault controlled pressure modelling in sedimentary basins*. [Ph.D. thesis]. Norwegian University of Science and Technology.

Bouneffouf, D. (2016). Exponentiated gradient exploration for active learning. *Computers*, *5*, 1.

Brochu, E., Cora, V. M., & De Freitas, N. (2010) A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*.

Chiles, J.-P., & Delfiner, P. (2012). *Geostatistics: Modeling spatial uncertainty*. Wiley.

Diggle, P., & Lophaven, S. (2006). Bayesian geostatistical design. *Scandinavian Journal of Statistics*, *33*, 53–64.

Dobbie, M. J., Henderson, B. L., & Stevens, D. L., Jr. (2008). Sparse sampling: Spatial design for monitoring stream networks. *Statistics Surveys*, *2*, 113–153.

Drovandi, C. C., Holmes, C., McGree, J. M., Mengersen, K., Richardson, S., & Ryan, E. G. (2017). Principles of experimental design for big data analysis. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, *32*, 385.

Drovandi, C. C., McGree, J. M., & Pettitt, A. N. (2013). Sequential Monte Carlo for Bayesian sequentially designed experiments for discrete data. *Computational Statistics & Data Analysis*, *57*, 320–335.

Dubuisson, M.-P., & Jain, A. K. (1994) A modified Hausdorff distance for object matching. *Proceedings of 12th International Conference on Pattern Recognition* (Vol. 1, pp. 566–568). IEEE.

Eidsvik, J., Martinelli, G., & Bhattacharjya, D. (2018). Sequential information gathering schemes for spatial risk and decision analysis applications. *Stochastic Environmental Research and Risk Assessment*, *32*, 1163–1177.

Eidsvik, J., Mukerji, T., & Bhattacharjya, D. (2015). *Value of information in the earth sciences: Integrating spatial modeling and decision analysis*. Cambridge University Press.

Evangelou, E., & Eidsvik, J. (2017). The value of information for correlated GLMs. *Journal of Statistical Planning and Inference*, *180*, 30–48.

Eyvindson, K., Hakanen, J., Mönkkönen, M., Juutinen, A., & Karvanen, J. (2019). Value of information in multiple criteria decision making: An application to forest conservation. *Stochastic Environmental Research and Risk Assessment*, *33*, 2007–2018. https://doi.org/10.1007/s00477-019-01745-4

Eyvindson, K. J., Petty, A. D., & Kangas, A. S. (2017). Determining the appropriate timing of the next forest inventory: Incorporating forest owner risk preferences and the uncertainty of forest data quality. *Annals of Forest Science*, *74*, 2.

Frazier, P. I. (2018) A tutorial on Bayesian optimization. *arXiv preprint arXiv:1807.02811*.

Fujita, O. (2013). Metrics based on average distance between sets. *Japan Journal of Industrial and Applied Mathematics*, *30*, 1–19.

García-Ródenas, R., García-García, J. C., López-Fidalgo, J., Ángel Martín-Baos, J., & Wong, W. K. (2020). A comparison of general-purpose optimization algorithms for finding optimal approximate experimental designs. *Computational Statistics & Data Analysis*, *144*, 106844.

Ginsbourger, D., Baccou, J., Chevalier, C., & Perales, F. (2016). *Design of computer experiments using competing distances between set-valued inputs*. In *mODa 11-advances in model-oriented design and analysis* (pp. 123–131). Springer.

Goldberg, D. E. (1989). *Genetic algorithms in search, optimization and machine learning* (1st ed.). Addison-Wesley Longman Publishing Co., Inc.

Grafström, A., Lundström, N. L., & Schelin, L. (2012). Spatially balanced sampling through the pivotal method. *Biometrics*, *68*, 514–520.

Gramacy, R. B. (2020). *Surrogates: Gaussian process modeling, design, and optimization for the applied sciences*. CRC Press.

Harman, R., Filová, L., & Richtárik, P. (2020). A randomized exchange algorithm for computing optimal approximate designs of experiments. *Journal of the American Statistical Association*, *115*, 348–361.

Huan, X., & Marzouk, Y. M. (2013). Simulation-based optimal Bayesian experimental design for nonlinear systems. *Journal of Computational Physics*, *232*, 288–317.

Huttenlocher, D. P., Rucklidge, W. J., & Klanderman, G. A. (1992) Comparing images using the Hausdorff distance under translation. *Proceedings 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 654–656). IEEE.

Kangas, A., Kangas, J., & Kurttila, M. (2008). *Decision support for forest management* (Vol. *16*). Springer.

Karvanen, J., Kulathinal, S., & Gasbarra, D. (2009). Optimal designs to select individuals for genotyping conditional on observed binary or survival outcomes and non-genetic covariates. *Computational Statistics & Data Analysis*, *53*, 1782–1793.

Krause, A., & Golovin, D. (2014). Submodular function maximization. *Tractability*, *3*, 71–104.

Levandowsky, M., & Winter, D. (1971). Distance between sets. *Nature*, *234*, 34–35.

Lothe, A. E. (2004) *Simulations of hydraulic fracturing and leakage in sedimentary basins*. [Ph.D. thesis]. University of Bergen.

Lothe, A. E., Cerasi, P., & Aghito, M. (2019). Digitized uncertainty handling of pore pressure and mud-weight window ahead of bit: North Sea example. *SPE Journal*, *25*, 24. https://doi.org/10.2118/189665-PA

Madan, V., Singh, M., Tantipongpipat, U., & Xie, W. (2019) Combinatorial algorithms for optimal design. *Proceedings of the Conference on Learning Theory* (pp. 2210–2258). PMLR.

Min, D., Zhilin, L., & Xiaoyong, C. (2007). Extended Hausdorff distance for spatial objects in GIS. *International Journal of Geographical Information Science*, *21*, 459–475.

Mitchell, T. J. (1974). An algorithm for the construction of "d-optimal" experimental designs. *Technometrics*, *16*, 203–210.

Moos, D., Peska, P., Ward, C., and Brehm, A. (2004) Quantitative risk assessment applied to pre-drill pore pressure, sealing potential, and mud window predictions from seismic data. *Proceedings of the 6th North America Rock Mechanics Symposium (NARMS) Gulf Rocks 2004*. American Rock Mechanics Association.

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

Paglia, J., Eidsvik, J., & Cerasi, P. (2020). Workflow for sensitivity analysis and decision making on the lower limit of the mud-weight window in an overpressured formation. *SPE Journal*, *25*, 203830. https://doi.org/10.2118/203830-PA

Paglia, J., Eidsvik, J., Grøver, A., & Lothe, A. E. (2019). Statistical modeling for real-time pore pressure prediction from predrill analysis and well logs. *Geophysics*, *84*, ID1–ID12.

Reinikainen, J., Karvanen, J., & Tolonen, H. (2016). Optimal selection of individuals for repeated covariate measurements in follow-up studies. *Statistical Methods in Medical Research*, *25*, 2420–2433.

Royle, J. A. (2002). Exchange algorithms for constructing large spatial designs. *Journal of Statistical Planning and Inference*, *100*, 121–134.

Ryan, E. G., Drovandi, C. C., McGree, J. M., & Pettitt, A. N. (2016). A review of modern computational algorithms for Bayesian optimal design. *International Statistical Review*, *84*, 128–154.

Schabenberger, O., & Gotway, C. A. (2017). *Statistical methods for spatial data analysis*. CRC Press.

Settles, B. (2012). Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *6*, 1–114.

Stein, M. L. (2012). *Interpolation of spatial data: Some theory for kriging*. Springer Science & Business Media.

Stevens, D. L., Jr., & Olsen, A. R. (2004). Spatially balanced sampling of natural resources. *Journal of the American Statistical Association*, *99*, 262–278.

Weaver, B. P., Williams, B. J., Anderson-Cook, C. M., & Higdon, D. M. (2016). Computational enhancements to Bayesian design of experiments using Gaussian processes. *Bayesian Analysis*, *11*, 191–213.

How to cite this article: Paglia, J., Eidsvik, J., & Karvanen, J. (2021). Efficient spatial designs using Hausdorff distances and Bayesian optimization. *Scandinavian Journal of Statistics*, 1–25. https://doi.org/10.1111/sjos.12554