Simulation-regression approximations for value of information analysis of geophysical data

Jo Eidsvik[1], **Geetartha Dutta**[2], **Tapan Mukerji**[2,3] **and Debarun Bhattacharjya**[4]

1) Department of Mathematical Sciences, NTNU, Norway

2) Department of Energy Resources Engineering, Stanford University, California

3) Department of Geophysics, Stanford University, California

4) Cognitive Computing Research, IBM T. J. Watson Research Center, New York

Corresponding author: Jo Eidsvik (joeid@math.ntnu.no), Department of Mathematical Sciences, NTNU, 7491 Trondheim, NORWAY

Running title: Approximations for VOI analysis

Keywords: value of information, decision analysis, geophysical data, spatial decision situations, simulation and regression, petroleum geostatistics

Abstract

Value of information analysis is useful for helping a decision maker evaluate the benefits of acquiring or processing additional data. Such analysis is particularly beneficial in the petroleum industry, where information gathering is costly and time-consuming. Furthermore, there are often abundant opportunities for discovering creative information gathering schemes, involving the type and location of geophysical measurements. A consistent evaluation of such data requires spatial modeling that realistically captures the various aspects of the decision situation: the uncertain reservoir variables, the alternatives and the geophysical data under consideration. The computational tasks of value of information analysis can be daunting in such spatial decision situations; in this paper, a regression-based approximation approach is presented. The approach involves Monte Carlo simulation of data followed by linear regression to fit the conditional expectation expression that is needed for value of information analysis. Efficient approximations allow practical value of information analysis for the spatial decision situations that are typically encountered in petroleum reservoir evaluation. Applications are presented for seismic amplitude data and electromagnetic resistivity data, where one example includes multi-phase fluid flow simulations.

Keywords: value of information, decision analysis, geophysical data, spatial decision situations, simulation and regression, petroleum geostatistics

1 Introduction

Petroleum reservoir decisions are characterized by large development and production costs as well as potentially huge revenues associated with oil or gas production. Making good reservoir decisions can be challenging due to notable uncertainty in our knowledge of the subsurface and the future costs and prices. In most situations, it may be valuable to gather data to learn more about the uncertain variables. Geophysical information can lead to an improved understanding of the subsurface distribution of reservoir variables, but this data usually comes at a considerable price. The decision-theoretic notion of value of information (VOI) allows decision makers to evaluate various sources of data and determine whether the information is worthwhile before it is purchased. When the decision maker can represent the value from the underlying decision situation in monetary units, then VOI is also in monetary units. This may have tremendous practical ramifications since the decision maker can then determine whether a particular dataset should be purchased depending on whether the VOI exceeds the price of the data.

The focus of this paper is on methods for conducting VOI analysis of geophysical data for oil and gas applications; specifically, the paper considers the potential purchase of seismic and electromagnetic data for making better drilling decisions. VOI is used to quantify the additional value of various information gathering opportunities. Compared with the traditional use of VOI in the petroleum industry (Bratvold et al., 2009), the current paper focuses on models that stress the spatial multivariate aspects of the uncertain reservoir variables, the alternatives, and the potential information gathering schemes. In most reservoir decision situations there is often a wealth of opportunities for information gathering that involves, for instance, a choice between different types of data (production tests, seismic data, electromagnetic data, etc.), determining locations of data acquisition, etc. The value of multivariate spatial data can be evaluated by consistently integrating them in models that incorporate spatial characteristics (Eidsvik et al., 2015).

VOI analysis for complex multivariate spatial models can however be computationally intensive. In this paper, approximate VOI computations are presented, using statistical regression techniques along with Monte Carlo simulations for applications involving geophysical data. While similar methods have been applied in medical applications (Strong et al., 2014), here the spatial aspects that are typical in decision situations in the earth sciences are stressed, illustrating how the approximation equations work in this setting. Results obtained by simulation and linear regression are compared with fully analytical solutions wherever possible, such as for linearized Gaussian models along with some working assumptions pertaining to the underlying decision situation (Bhattacharjya et al., 2013).

The remainder of the paper is organized as follows. Section 2 provides the notation, assumptions and methods for VOI analysis in spatial decision making. Section 3 outlines the computational challenges of VOI analysis, and presents simulation - regression methods for efficient approximation of the VOI. Section 4 shows two applications of VOI analysis of geophysical data that deploy the proposed approximate approaches. Concluding remarks are in Section 5.

2 Background on Decision Analysis and VOI Analysis

Decision analysis applies the principles of decision theory to practical decision situations with the eventual goal of providing the decision maker with clarity of action (Howard and Abbas, 2015). The field has a rich history of effective application in the oil and gas industry, see Bratvold and Begg (2010) and Newendorp and Schuyler (2013), and VOI in particular has raised a lot of interest (Bratvold et al., 2009).

2.1 Prior value

This presentation starts by the notation, theory and equations for VOI analysis (Table 1). In reservoir applications, the uncertainties, alternatives and data tend to be vector-valued. The alternatives are denoted $\boldsymbol{a} = \{a_j : j = 1, 2, ..., N\}$, where entries a_j are associated with different spatial locations. The spatial locations could for instance represent reservoir segments, or they could indicate sets of injection and production wells for reservoir depletion. The decision maker must choose alternatives from an available set $\boldsymbol{a} \in \boldsymbol{A}$. The uncertain spatial variables are denoted $\boldsymbol{x} = \{x_i : i = 1, 2, ..., n\}$, where index i is associated with spatial location. When $N = n$, there are identical spatial modeling scales for the uncertain variable and the alternatives, but this need not be the case in general. From a computational perspective, a large set of alternatives can be difficult to handle. This is exemplified by the decision of whether to drill or not at $N = n$ reservoir segments, yielding 2^n alternatives in total. In situations with fewer alternatives, one can enumerate the alternatives more easily. Such a case is illustrated by the decision of choosing among a few drilling plans.

Denote by $v(\boldsymbol{x}, \boldsymbol{a})$ the value obtained by the decision maker. The explicit functional form clarifies that the value depends on the uncertain variable $\boldsymbol{x}$ as well as the chosen alternative $\boldsymbol{a}$. Assuming a risk-neutral decision maker who chooses alternatives based on optimal expected value, the prior value (PV) is:

$$\text{PV} = \max_{\boldsymbol{a} \in \boldsymbol{A}} \left\{ E[v(\boldsymbol{x}, \boldsymbol{a})] \right\}. \tag{1}$$

The expectation in Eq. (1) is over the uncertain reservoir variable $\boldsymbol{x}$. Its probability distribution is defined by prior probability density function $p(\boldsymbol{x})$ for a continuous sample space, and a probability mass function $p(\boldsymbol{x})$ for a discrete sample space. The prior distribution incorporates all currently available information and the expected values with respect to this variable take the form of integrals or sums. For a more general discussion about utility functions that represent risk-averse or risk seeking behavior and related VOI analysis, the reader should consult the decision analysis literature, for instance Howard (1966) and Howard and Abbas (2015).

A distinction is made between coupled and decoupled value. When there is coupling, the value computation depends simultaneously on several of the spatial elements and cannot be split into different components. An example of such a model is where fluid flow simulation is used to evaluate recoverable oil (and hence ultimate value), in which case the value function involves complex coupling of reservoir properties and the chosen drilling alternatives. When the value function decouples, it can be split into a sum over different spatial elements. The situation is exemplified by drilling choices at different non-communicating reservoir segments where the optimization of expected value can be done for individual segments. For this decoupled situation with $N = n$, the

PV expression can be simplified to:

$$\text{PV} = \sum_{i=1}^{n} \max_{a_i} \left\{ E \left[v_i \left(x_i, a_i \right) \right] \right\}, \tag{2}$$

where v_i is the value function at location i.

To illustrate the basic ideas, let us consider a simple situation where one must decide whether to drill at a petroleum prospect ($n = 1$) with uncertain outcome $x \in \{0, 1\}$. The probability of a wet well (discovery) is assumed to be $p(x = 1) = 0.1$ and therefore the probability of a dry well is $p(x = 0) = 0.9$. In this situation, the decision maker has two alternatives: $a = 0$ (do not drill) and $a = 1$ (drill). The decision maker obtains no value if she does not drill; $v(x, a = 0) = 0$, while the value depends on the uncertain outcome if the well is drilled. For this example, assume $v(x = 0, a = 1) = -0.5$ (because there is a cost incurred for drilling) and $v(x = 1, a = 1) = 3 - 0.5 = 2.5$. (All values are in generic monetary units.) A decision tree for this situation is shown in Fig. 1 (top left). Here, squares represent decisions while circles represent uncertainties. A priori, the expected value and the PV of the drilling alternative are

$$E \left[v \left(x, a = 1 \right) \right] = -0.5 \cdot 0.9 + 2.5 \cdot 0.1 = 3 \cdot 0.1 - 0.5 = -0.2, \tag{3}$$

$$\text{PV} = \max \left\{ E \left[v \left(x, a = 0 \right) \right], E \left[v \left(x, a = 1 \right) \right] \right\} = \max \{0, -0.2\} = 0. \tag{4}$$

A geoscientist would assert that decision situations are often more complex than the simple univariate problem described above. Fig. 2 illustrates an example that extends the setting to involve two prospects. Note that this is still highly simplified, but it demonstrates some of the additional aspects that must be considered for multivariate and spatial analysis. The statistical model is defined by a Bayesian network (Darwiche, 2009) consisting of five nodes (Fig. 2). The top node is a common parent of the outcome at both prospects, inducing dependence in the reservoir outcomes as a result of common geological mechanisms for the two prospects. The parent node would indicate the presence or absence of a source or a petroleum system. The edges from the common parent node to the prospect nodes represent conditional probabilities. Here, there is probability one of propagating a dry outcome; $p(x_i = 0 | x_0 = 0) = 1$, $i = 1, 2$. Furthermore, there is a positive probability of failure in the propagation of a success, $p(x_i = 1 | x_0 = 1) = 0.5$, $i = 1, 2$. This is natural for some geological modeling situations, for instance where a hydrocarbon source migrates from a source rock with a chance of failure in the migration (Martinelli et al. (2011) and Lilleborge et al. (2016)). The bivariate probability mass function for prospect 1 and 2, when marginalizing over the common parent node, is provided in Table 2.

The alternatives that are available in this two-prospect situation depend on the particulars of the multivariate decision situation. In this case, an assumption of decoupling value function (Eq. (2)) is made, and the decision maker is free to select as many prospects as desired, therefore there are four available alternatives: choose none, choose only Prospect 1, choose only Prospect 2, or choose both. Note that one could potentially have opportunities for sequential selection in a situation like this one, where the decision maker is allowed to choose one prospect, observe the result, and then decide whether to choose the next prospect (Bickel and Smith (2006) and Martinelli et al. (2014)). Even though the VOI framework also works in such a context with sequential selection, the solution requires dynamic programming and it is computationally more difficult to solve for large-size models. Here, the focus is solely on static selection.

The revenue from each prospect is specified as 3 money units when it is wet (hydrocarbon discovery), while the cost of development is varied. There are no shared costs. The cost is incurred no matter if the prospect is wet or dry. (Note that it was set to 0.5 in the single-prospect analysis above.) The PV can be calculated as

$$\text{PV} = \sum_{i=1}^{2} \max_{a_i \in \{0,1\}} \left\{ \sum_{x_i=0}^{1} v_i\left(x_i, a_i\right) p\left(x_i\right) \right\} = 2 \cdot \max\left\{0, (3 \cdot 0.1 - \text{Cost})\right\}. \tag{5}$$

It is optimal to select both prospects if Cost < 0.3. As a function of cost, the PV decreases with slope 2 from 0.6 at zero cost, and becomes 0 for Cost ≥ 0.3.

2.2 Posterior value and VOI

Suppose next that the decision maker can purchase data before making decisions. The data are modeled by a likelihood model $p(\boldsymbol{y}|\boldsymbol{x})$ connecting the uncertain variables to the information $\boldsymbol{y}$. This likelihood model of course also depends on the experimental design or the data gathering scheme. The posterior value (PoV) is

$$\mathrm{PoV}(\boldsymbol{y}) = \sum_{\boldsymbol{y}} \max_{\boldsymbol{a} \in \boldsymbol{A}} \left\{ E\left[v\left(\boldsymbol{x}, \boldsymbol{a}\right) | \boldsymbol{y} \right] \right\} p\left(\boldsymbol{y}\right), \tag{6}$$

assuming a discrete sample space for the data. Compared with the PV in Eq. (1) the choice among the alternatives can now vary depending on the outcomes of data. This means that the PoV is larger or equal to the PV - the decision maker can choose more wisely after getting more information. When the value function decouples over the $n = N$ locations, the PoV becomes

$$\mathrm{PoV}(\boldsymbol{y}) = \sum_{\boldsymbol{y}} \sum_{i=1}^{n} \max_{a_i} \left\{ E\left[v_i\left(x_i, a_i\right) | \boldsymbol{y} \right] \right\} p\left(\boldsymbol{y}\right). \tag{7}$$

For a risk-neutral decision maker the VOI is the difference between posterior and prior value:

$$\mathrm{VOI}(\boldsymbol{y}) = \mathrm{PoV}(\boldsymbol{y}) - \mathrm{PV}. \tag{8}$$

The VOI should be compared with the price of the data to be acquired and processed. If the VOI exceeds the price, it is worthwhile to purchase the data. The decision maker can use the VOI results of different tests to evaluate various opportunities for information gathering, and this analysis is done before obtaining the data.

Consider again the situation with one prospect. Suppose the decision maker can perform tests to get information about the uncertain outcome at the prospect. A test can be considered imperfect (measured with noise or without direct relation) or perfect (no noise and direct relation). The calculation of PoV with perfect information is shown in Fig. 1 (bottom left); this decision tree represents the situation where the decision maker has the opportunity to observe the prospect outcome before deciding whether to drill. Perfect information about a discovery would induce the decision maker to drill the well, while she would choose not to drill with dry perfect information. The PoV is calculated as

$$\mathrm{PoV}(x) = \sum_{x=0}^{1} \max\left\{0, v\left(x, a=1\right)\right\} p\left(x\right) = 0 \cdot 0.9 + 2.5 \cdot 0.1 = 0.25. \tag{9}$$

By direct use of Eq. (8), and using that PV $= 0$, the VOI $(x) = 0.25 - 0 = 0.25$.

This value of perfect information forms an upper bound for the value of any imperfect test. Suppose the decision maker can gather imperfect data, say a seismic test, modeled using a likelihood model $p\left(y|x\right)$. For this case, consider a seismic test with binary outcome; $y \in \{0, 1\}$ where a correct negative seismic test occurs with likelihood $p(y=0|x=0)=0.9$ and a correct positive seismic test has likelihood $p(y=1|x=1)=0.9$. From Eq. (7), the decision maker now chooses the optimal alternatives based on the conditional expected value, given the data realization y. The values are averaged over all possible data using probabilities $p(y)$ (see Fig. 1 right). Here, the optimal decision is to drill for a positive test, and not drill for a negative test. The PoV with this imperfect information is

$$\begin{aligned}\text{PoV}(y) &= \sum_y \max_a \{E[v(x,a)|y]\} p(y) \\ &= \max\{0,-0.47\}\,0.82+\max\{0,1\}\,0.18=0.18. \end{aligned} \tag{10}$$

The seismic test has VOI $(y)=0.18-0=0.18$, and if the price of the seismic test is less than this, the decision maker should purchase the test.

The extension to the situation with two prospects entails more testing options; both (total), only one (partial), perfect (drilling) or imperfect (seismic). Note that one could also perform sequential testing, where one prospect is tested first, with the option of continued testing at the other. This is computationally more demanding for large-size models, and only static testing is considered here. The value of perfect total information is

$$\text{PoV}(\boldsymbol{x}) = 2 \cdot 0.1 \cdot \max\{0,(3-\text{Cost})\}. \tag{11}$$

This PoV of perfect information starts at 0.6 for zero cost and decreases with slope 0.2 until it reaches 0 at cost equal to 3. The bottom nodes in Fig. 2 represent the results of seismic tests at both prospects, which would give imperfect information. These test results are assumed to be conditionally independent given the outcome of the prospect. The likelihood model is defined by $\gamma = p(y_i=k|x_i=k)$, $k=0,1$, and for prospects $i=1,2$. This means that the chance of a positive test for a wet reservoir equals that of a negative test for a dry reservoir at both prospects.

Fig. 3 shows the VOI for different information gathering schemes in this example. The results for partial testing (at just one prospect, univariate data) are shown in the top display, while those for total testing (at both prospects; bivariate data) are at the bottom. In both displays $\gamma=0.9$, and the VOI is plotted as a function of the drilling cost at a prospect. Note that the VOI curves all start near 0 for very small development cost where the decision maker would surely drill prospects and there is no need for data. The curves then increase for intermediate costs where data are likely to influence drilling decisions, and decrease back to small values for large costs where the decision maker would surely avoid drilling because it is too expensive; again, in this region, there is little help from additional data. In both displays, the VOI of imperfect information is less than that of perfect information (the dashed line is below the solid line), but note that the VOI of imperfect information at both prospects could be higher or lower than the VOI of perfect information at one prospect.

For perfect information at both prospects (denoted by the solid line in the bottom display) the VOI $= 0.5$ for cost 0.5. This is twice of what was observed in the single-prospect case in Eq. (9). For partial perfect information (the solid line in the top display) the corresponding VOI $= 0.35$, which is still larger than for the single prospect. This increased VOI is due to the correlation in the model; observing that Prospect 1 is a discovery has a clear influence on Prospect 2. Here, the conditional probability is $p\left(x_2 = 1|x_1 = 1\right) = 0.5$, and the positive observation at Prospect 1 encourages development at Prospect 2. The piecewise linear curves in Fig. 3 can be understood by relating the conditional probabilities to costs; a change in the slope implies two different cost regimes, where the conditional expectations exceed the cost for the various data outcomes (different positive elements for the sum over data in Eq. (10)).

2.3 A workflow for spatial VOI analysis

By embedding VOI analysis in a unified framework involving spatial uncertainties and spatial alternatives, one can address and evaluate a range of spatial information gathering schemes with high-fidelity models that realistically capture the various aspects of the decision situation. Eidsvik et al. (2015) recommend a workflow involving the following steps:

1. Frame the spatial decision situation.

2. Study potential spatial information gathering schemes.

3. Build a spatial model for variables and data.

4. Conduct VOI analysis.

Step 1. in the workflow is for the decision maker to frame their decision situation by identifying the decisions (and associated alternatives) as well the most pertinent uncertainties.

For step 2. in the workflow: In spatial situations, there are a number of possibilities, including acquiring or processing data over the entire spatial domain of interest (total test) as well as performing partial testing, that is only over a subset of the spatial domain. Modeling the accuracy of the test is critical for VOI analysis. With perfect information, either $\boldsymbol{y} = \boldsymbol{x}$ (total information) or $\boldsymbol{x}_{\mathbb{K}}$ (partial information), where $\mathbb{K}$ denotes a subset of the set $i = 1, \ldots, n$. (In the prospects example this subset was only one prospect of two possible prospects.) In many reservoir studies companies must also evaluate various data sources that could help make better drilling decisions. The value of gathering say seismic and / or electromagnetic data depends on their likelihood as well as the prior model for reservoir variables and the alternatives.

The models for the uncertain resevoir variables and the data variables should include realistic spatial dependencies (step 3. in the workflow). In the application in Sect. 4.1, a Gaussian random fields is used for the reservoir profits, having a linear likelihood model with conditional independence between the selected measurement locations. In Sect. 4.2, multiple reservoir variables are built from a suite of geostatistical models, and the value function computation involves a fluid flow simulator calculating the amount of recoverable oil for each sample of the uncertain outcomes. The data are modeled by seismic forward models.

Step 4. of the workflow involves the actual computation of the VOI and the practical interpretation of results. Decision makers' interests often lie in comparing the difference between the VOI of different acquisition or processing schemes for spatial data with their price. It is further useful to study the sensitivity of the VOI to different designs or to specific input parameters such as the spatial heterogeneity or the variability in the prior model or the likelihood.

3 Approximations for VOI

VOI computation is often demanding and therefore there is a need to develop efficient approximations for real-world models; various approaches for computing and approximating the VOI are presented in the following sections.

3.1 Review of methods

The PoVs (Eqs. (6) or (7)) are calculated from maximum expected values conditional on different data outcomes (referred to as the inner sum or integral), and then weighted over all possible data (referred to as the outer sum or integral). There are rarely closed form solutions for the inner expression $E\left[v\left(\boldsymbol{x},\boldsymbol{a}\right)|\boldsymbol{y}\right]$ in Eq. (6). Moreover, the integration of these maximum conditional values with the marginal probability of the data in the outer expression is even less likely to be available in closed form.

One generally applicable computational approach for the PoV (and the VOI) is that of double Monte Carlo sampling. This entails using Monte Carlo sampling of data $\boldsymbol{y}$ for the outer expression, and another round of Monte Carlo sampling (for each data realization) for the inner expression. But this is extremely time-consuming and several approaches have been suggested in Earth science applications to overcome such a costly double Monte Carlo approach.

Trainor-Guitton et al. (2011) and Trainor-Guitton et al. (2013) suggested a dimension reduction of the uncertain variable that effectively discretizes the sample space for the uncertain variables. Similarly, the data variables are inverted to the same scale and PoV is available as a sum over few terms. Trainor-Guitton et al. (2014) extended this approach and applied it to a geo-thermal application with uncertain clay-cap throat, where the optimal drilling decisions (location) depend on the magnetotellurics data. Albeit clearly useful in practice, in particular when the main uncertainty is related to geologic scenario or special geographic structures, such a discretization might be unnatural if the decision maker has multi-variable spatial alternatives available.

Eidsvik et al. (2008), Bhattacharjya et al. (2010) and Rezaie et al. (2014) used parametric models and associated methods to analytically calculate the conditional expectation $E\left[v_i\left(x_i,a_i\right)|\boldsymbol{y}\right]$, given the data. These approaches resulted in a closed form solution for the inner expression in Eq. (7), while the outer expression was solved by Monte Carlo sampling.

For Gaussian linear models, the outer integral over the data can also be computed exactly for decisions with two alternatives, and there are analytical results for the PoV and VOI. The key result is that for a Gaussian variable x with mean m and variance r^2:

$$E\left(\max\left\{0,x\right\}\right)=\int_0^\infty xp\left(x\right)dx=m\Phi\left(m/r\right)+r\phi\left(m/r\right), \tag{12}$$

where the density is $\phi\left(z\right)=\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$, and the cumulative function is $\Phi(z)=\int_{-\infty}^{z}\phi(x)dx$ (Bickel, 2008).

Bhattacharjya et al. (2013) extended this result for multivariate models and spatial decision situations with two alternatives at every spatial location. To illustrate, assume a multivariate Gaussian prior distribution for the values, say, $v_i\left(x_i,a_i=1\right)=x_i$, $i=1,\dots,n$;

$$p\left(\boldsymbol{x}\right)=N\left(\boldsymbol{\mu},\boldsymbol{\Sigma}\right)=\frac{1}{(2\pi)^{n/2}}|\boldsymbol{\Sigma}|^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)\right). \tag{13}$$

Assume further a linear and Gaussian likelihood model $\boldsymbol{y}=\boldsymbol{F}\boldsymbol{x}+\boldsymbol{e}$, where $p\left(\boldsymbol{e}\right)=N\left(\underline{0},\boldsymbol{T}\right)$, and $\boldsymbol{x}$ and $\boldsymbol{e}$ are independent. Then $p\left(\boldsymbol{y}\right)=N\left(\boldsymbol{F}\boldsymbol{\mu},\boldsymbol{F}\boldsymbol{\Sigma}\boldsymbol{F}^T+\boldsymbol{T}\right)$, and the conditional distribution for the values is Gaussian with mean and covariance matrix:

$$\boldsymbol{\mu}_{x|y}=\boldsymbol{\mu}+\boldsymbol{\Sigma}\boldsymbol{F}^T\left(\boldsymbol{F}\boldsymbol{\Sigma}\boldsymbol{F}^T+\boldsymbol{T}\right)^{-1}\left(\boldsymbol{y}-\boldsymbol{F}\boldsymbol{\mu}\right),\ \ \boldsymbol{\Sigma}_{x|y}=\boldsymbol{\Sigma}-\boldsymbol{\Sigma}\boldsymbol{F}^T\left(\boldsymbol{F}\boldsymbol{\Sigma}\boldsymbol{F}^T+\boldsymbol{T}\right)^{-1}\boldsymbol{F}\boldsymbol{\Sigma}. \tag{14}$$

The posterior value can be computed as

$$\text{PoV}\left(\boldsymbol{y}\right)=\sum_{i=1}^{n}\int\max\left\{0,E\left(x_i|\boldsymbol{y}\right)\right\}p\left(\boldsymbol{y}\right)d\boldsymbol{y}=\sum_{i=1}^{n}\left[\mu_i\Phi\left(\mu_i/r_i\right)+r_i\phi\left(\mu_i/r_i\right)\right], \tag{15}$$

where r_i^2 are the diagonal elements of the reduction in covariance $\boldsymbol{\Sigma}\boldsymbol{F}^T\left(\boldsymbol{F}\boldsymbol{\Sigma}\boldsymbol{F}^T+\boldsymbol{T}\right)^{-1}\boldsymbol{F}\boldsymbol{\Sigma}$.

Although these results are extremely fast to compute and can provide useful VOI results in many application domains (Sect. 4.1), they do rely heavily on the Gaussian linear assumptions. Such an approach should therefore only be used when these assumptions are reasonable.

3.2 Simulation-regression approach

In this section, a simulation-regression approach is described for approximating the PoV and hence the VOI in spatial decision situations (Eqs. (6) or (7)). This approach is based on Monte Carlo sampling of data (outer sum or integral) and regression fitting of conditional expected values (inner sum or integral). Strong et al. (2014) and Heath et al. (2016) present similar simulation-regression approaches for VOI analysis in other contexts. One goal of the current paper is to demonstrate that these approaches are also very useful for petroleum geostatistics applications. The spatial decision making context introduces new elements such as spatial models for the uncertainties, the coupling or decoupling of potentially complex value functions and the potential combinatorial proliferation of alternatives.

The Monte Carlo and regression technique builds on the structure of Fig. 4, illustrating the connection between decisions, uncertain variables of interest, values and data in the form of influence diagrams (Howard and Matheson, 1984). These graphical representations are extensions of Bayesian networks; they represent a decision maker's decision situation. The original diagram (left) is as specified by the model. The approximation converts it to the diagram where the distribution over data $\boldsymbol{y}$ is represented by Monte Carlo samples and the new value function is based on the regression model (right). For VOI analysis the influence of data on the conditional expected value is critical. The idea of regressing value on the data is in some sense similar to the prediction-focused view of Scheidt et al. (2015), but they phrase this aspect only for faster prediction and not in a decision analytic framework.

The simulation-regression approach starts by simulating B variables $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^B$ from the prior $p(\boldsymbol{x})$. For each alternative $\boldsymbol{a} \in \boldsymbol{A}$ and for each realization of variables, value samples $v_a^b = v\left(\boldsymbol{x}^b, \boldsymbol{a}\right)$ are generated, as well as data samples $\boldsymbol{y}^b$ using the likelihood model $p\left(\boldsymbol{y}^b | \boldsymbol{x}^b\right)$. Next, a regression model is used to fit $E\left[v\left(\boldsymbol{x}, \boldsymbol{a}\right) | \boldsymbol{y}\right]$ based on the samples $\left(\boldsymbol{y}^b, v_a^b\right)$, $b = 1, \ldots, B$. The focus here is on fitting a functional parametric form $g_{\boldsymbol{a}}\left(\boldsymbol{y}; \boldsymbol{\beta}_{\boldsymbol{a}}\right)$ to the conditional expected value. The posterior value is then approximated by

$$\text{PoV}\left(\boldsymbol{y}\right) \approx \frac{1}{B} \sum_{b=1}^{B} \max_{\boldsymbol{a} \in \boldsymbol{A}} \left\{ g_{\boldsymbol{a}}\left(\boldsymbol{y}^b; \hat{\boldsymbol{\beta}}_{\boldsymbol{a}}\right) \right\}. \tag{16}$$

To ensure that the VOI is always non-negative, it is considered good practice to consistently use the approximation for the PV as well. This involves a double expectation:

$$E_{\boldsymbol{x}}\left[v\left(\boldsymbol{x}, \boldsymbol{a}\right)\right] = E_{\boldsymbol{y}}\left(E_{\boldsymbol{x}}\left[v\left(\boldsymbol{x}, \boldsymbol{a}\right) | \boldsymbol{y}\right]\right).$$

The VOI is then approximated by the key formula:

$$\text{VOI}\left(\boldsymbol{y}\right) \approx \frac{1}{B} \sum_{b=1}^{B} \max_{\boldsymbol{a} \in \boldsymbol{A}} \left\{ g_{\boldsymbol{a}}\left(\boldsymbol{y}^b; \hat{\boldsymbol{\beta}}_{\boldsymbol{a}}\right) \right\} - \max_{\boldsymbol{a} \in \boldsymbol{A}} \left\{ \frac{1}{B} \sum_{b=1}^{B} g_{\boldsymbol{a}}\left(\boldsymbol{y}^b; \hat{\boldsymbol{\beta}}_{\boldsymbol{a}}\right) \right\}. \tag{17}$$

The approximation improves for larger sample sizes B but it also depends on the selection of model $g_{\boldsymbol{a}}\left(\boldsymbol{y}^b;\boldsymbol{\beta}_{\boldsymbol{a}}\right)$. The choice would be very case specific. Strong et al. (2014) recommend using generalized additive models or Gaussian processes. In the current paper linear regression approaches are demonstrated such that $g_{\boldsymbol{a}}\left(\boldsymbol{y}^b;\boldsymbol{\beta}_{\boldsymbol{a}}\right) = \beta_{0,\boldsymbol{a}} + \boldsymbol{\beta}_{1,\boldsymbol{a}}^T\boldsymbol{y}$, for the applications in Sect. 4. For high dimensional data values are regressed on only the most important components of the data. In one of the applications, Sect. 4.2, partial least squares regression (PLSR) is used for this purpose. More complex regression models are certainly interesting, such as non-parametric regression techniques, especially when the value function is complex and with non-linear likelihood models, but this is left out for further work. A comparison of value approximations using different regression methods for medical decisions involving only a few alternatives is provided in Heath et al. (2015).

Let us illustrate the simulation-regression approach for the prospect example in Sect. 2. First, simulate variables x_i^b, $i = 1, 2$ and $b = 1, \ldots, B$, and data $p\left(y_i^b|x_i^b\right)$. Next, set values: $v_{a_i=0}^b = 0$, and $v_{a_i=1}^b = 3I\left(x_i^b = 1\right) - \text{Cost}$. With such categorical data, it is natural to fit the conditional expectation from sample averages of values in the different groups of data. Data are binary and groups; $\left(y_1^b, y_2^b\right)$ could be $(0,0), (0,1), (1,0), (1,1)$.

If the spatial decision situation involves a decoupling of values, the regression must be done for each of the locations. This entails data and values $\left(\boldsymbol{y}^b, v_{a_i}^b\right)$, $b = 1, \ldots, B$, $i = 1, \ldots, n$. In calculations where there is one deterministic input and one uncertain input, such as $\max\left\{0, v\left(x_i, a_i\right)\right\}$, an upward bias might occur while approximating the VOI. For instance, when using linear regression, the line could fit insignificant effects and induce a falsely positive element in the PoV, but because of the $\max$ expression there will never be falsely negative elements. The simulation-regression approach will nevertheless converge to the best linear approximate solution when $B \to \infty$, but for small sample sizes, it could be useful to reduce the bias by removing insignificant effects in the linear regression. These effects are greatest in situations with many decision locations and relatively high measurement noise levels.

In this paper, bootstrapping is suggested to screen insignificant effects in the PoV. This is done for the value at one site at a time. It is illustrated for one site and decoupled value:

1. Repeat for C bootstrap replicates:
   - Draw a sample of size B with replacement from the indices $b* = 1, \ldots, B$.
   - Fit a regression line from samples $\left(\boldsymbol{y}^{b*}, v_{a_i}^{b*}\right)$, $b* = 1, \ldots, B$.
2. Sort the C bootstrap replicates to construct a $(1-\alpha)$ coverage interval for the fitted value $\hat{v}_{a_i}\left(\boldsymbol{y}^b\right)$ at each $\boldsymbol{y}^b$, where an $\alpha/2$ fraction of the values are cut at both the lower and higher end. If this interval covers the overall prior mean $\bar{v}_{a_i}$, screen this fitted value to be insignificant; $\hat{v}_{a_i}\left(\boldsymbol{y}^b\right) = \bar{v}_{a_i}$.

This bootstrap screening approach seems to clearly reduce bias in the test runs, but the bias is not removed entirely. Localized approximations were also tested, where values at sites far from the data are set identical in the prior and posterior value calculation. This procedure seems reasonable in the spatial context, where one can tune the influence range from the spatial correlation in some cases.

Note that although the regression approximations simplify the computation of conditional expected values for each alternative, this must still be done for all alternatives $\boldsymbol{a} \in \boldsymbol{A}$. This issue may be addressed at least partially by reducing the number of alternatives explored by clustering them - this is not considered in the current paper.

4 Applications

The proposed methods are illustrated on two applications of VOI analysis involving geophysical data. The description follows the workflow described in Sect. 2.3.

4.1 Electromagnetic resistivity data

Consider a two-dimensional lateral domain covering the top of a petroleum reservoir that has been mapped out from geological understanding, and where there is some prior knowledge about the reservoir variables from interpreted seismic data (Avseth et al., 2005). The prediction of reservoir variables such as facies, saturation and porosity at the lateral reservoir segments is indicative of the profits and therefore critical for reliable decision making. (See also Chapter 6.3 in Eidsvik et al. (2015)).

A grid of 25×25 reservoir segments is defined at this top-reservoir zone. The decision maker can freely select to drill or not at each of the $N = 625$ segments. Then the alternatives are $a_i \in \{0, 1\}$, $i = 1, \ldots, 625$. This means there are 2^{625} alternatives. The value function is assumed to be the sum over segment values, so it decouples. If the decision maker decides to drill a well at a particular cell i, the uncertain value is $v(x_i, a_i = 1)$, where x_i represents the uncertain reservoir properties at segment i. The decision maker gets 0 value if there is no drilling at a segment. The particular form of the value function is discussed further below.

Since extensive three-dimensional seismic interpretation has been done for this reservoir zone, several reservoir variables such as porosity and thickness are quite accurately determined. However, there is remaining uncertainty about the saturation and hence the profits. It could be worthwhile for the petroleum company to acquire and process electromagnetic (EM) data at this stage. Hydrocarbon saturated rocks have higher resistivity than brine saturated reservoir rocks, and these EM data would give imperfect information about the reservoir saturation and hence the profits, leading to improved decision making. The VOI of EM resistivity data depends on the acquisition design, rock and fluid properties, and the data quality compared with the prior understanding of the reservoir profits. North-south survey lines are considered here for the EM data. Three possible lines are considered; west, center and east (Fig. 5). All three tests thus provide partial information; specifically they pertain to only 25 out of the 625 segments.

The prior model is constructed from samples of porosity and saturation obtained from a probabilistic inversion method given the currently available seismic amplitude data. The uncertain revenue at location i is $r_0 h_i s_i \phi_i$, where r_0 is a fixed factor including oil price and recovery rates, h_i is the reservoir thickness, s_i the oil saturation and ϕ_i is the porosity. Based on samples, the variability in revenues largely depends on the sampling variability in saturation. From the samples a Gaussian prior is fit directly for profits $v_i(x_i, a_i = 1)$, where the cost of drilling is subtracted from the revenues. The means and variances are spatially varying and there is spatial correlation. The prior means of the profits are shown in Fig. 5.

The likelihood model for EM resistivity data is based on Archie's relation (Mavko et al., 2009) that relates the rock bulk resistivity to the fluid saturation and fluid resistivity. The likelihood for log-resistivity data $\boldsymbol{y} = (y_1, \ldots, y_{25})$ is specified by having a linearized mean as a function of uncertain saturation (and hence as a function of the profits, since the other parameters are specified from the seismic data). The slope and intercepts of the linearized model still depend on the porosity properties at the segments but they are fixed from the seismic inversion. This linearization of Archie's relation is an approximation which seems to work rather well when the linearization is performed at the current estimates of porosity. The likelihood model has conditionally independent errors.

For these linear Gaussian modeling assumptions, the analytical result in Sect. 3.1 holds (Eq. (15)). The analytical solutions can be compared with the approximations obtained by Monte Carlo simulation and regression techniques.

Since there is a decoupled value function, the regression of value is done for each cell, individually, on the data. For each segment $i = 1, \ldots, 625$, samples $(v_{a_i}^b, \boldsymbol{y}^b)$, $b = 1, \ldots, B$ are used to fit a linear regression model for the conditional expectation $\hat{E}\left(v_{a_i}|\boldsymbol{y}^b\right)$. As mentioned in Sect. 3.2, a straightforward linear regression approach tends to overestimate the PoV in this situation, therefore bootstrap screening is used to reduce the bias. Fig. 6 illustrates this bootstrap idea by showing the results of the simulation-regression procedure for values at sites near (top) and far (bottom) from the data locations (west line). For illustration, the plot is only against the first component of the data (south-west data location). The displays hence show fits of $g_{a_i=1}\left(\boldsymbol{y}; \boldsymbol{\beta}_{a_i=1}\right) = \beta_{0,i} + \beta_{1,i} y_1$ (displayed as a function of data outcome y_1). The near site (1,1) and far site (1,23) are indicated in Fig. 5. The left displays show $B = 50$ samples and the fitted regression lines. There is a connection between data and value for the near site, while the far site is not very influenced by the data. By using bootstrap replicates of the data and values one can study the uncertainty of the linear regression. The variety in fitted lines (right displays) shows that the near site (top) is significantly affected by data, but not the far site (bottom), which has posterior value just like the prior value (0 in this case).

Table 3 shows the VOI results for three possible north-south acquisition lines (Fig. 5), for two different measurement accuracies. The prior value is the sum over all cells with non-negative expected profit, without any EM information. For this example the prior value is 410 million. The VOI is the expected additional value obtained by an EM survey. The 80 percent coverage intervals for the VOI were computed by using 50 repetitions of VOI approximations, sorting the VOI results from smallest to largest, and picking the 10th and 90th percentiles.

The MC+Gauss approach generates data samples from the Gaussian marginal distribution of the EM resistivity data $\boldsymbol{y}^b$, $b = 1, \ldots, B$. For each sample, the conditional expectation is computed by the closed form Gaussian expression (Eq. (14)). The non-zero expected values are then averaged over all data samples. This approach is unbiased and converges to the exact analytical solution for the posterior value when the sample size increases. Note that this approach is only applicable since the actual expression for the conditional expectation is known.

The MC+regress approach is based on sampling both profits and EM data samples, and then using linear regression to fit the conditional expectation going into the VOI calculation (Eq. (17)). Bootstrap screening is used to remove insignificantly different values in the posterior, as described above.

The Monte Carlo results for the VOI in Table 3 are representative of the analytical values but with some Monte Carlo variability. The westernmost line is the most valuable because the prior profits are very near 0 in this region, and acquiring EM data in this spatial part of the domain will clearly inform the decision maker whether to drill or not at these western segments of the reservoir. Thus the western line would be selected for data gathering if the price of data is smaller than the VOI at this line. The decision about data acquisition and processing depends on the price of data. If the price is more than the VOI, purchasing the additional data is hard to justify. The VOI depends on the accuracy of the EM data which is quantified by the variance Var $(\boldsymbol{y}|\boldsymbol{x}) = \tau^2 \boldsymbol{I}_{25}$. When this accuracy is halved (standard deviation is doubled, $\tau = 2$), the VOI is less than half of the VOI when $\tau = 1$ for all survey lines.

The Monte Carlo results are clearly more accurate for a larger sample size (5000 compared with 500). For the regression approach there is now less upwards skewness in the results. This occurs because the larger sample size gives a better regression fit, with more reliable effects and less spurious correlations that sometimes induce positive tendencies that are really just a result of random Monte Carlo variability. Note that if no bootstrap screening of the regression effects is applied, this skewness effect is much larger. For instance, the VOI interval of the westernmost line is then (7.55-8.71) for low accuracy and 500 samples, while it is (5.08-5.59) when using 5000 samples. With high accuracy in the EM data the VOI interval is then (12.7-14.1) with 500 samples and (10.9-11.3) with 5000 samples.

4.2 Seismic data for reservoir development decisions

Next, a complex value function incorporating fluid flow in a heterogeneous medium is studied, but the scope is limited to a relatively simplified decision situation where the decision maker can develop the reservoir by drilling a fixed configuration of wells. There are nine well configurations that are possible; $a \in \{1, \dots, 9\}$. The decision maker can also avoid development ($a = 0$). A petroleum company could have such limited alternatives for various reasons, say there exists infrastructure making a (few) well configuration(s) much less expensive than others. The drilling alternatives considered here are shown in Fig. 7 through a map view indicating the location of the vertical wells.

The decision maker has the option to collect seismic data before choosing a particular drilling alternative. The seismic data would on average reduce spatial uncertainty and thereby help in better decision making. One might be interested in evaluating the value of seismic data in the context of this decision to justify collecting the data. For the baseline dataset a frequency bandwidth 10-40 Hz in the seismic data is used. Results with shorter frequency band (10-20 Hz) and wider frequency band (10-60 Hz) are also shown.

The prior model is represented using an ensemble of 500 realizations of reservoir properties: facies, porosity and permeability. The geological uncertainty is two-fold: scenario uncertainty with two possible scenarios: channel and delta, and stochastic variations within each scenario. The facies realizations are generated using the geostatistical algorithm SNESIM (Strebelle, 2002) with two different training-images for each scenario. Fig. 8 shows one facies realization each of the channel scenario and the delta scenario at the top. The porosity realizations are generated using sequential Gaussian simulation conditioned on the facies. The permeability realizations are obtained using the Kozeny-Carman equation (Mavko et al., 2009) relating porosity to permeability.

The seismic data is generated independently for each realization. The acoustic impedance (AI) is first computed at the geostatistical scale by applying rock physics modeling on the reservoir properties, and then a Fourier-domain Born filter is applied to represent the AI at the seismic scale. The AI at the seismic scale approximates what one would obtain by inverting the seismic data. Fig. 8 (bottom) shows the AI realizations corresponding to each of the two geological scenario.

To compute the prospect values, flow simulation is run on each realization of reservoir variables, for each well configuration. The oil and water production rates for the ninth drilling alternative are shown in Fig. 9. The variability in oil and water production is a result of the uncertainty in the a priori reservoir model. For each of the nine drilling alternatives, the net present value (NPV) v_a^b for each realization is computed using the following function:

$$v_a^b = \int_0^T \frac{q_{o,a}\left(t,b\right) r_o - q_{w,a}\left(t,b\right) r_w}{\left(1+r\right)^t} dt, \tag{18}$$

where time is t going from the start of production until production life T (truncation of simulator). The oil production rate is $q_{o,a}$ while the water production rate is $q_{w,a}$. The oil price is r_0 (set to USD 40/barrel), cost of water produced is r_w (set to USD 5/barrel), and the discount rate is r (set to 10 percent per year).

First, consider the PV and the PoV with perfect information. These two are estimated using straightforward non-parametric bootstrapping over the 500 realizations. The 10th percentile, median and 90th percentile of the PV are found to be USD 3,590 million, USD 3,610 million and USD 3,630 million respectively. The corresponding percentiles of the VOI of perfect information are USD 650 million, USD 680 million and USD 705 million.

To estimate the PoV with imperfect seismic data, the expected value given the data is approximated using a regression model. The seismic data is represented on a large spatial grid and it is high-dimensional. When regressing values on the data, the effective dimension of the data is reduced by partial least squares regression (PLSR), or related high-dimensional regression techniques (Hastie, 2009). PLSR effectively chooses the dimensions of the data that are important for value approximation. It appears to be an effective method for estimating the PoV with imperfect information. In mathematical terms, PLSR constructs linear combinations of data that have high variance and large correlation with the values. In comparison, principal component regression is known to focus more on the variance in the data. PLSR is an iterative method and when this procedure is stopped after a few iterations, a reduced regression is obtained, which is often a good trade-off between model fitting and predictive power.

One PLSR model is built for each decision alternative and the values are fitted at each data point. In this application, with only a few alternatives, no screening of regression effects is done, except what is already achieved by the PLSR data reduction technique. Ten-fold cross validation is used to find the optimum number of PLSR components to use in our regression model, using the Predicted Residual Error Sum of Squares (PRESS) criterion (Abdi, 2010). It is found that the smallest PRESS and hence the optimum number of PLSR components range from 1 to 5 for the different alternatives, as shown in Fig. 10. Here, the value corresponding to the number of components 0 represents the PRESS when the regression model is just the mean of the training data. The clear reduction in PRESS in going from 0 to 1 component in the PLSR is obtained by the more refined regression model capturing most of the variability between the two geological scenarios. With more than one component, the PLSR fit also starts to pick some more spatial detail within each scenario, and this could increase the predictive power a little in this case, at least for some alternatives.

As with the PV and the PoV with perfect information, bootstrapping is employed to estimate the distribution of the PoV with imperfect seismic information. This is done by taking a constant number of PLSR components for all alternatives, ranging from 1 to 5 (see Fig. 10). The VOI results are summarized in Table 4. The VOI interval for the baseline case with 10-40 Hz seismic data is USD 533-566 million for one component in the PLSR regression and increases to USD 564-598 million for five components. The tendency of increasing VOI occurs because fewer components could yield overly smooth conditional expectations, while too many components could induce potential overfitting of the conditional expectation to the data realizations. If the number of PLSR components gets very large, the VOI of imperfect information tends towards the VOI of perfect information in this case.

The VOI result should be compared with the price of acquiring and processing seismic data. The price of the seismic data of course increases with higher resolution in the data. The VOI is computed for different seismic frequency bandwidths: 10-40 Hz (baseline), 10-20 Hz and 10-60 Hz. From Table 4, one sees that for wider bandwidth, the VOI results are larger because the high-resolution seismic data carry more information. However, the effect of going from 40 to 60 Hz is very small in our case. This indicates that the additional costs (using broader band geophones) required to acquire the broader band data might not be valuable in this situation.

Recall that in this example the design strategies for drilling are fixed. In future work, a goal is to extend this to situations with more alternatives and involving closed loop reservoir management (Barros et al., 2016).

5 Closing Remarks

There are plenty of opportunities for creative and effective spatial information gathering schemes in the petroleum industry. The value of such data is most consistently gauged by framing the spatial decision situation in a unified context including geostatistics, geophysical modeling and decision analysis. This paper describes these aspects and show examples of VOI analysis for comparing different spatial experiments, such as seismic data or electromagnetic data.

When one incorporates more realistic models for the uncertain variables and for the spatial decision models, the VOI computations become demanding, and this hinders effective VOI analysis. In this paper attempts are made to overcome this challenge by constructing efficient approximations for the VOI. Simulation-regression approaches are used to solve parts of the computational problems and provide useful approximations to the VOI. In a Gaussian linearized example with electromagnetic data, VOI results obtained by simulation-regression techniques are compared with the available analytical results. Modern regression techniques such as bootstrap can screen spurious correlation effects and provide accurate approximations. In a more realistic nonlinear example including fluid flow simulation for the value calculation and for the seismic amplitude data, PLSR techniques for the VOI analysis are demonstrated. Such new simulation-regression methods and related approaches are likely to be important for efficient VOI analysis, while maintaining flexibility of spatial alternatives and realistic value functions.

6 References

Abdi H (2010) Partial least squares regression and projection on latent structure regression (PLS Regression). WIRES Comp Stat 2: 97-106.

Avseth P, Mukerji T, Mavko G (2005) Quantitative Seismic Interpretation. Cambridge University Press.

Barros EGD, van den Hof PMJ, Jansen JD (2016) Value of information in closed loop reservoir management. Computational Geosciences 20: 737-749.

Bhattacharjya D, Eidsvik J, Mukerji T (2010) The value of information in spatial decision making. Mathematical Geosciences 42: 141-163.

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

Bickel J, Smith J (2006) Optimal sequential exploration: A binary learning model. Decision Analysis 3: 16-32.

Bickel JE (2008) The relationship between perfect and imperfect information in a two-action risk-sensitive problem. Decision Analysis 5: 116-128.

Bratvold RB, Bickel JE, Lohne HP (2009) Value of information in the oil and gas industry: Past, present, and future. SPE: Reservoir Evaluation & Engineering 12: 630-638.

Bratvold RB, Begg SH (2010) Making Good Decisions. Society of Petroleum Engineers.

Darwiche A (2009) Modeling and Reasoning with Bayesian Networks. Cambridge University Press.

Eidsvik J, Bhattacharjya D, Mukerji T (2008) Value of information of seismic amplitude and CSEM resistivity. Geophysics 73: R59-R69.

Eidsvik J, Mukerji T, Bhattacharjya D (2015) Value of Information in the Earth Sciences. Cambridge University Press.

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning. Springer.

Heath A, Manolopoulou I, Baio G (2016) Estimating the expected value of partial perfect information in health economic evaluations using integrated nested Laplace approximation. Statistics in Medicine 35: 4264-4280.

Heath A, Manolopoulou I, Baio G (2015) A review of methods for the analysis of the expected value of information. ArXiv:1507.02513.

Howard R (1966) Information value theory. IEEE Transactions on Systems Science and Cybernetics 2: 22-26.

Howard RA, Abbas A (2015) Foundations of Decision Analysis. Prentice Hall.

Howard R, Matheson J (1984) Influence diagrams. In R. Howard and J. Matheson, eds, The Principles and Applications of Decision Analysis, Vol. II. Strategic Decision Group 721-762.

Lilleborge M, Hauge R, Eidsvik J (2016) Information gathering in Bayesian networks applied to petroleum prospecting. Mathematical Geosciences 48: 233-257.

Martinelli G, Eidsvik J, Hauge R, Førland MD (2011) Bayesian networks for prospect analysis in the North Sea. AAPG Bulletin 95: 1423-1442.

Martinelli G, Eidsvik J, Hokstad K, Hauge R (2014) Strategies for petroleum exploration based on Bayesian networks: A case study. SPE Journal 19: 564-575.

Mavko G, Mukerji T, Dvorkin J (2009) The Rock Physics Handbook: Tools for Seismic Analysis of Porous Media. 2nd ed. Cambridge University Press.

Newendorp PD, Schuyler JR (2013) Decision Analysis for Petroleum Exploration. Planning Press.

Rezaie J, Eidsvik J, Mukerji T (2014) Value of information analysis and Bayesian inversion for closed skew-normal distributions: Applications to seismic amplitude variation with offset data. Geophysics 79: R151-R163.

Scheidt C, Renard P, Caers J (2015) Prediction-focused subsurface modeling: Investigating the need for accuracy in flow-based inverse modeling. Mathematical Geosciences 47: 173-191.

Strebelle S (2002) Conditional simulation of complex geological structures using multiple-point statistics. Mathematical Geology 34: 1-21.

Strong M, Oakley J, Brennan A (2014) Estimating multiparameter partial expected value of perfect information from a probabilistic sensitivity analysis sample: A nonparametric regression approach. Medical Decision Making 34: 311-326.

Trainor-Guitton WJ, Caers J, Mukerji T (2011) A methodology for establishing a data reliability measure for value of spatial information problems. Mathematical Geosciences 43: 929-949.

Trainor-Guitton WJ, Mukerji T, Knight R (2013) A methodology for quantifying the value of spatial information for dynamic Earth problems. Stochastic Environmental Research and Risk Assessment 27: 969-983,

Trainor-Guitton WJ, Hoversten GM, Ramirez A, Roberts J, Juliusson E, Key K, Mellors R (2014) The value of spatial information for determining well placement: A geothermal example. Geophysics 79: W27-W41.

Tables

Table 1: Nomenclature.

Uncertainties	$\boldsymbol{x}$
Alternatives	$\boldsymbol{a}$
Value	$v(\boldsymbol{x}, \boldsymbol{a})$
Data	$\boldsymbol{y}$
Prior value	PV
Posterior value	PoV
Value of information	VOI

Table 2: Probability mass function for the case with two dependent petroleum prospects.

	Prospect 1 dry	Prospect 1 wet	Marginal
Prospect 2 dry	0.85	0.05	0.9
Prospect 2 wet	0.05	0.05	0.1
Marginal	0.9	0.1	1

Table 3: VOI in millions (80 % coverage intervals) for three different electromagnetic survey lines (West, Center, East), for two different accuracies in the log-resistivity measurements and two sample sizes.

$B = 500$	$\tau = 1$			$\tau = 2$		
	MC+Gauss	MC+regress	Analytical	MC+Gauss	MC+regress	Analytical
West	10.4-11.0	10.9-11.9	10.8	4.71-5.22	4.97-6.10	5.00
Center	7.15-7.92	7.82-8.98	8.11	3.24-3.60	3.43-4.71	3.59
East	7.38-7.94	7.75-8.89	7.97	3.31-3.82	3.52-4.50	3.74
$B = 5000$	$\tau = 1$			$\tau = 2$		
	MC+Gauss	MC+regress	Analytical	MC+Gauss	MC+regress	Analytical
West	10.7-10.9	10.7-11.2	10.8	4.94-5.37	4.90-5.37	5.00
Center	7.90-8.18	7.87-8.57	8.11	3.47-3.67	3.59-4.00	3.59
East	7.82-8.12	7.90-8.27	7.97	3.69-3.82	3.70-4.06	3.74

Table 4: VOI of seismic data in millions (80 % coverage intervals) for the complex value function of reservoir simulation. The results are displayed for different number of components used in the Partial Least Squares Regression (PLSR) and for different frequency bandwidths in the seismic data.

PLSR order	Freq. bandwidth (Hz)	VOI (10 %)	VOI (50 %)	VOI (90 %)
1	10-20	533	551	566
2	10-20	535	556	574
3	10-20	554	575	589
4	10-20	555	578	592
5	10-20	560	581	595
1	10-40	533	551	566
2	10-40	558	576	595
3	10-40	564	583	600
4	10-40	564	583	599
5	10-40	564	583	598
1	10-60	533	551	566
2	10-60	562	578	593
3	10-60	566	585	601
4	10-60	565	583	598
5	10-60	563	583	598

Figure captions

Fig. 1: Decision trees for the situation of drilling /not drilling at one petroleum prospect with uncertain outcome. Squares represent decisions while circles represent uncertainties. Top left: Prior value. Bottom left: Posterior value of perfect information. Right: Posterior value with imperfect information

Fig. 2: A simple Bayesian network model for two petroleum prospects with a common parent node. Information can be obtained by exploration wells (perfect information) or by seismic tests (imperfect information), at one of the prospect (partial information) or at both (total information)

Fig. 3: Value of information results as a function of drilling cost (first axis) for the situation with two petroleum prospects. Information at one of the prospects (top) and both prospects (bottom)

Fig. 4: Left: The value is a function of the selected alternative and the uncertain variable. A likelihood model relates the uncertain variable to the data. Right: Simulation-regression approximation involves sampling of data and values, and regressing value on data to fit the conditional mean

Fig. 5: Prior means for monetary profits at the reservoir segments. The three lines indicate possible survey lines for electromagnetic data. The dots indicate the near and far site, when data is collected along the westernmost line

Fig. 6: Illustration of the simulation-regression approach for a site near data (top) and a site far from data (bottom). The displays show 50 sets of values and data generated by Monte Carlo sampling, and the fitted regression line (left). Bootstrap replicates over data is one way of screening insignificant effects (right)

Fig. 7: The nine drilling design alternatives in the example with flow simulation and seismic data for improved reservoir development decisions

Fig. 8: One facies realization each of (a) the channel scenario, and (b) the delta scenario. The acoustic impedance (AI) at the seismic scale for (c) the facies realization shown in (a), and (d) the facies realization shown in (b)

Fig. 9: The oil production rates (a) and the water production rates (b) for the ninth drilling alternative

Fig. 10: Predicted Residual Sum of Squares (PRESS) versus the number of Partial Least Squares Regression (PLSR) components for each of the nine decision alternative

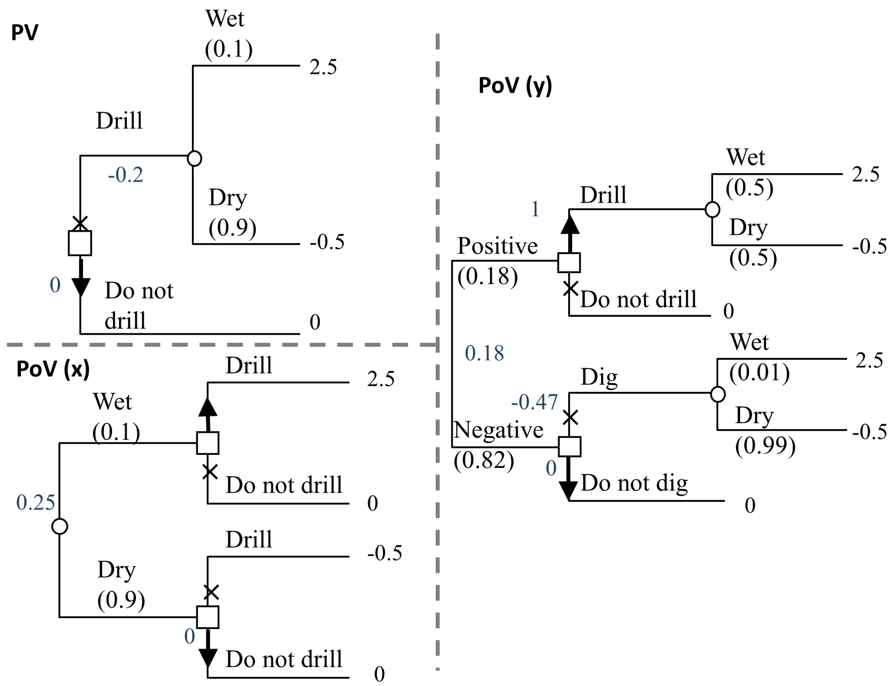

PV
Wet
(0.1)
2.5
Drill
-0.2
Dry
(0.9)
-0.5
0
Do not
drill
0
PoV (x)
Drill
Wet
(0.1)
2.5
Do not drill
0
0.25
Drill
-0.5
Dry
(0.9)
0
Do not drill
0
PoV (y)
Wet
(0.5)
2.5
1
Drill
Dry
(0.5)
-0.5
Positive
(0.18)
Do not drill
0
0.18
Wet
(0.01)
2.5
Dig
-0.47
Dry
(0.99)
-0.5
Negative
(0.82)
0
Do not dig
0

Figure 1:

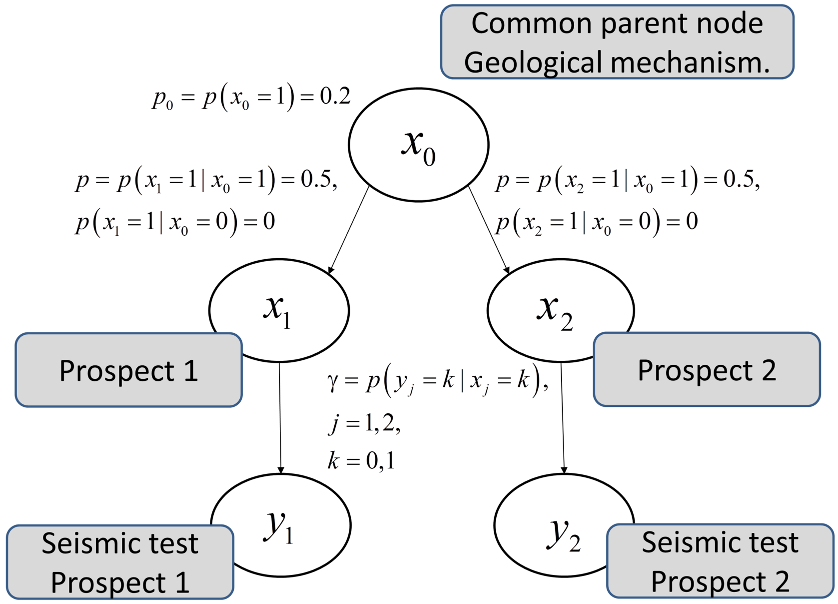

Common parent node
Geological mechanism.
$p_0 = p(x_0 = 1) = 0.2$
x_0
$p = p(x_1 = 1 | x_0 = 1) = 0.5,$
$p(x_1 = 1 | x_0 = 0) = 0$
$p = p(x_2 = 1 | x_0 = 1) = 0.5,$
$p(x_2 = 1 | x_0 = 0) = 0$
x_1
x_2
Prospect 1
Prospect 2
$\gamma = p(y_j = k | x_j = k),$
$j = 1,2,$
$k = 0,1$
y_1
y_2
Seismic test
Prospect 1
Seismic test
Prospect 2

Figure 2:

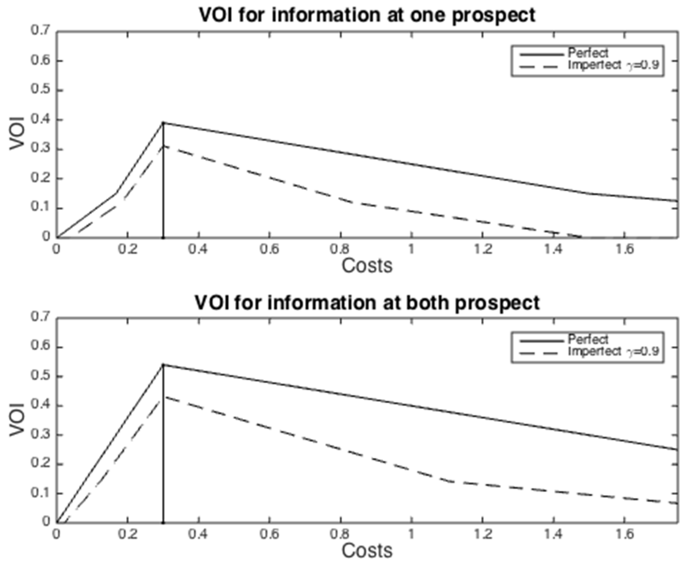

VOI for information at one prospect
Perfect
Imperfect γ=0.9
VOI
Costs
VOI for information at both prospect
Perfect
Imperfect γ=0.9
VOI
Costs

Figure 3:

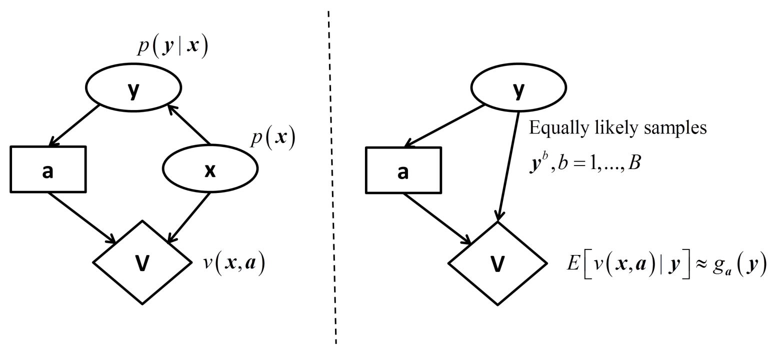


Figure 4:

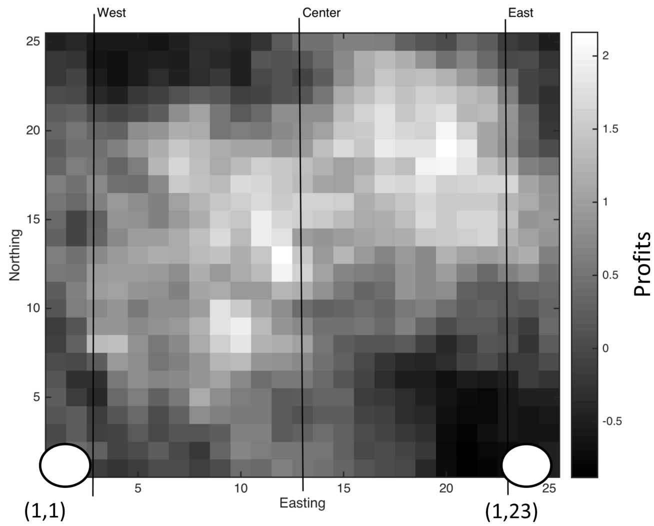
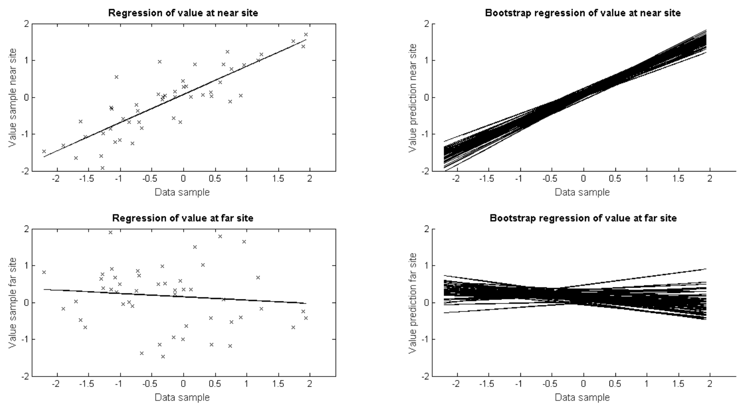

West
Center
East
25
20
15
10
5
Northing
5
10
15
20
25
Easting
(1,1)
(1,23)
2
1.5
1
0.5
0
-0.5
Profits

Figure 5:



Regression of value at near site
Bootstrap regression of value at near site
Regression of value at far site
Bootstrap regression of value at far site
Value sample near site
Value prediction near site
Value sample far site
Value prediction far site
Data sample

Figure 6:

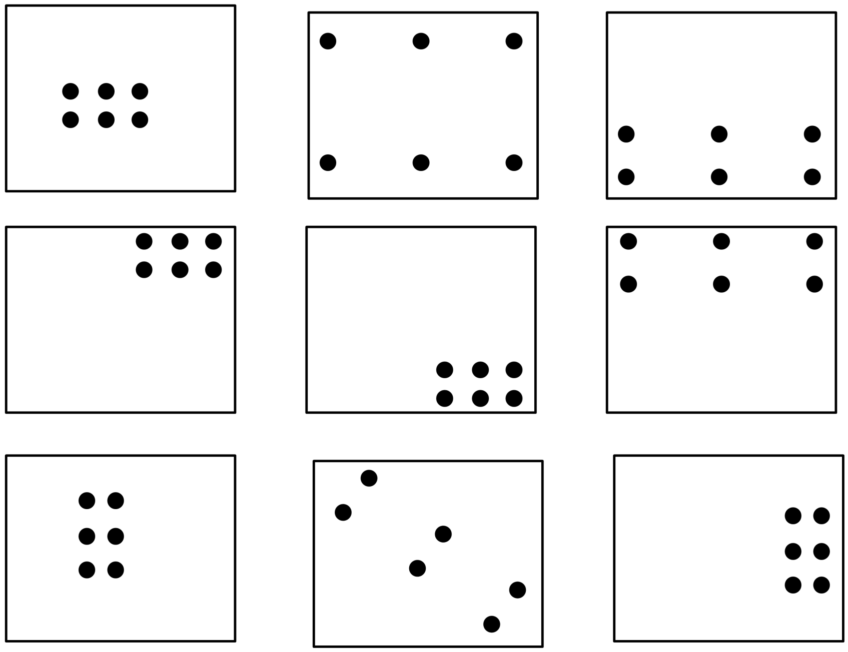


Figure 7:

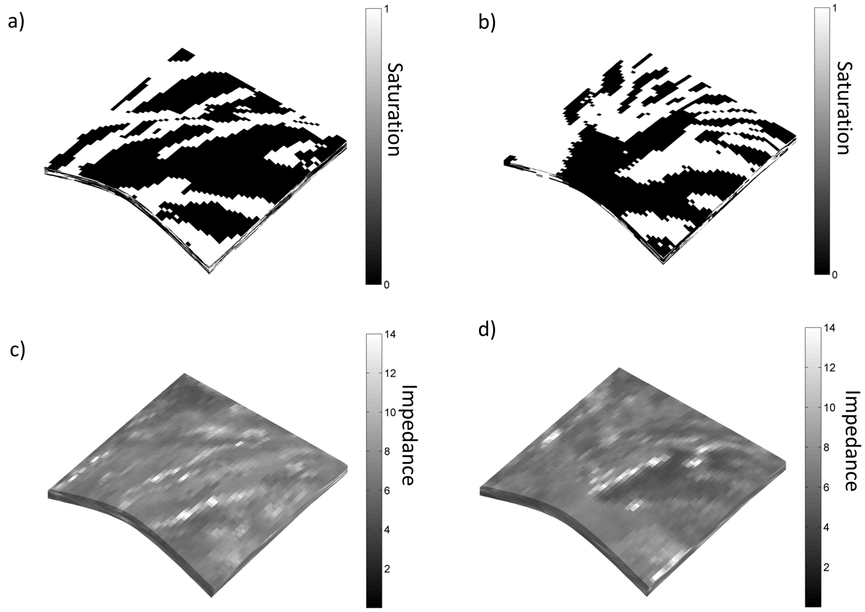
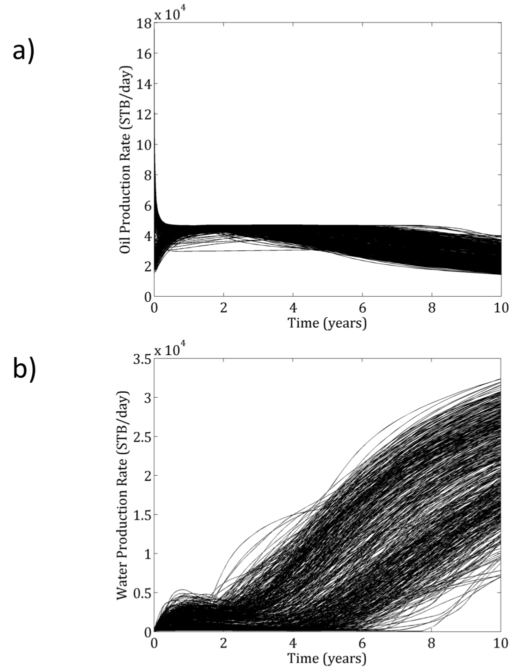

a)
Saturation
1
0
b)
Saturation
1
0
c)
Impedance
14
12
10
8
6
4
2
d)
Impedance
14
12
10
8
6
4
2

Figure 8:



a)
x 10^4
Oil Production Rate (STB/day)
Time (years)
b)
x 10^4
Water Production Rate (STB/day)
Time (years)

Figure 9:

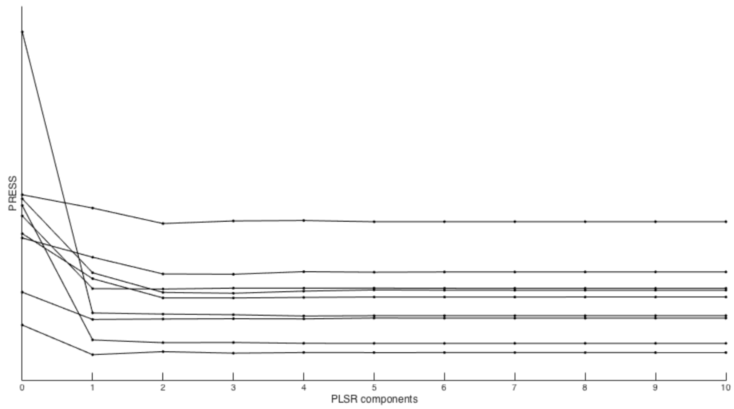


Figure 10: