JYU DISSERTATIONS 855

Vilja Koski

Applying Value of Information and Subsample Selection to Cost-Efficient Lake Monitoring

UNIVERSITY OF JYVÄSKYLÄ
FACULTY OF MATHEMATICS
AND SCIENCE

JYU DISSERTATIONS 855

Vilja Koski

Applying Value of Information and Subsample Selection to Cost-Efficient Lake Monitoring

Esitetään Jyväskylän yliopiston matemaattis-luonnontieteellisen tiedekunnan suostumuksella
julkisesti tarkastettavaksi yliopiston Agora-rakennuksen Lea Pulkkisen salissa
joulukuun 14. päivänä 2024 kello 12.

Academic dissertation to be publicly discussed, by permission of
the Faculty of Mathematics and Science of the University of Jyväskylä,
in building Agora, Lea Pulkkinen hall, on December 14, 2024 at 12 o'clock noon.

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2024

Editors
Salme Kärkkäinen
Department of Mathematics and Statistics, University of Jyväskylä
Ville Korkiakangas
Open Science Centre, University of Jyväskylä

ISBN 978-952-86-0408-2 (PDF)
URN:ISBN:978-952-86-0408-2
ISSN 2489-9003

Permanent link to this publication: http://urn.fi/URN:ISBN:978-952-86-0408-2

ABSTRACT

Koski, Vilja
Applying value of information and subsample selection to cost-efficient lake monitoring
Jyväskylä: University of Jyväskylä, 2024, 48 p. (+included articles)
(JYU Dissertations
ISSN 2489-9003; 855)
ISBN 978-952-86-0408-2 (PDF)

This thesis employs decision analysis and subsample selection tools to support environmental management decision-making under uncertainty. The practical motivation arises from management of lake water quality within in Finland. If the ecological status of a lake based on monitoring data is weak, the European Union's Water Framework Directive obliges its member countries to implement management actions to improve the status. Though legally and biologically principled, demonstrating the cost-efficiency and value of monitoring remains challenging.

In this thesis, value of information (VOI) is used to quantify the value of lake monitoring data. VOI is a concept of decision analysis to assess the value of additional information before it is collected. This thesis is one of the first attempts to apply VOI to real-life environmental monitoring data. A risk averse decision-maker and its effect on VOI in lake management is also considered. This has often been ignored in practical applications. In addition, heuristic subsample selection algorithms are applied to identify subsamples that either a) maximize the VOI of the sample, or b) maximize the D-optimality criterion, with the aim of finding a sample where the fitted statistical model estimates the model parameters as precisely as possible.

The findings indicate that VOI can be effectively applied to lake monitoring data. From a lake management perspective, the primary conclusion is that current monitoring is cost-efficient. Monitoring should focus on lakes that, based on preliminary data, are not expected to require immediate management actions, as well as those where the ecological status remains uncertain. This study encourages further application of VOI analysis to address environmental challenges.

Keywords: decision-making, lake management, optimal design, optimality criteria, risk aversion, utility function, value of information

TIIVISTELMÄ (ABSTRACT IN FINNISH)

Koski, Vilja
Informaatioarvon ja osaotoksen valitsemisen soveltaminen kustannustehokkaaseen järvien seurantaan
Jyväskylä: University of Jyväskylä, 2024, 48 s. (+artikkelit)
(JYU Dissertations
ISSN 2489-9003; 855)
ISBN 978-952-86-0408-2 (PDF)

Tässä väitöskirjassa hyödynnetään päätösanalyysin työkaluja ja osaotoksen valitsemista ympäristönsuojeluun liittyvässä päätöksenteossa. Käytännön tutkimuskysymys liittyy Suomen järvien hoitoon. Jos seuranta-aineistoon perustuva järven ekologinen tilaluokka on heikko, Euroopan Unionin vesipuitedirektiivi velvoittaa jäsenmaitaan toteuttamaan hoitotoimenpiteitä tilan parantamiseksi. Vaikka seuranta on sekä laillisesti että biologisesti perusteltua, sen kustannustehokkuus ja hyöty on vaikea osoittaa.

Väitöskirjassa käytetään informaatioarvoa (engl. VOI) arvioimaan järvien seuranta-aineiston arvoa. Informaatioarvo on päätösanalyysin käsite, jonka tarkoitus on arvioida lisäaineiston arvo jo ennen kuin aineistoa on kerätty. Tämä väitöskirja on yksi ensimmäisistä yrityksistä soveltaa informaatioarvon käsitettä todelliseen ympäristön seuranta-aineistoon. Riskineutraalin päätöksentekijän lisäksi tarkastellaan riskinkarttajaa ja sen vaikutusta informaatioarvoon järvien hoitoon liittyvässä kysymyksessä. Tätä ei ole usein huomioitu käytännön sovelluksissa. Lisäksi väitöskirjassa sovelletaan heuristisia osaotoksen valinta-algoritmeja tunnistamaan sellaisia osaotoksia, jotka joko a) maksimoivat otoksen informaatioarvon, tai b) maksimoivat D-optimaalisuuskriteerin tavoitteena löytää otos, johon sovitettu tilastollinen malli estimoi mallin parametrit mahdollisimman tarkasti.

Tulokset osoittavat, että informaatioarvon käsitettä voidaan onnistuneesti soveltaa järvien seuranta-aineistoon. Ympäristönsuojelun näkökulmasta tärkein johtopäätös on, että seuranta on kustannustehokasta. Seurannan tulisi ensisijaisesti keskittyä järviin, jotka eivät ennakkotiedon perusteella tarvitse hoitotoimenpiteitä sekä järviin, joiden tila on vielä epävarma. Tutkimus kannustaa jatkamaan informaatioarvon käsitteen soveltamista ympäristökysymyksiin.

Avainsanat: päätöksenteko, järvien hoito, optimaalinen asetelma, optimaalisuuskriteeri, riskin karttaminen, hyötyfunktio, informaatioarvo

Author

Vilja Koski
Department of Mathematics and Statistics
University of Jyväskylä

Supervisors

Dr. Salme Kärkkäinen
Department of Mathematics and Statistics
University of Jyväskylä

Professor Juha Karvanen
Department of Mathematics and Statistics
University of Jyväskylä

Dr. Niina Kotamäki
Finnish Environment Institute

Reviewers

Professor Claire Miller
School of Mathematics & Statistics
University of Glasgow

Dr. Evangelos Evangelou
Department of Mathematical Sciences
University of Bath

Opponent

Professor Jarno Vanhatalo
Department of Mathematics and Statistics, Faculty of Science
Organismal and Evolutionary Biology Research Programme, Faculty of Biological and Environmental Sciences
University of Helsinki

ACKNOWLEDGEMENTS

My heart is full of joy as I write this, because it means I have made it this far. While I am extremely proud of my achievement, there are several people and entities to thank for that.

First, I would like to express my gratitude to my two supervisors at the university, Dr. Salme Kärkkäinen and Professor Juha Karvanen. Both supervisors' energy and unwavering support has been necessary throughout my academic journey and encouraging in my personal life.

I am deeply thankful to Salme, who not only supervised my master's thesis but also encouraged me to pursue doctoral studies. Her bright attitude and exceptional talent in her field have been a constant source of inspiration. I have especially enjoyed our lunches together, which have been both delightful and motivating.

I am immensely grateful to Juha, whose guidance was essential for me to complete my last articles. His expertise, insightful advice, and emphatic attitude helped me to finish the thesis. The stories he shares about his career have been a great source of learning as well as entertainment.

I am also deeply grateful to my talented supervisor from the Finnish Environment Institute, Dr. Niina Kotamäki. I would like to thank Niina for providing me the practical application, interesting research questions and the data, while guiding me through the complexities of environmental and legislative issues.

During my studies, I had the privilege of collaborating with one of the foremost experts in my field, Professor Jo Eidsvik, who graciously invited me to visit in his university. I truly appreciate his warmth and generosity throughout my stay, which made the experience both productive and enjoyable. His guidance, encouragement, and expertise were instrumental in the completion of the third article in my thesis. Thank you, Jo, for your mentorship and support.

In addition, I want to extend my sincere thanks to Docent Kristian Meissner from the Finnish Environment Institute and Dr. Heikki Hämäläinen from the Department of Biological and Environmental Sciences at the University of Jyväskylä, whose expertise helped me navigate the challenges of my first article.

I wish to express my gratitude to Professor Jarno Vanhatalo for serving as the opponent and to Professor Claire Miller and Dr. Evangelos Evangelou for their work as examiners. Thank you for your time and energy you gave to my thesis.

My work was funded by the Emil Aaltonen Foundation, the Kone Foundation and the Department of Mathematics and Statistics at the University of Jyväskylä, which I am grateful. Also, thank you to the department for providing the comfortable working facilities during my work there. I want to acknowledge CSC – IT Center for Science for computational resources.

Thank you to my colleagues at the Department of Mathematics and Statistics. I have felt welcome from the very beginning of my studies. I am especially grateful to my fellow doctoral students, whose peer support and advice have

been invaluable throughout the years.

I would like to thank my family and friends. I am very lucky to have you in my life. Special thanks to my sweet little daughter, who sometimes made the workdays feel like holiday.

Jyväskylä, November 2024
Vilja Koski

CONTENTS

LIST OF INCLUDED ARTICLES

This thesis consists of an introductory part and the following publications, referred to as Articles I–IV in the text.

I Koski, V., Kotamäki, N., Hämäläinen, H., Meissner, K., Karvanen, J., & Kärkkäinen, S. The value of perfect and imperfect information in lake monitoring and management. *Science of the Total Environment, 726,* DOI: https://doi.org/10.1016/j.scitotenv.2020.138396, 2020.

II Koski, V., Kärkkäinen, S. & Karvanen, J. Subsample selection methods in the lake management. *JABES,* DOI: https://doi.org/10.1007/s13253-024-00630-0, 2024.

III Koski, V., & Eidsvik, J. Sampling design methods for making improved lake management decisions. *Environmetrics, e2842.* DOI: https://doi.org/10.1002/env.2842, 2024.

IV Koski, V. & Karvanen, J. Risk aversion in the value of information analysis: application to lake management. Submitted to *Stochastic Environmental Research and Risk Assessment*, 2024.

The author of this thesis was the main author of the joint Articles I–IV. In particular, she was mainly responsible for the writing of the articles and solely responsible for the data curation and visualization within all of the articles. Except for some spatial modelling results in Article III, she has implemented all the analyses in all the articles. The author had the main responsibility of writing the programming code in Articles I and IV, and modified earlier code and created new code based on it in Articles II and III. Also, she had the main responsibility of ensuring the correctness of the programming code in all the articles. The ideas of the research questions of the articles were formulated together with the co-authors.

1 INTRODUCTION

The present thesis applies statistical methods to address a practical problem in water management decision-making in Finland. Water management is a part of the implementation of the EU Water Framework Directive (WFD) (European Parliament, 2000). The goal is to achieve and secure at least a good status of surface and groundwater. The directive obliges its member countries to implement management actions to improve the status if needed. An essential part of the water management is the monitoring of ecological status of waters, which forms the basis for understanding changes in water systems and provides relevant information for challenging management decision-making (Kotamäki et al., 2024). The thesis focuses on the monitoring of Finnish lakes.

Although lake monitoring is justified on legal and ecological grounds, its cost-efficiency is difficult to quantify, which is also evidenced by the paucity of literature in this space (Wätzold and Schwerdtner, 2005; Nygård et al., 2016). According to Lovett et al. (2007), environmental monitoring has been criticized for being expensive and wasteful, and for never using the most of monitoring data. In their review, they list common criticisms while emphasizing the benefits of monitoring programs. From the lake management point of view, the primary goal of this thesis is to demonstrate the cost-efficiency and overall value of the lake monitoring in Finland. To achieve this, the following questions are asked:

1. What is the value of lake monitoring data?
2. How much monitoring data is needed?
3. How should the monitoring design be selected to achieve the maximum benefit at the lowest possible cost?
4. How does the decision-maker's attitude to risk affect decision-making in lake management?

To answer these practical questions, we apply existing statistical methods. A very central tool is value of information (VOI). It is a concept of decision theory, which can be used to evaluate the (monetary) value of the data even before it is collected. This assessment can aid to decide whether to acquire data or not: if the cost of data acquiring is lower than the VOI, then the acquiring is profitable. A relatively

recent and comprehensive presentation is given e.g. by Eidsvik et al. (2015), who introduce examples from the earth sciences.

Initially, VOI has been developed by economists (Raiffa and Schlaifer, 1961), but the analysis framework is widely used in a variety of different fields. The most well-known and widely used fields are finance (Lawrence, 1999), health care (Zonta et al., 2014), and the combination of these (Pozzi and Kiureghian, 2011). Other practical applications that have utilized VOI analysis are, for example, fishery management (Mäntyniemi et al., 2009), environmental health-risk management (Yokota and Thompson, 2004a), medical clinical trials (see, e.g., a review by Yokota and Thompson (2004b)) and optimization of manufacturing (Marchese et al., 2018). In the context of environmental monitoring, assessing the value and optimal level of monitoring, the interest has grown over the past years (Bouma et al., 2009; Williams et al., 2011; Canessa et al., 2015; Bolam et al., 2019), and indeed, there has been a demand for it (Colyvan, 2016). Also, new research on the topic has appeared just recently (Venus and Sauer, 2022; Luhede et al., 2024).

Usually in decision analysis, it is assumed that the decision-maker has a neutral attitude towards risk. This makes the calculations easier, but however, in real life, it is generally believed that humans are risk averse (Davies and Satchell, 2007). In fact, not addressing the risk aversion may even lead to errors in the analysis in some decision situations (see e.g. Keefer (1991)). The relation of VOI and risk has been actively studied in the economics and operations research (Hilton, 1981; Mehrez, 1985; Nadiminti et al., 1996; Bickel, 2008; Delquié, 2008; Abbas et al., 2013; Sun and Abbas, 2014), but it is rarely considered in other applications. In this thesis, we are also interested to study the relationship of VOI and the decision-maker's risk attitude in the context of lake monitoring.

Another essential topic of this thesis is the subsample selection in the case of an observational study. It is related to the optimal experimental design, with some differences. While in both problems, the aim is to find a design that maximizes (or minimizes) the selected optimality criterion, in subsample selection we assume a finite set of possible observation points. More importantly, in subsample selection, each observation point can be selected only once. An overview and a framework for optimal design of observational studies is presented by Karvanen et al. (2017). An increase in data availability typically reduces uncertainty, thereby facilitating more informed decision-making. However, practical constraints often limit the amount of data that can be collected (Brown et al., 2005). A simple random sampling approach is the most common choice, but sometimes a carefully selected nonrandom sample may provide benefits. There exist a number of different criteria for optimal designs, see Ryan et al. (2016) for a review. Our first choice is to select a design that has large VOI, which can be compared to the actual costs of the data gathering in order to find out how much data should be collected, but more traditional optimality criteria are alphabet criteria, e.g. D-optimality (Atkinson et al., 2007). The goal of D-optimality criterion is to minimize the determinant of the information matrix in order to find a design that estimates a model fitted to the subsample as precisely as possible.

After deciding the subsample selection criterion, the computational problem of finding the best design still exists. A review of different methods is offered by García-Ródenas et al. (2020). In recent works, the optimal approximate designs have been found using a greedy method (Reinikainen et al., 2016; Reinikainen and Karvanen, 2022), Bayesian optimization (Paglia et al., 2022) and exchange algorithms (e.g. Harman et al. (2020)). We utilize all these in Articles II and III.

This thesis consists of an introductory part with an aim to familiarize the reader with terms and notions of statistics and the lake management, and four articles, each of which provides tools for decision-making in the context of lake monitoring. Articles I, III and IV focus on VOI. Article I calculates the value of perfect and imperfect information, where the latter is approximated utilizing a Monte Carlo type method with empirical data. Article I addresses question 1 proposed earlier. Article IV expands the assumption of Article I of a risk neutral decision-maker and considers how the VOI changes if the decision-maker is risk averse instead. This addresses question 4. In Article III, the research question is how one should select a cost-efficient subsample, when considering the uncertainty of the information gained from the data as well as the costs of gathering it. Instead of addressing the value of information, Article II is the only one using the methodology of the optimal design. Similarly as in Article III, the question is to find an optimal subsample. More specifically, the question is how one should select a subsample, when the aim is to estimate the parameters of a regression model as precisely as possible using a D-optimality criterion. Articles II and III address questions 2 and 3. In all Articles I–IV, the research questions are approached using the lake monitoring data.

The structure of the rest of the introductory part is as follows. Chapter 2 introduces the lake management in Finland and the monitoring data used in the articles. Chapter 3 presents the notations and concepts of the decision analysis to form a basis to the theory of the value of information. Chapter 4 continues to describe the statistical methods used in the thesis by outlining the decision problem and the criteria to solve it. In addition to the value of information, the D-optimality from the theory of experimental design is introduced. Chapter 5 briefly discusses two main computational methods used in the thesis. First, the subsample selection methods are discussed. Second, the reader is familiarized with Gaussian processes which are used to form a basis to the VOI approximation discussed in Article III. Chapter 6 summarizes the research contribution of the articles to the research questions outlined above. Finally, Chapter 7 concludes with highlighting the results and the overall significance of the thesis and stating the future study directions.

2 LAKE MANAGEMENT IN FINLAND

In 2000, the European Union enacted the Water Framework Directive (WFD), the purpose of which is to protect and improve the quality of the inland waters in all EU countries (European Parliament, 2000). The legislation places clear responsibilities on national authorities, one of which is to monitor the status of the water in each basin. The quality of the water systems is based on several indicator variables reflecting the biotic structure of lakes (Fig. 1). Based on the pre-determined, undisturbed reference conditions of each parameter (Nõges et al., 2009), the water systems are classified into five ecological status classes: high, good, moderate, poor and bad. Moreover, the directive obligates the member countries to implement restoration actions to improve the status if it is moderate or weaker.

The key tool for implementing the directive is the River Basin Management Planning (RBMP) (Aroviita et al., 2019). It is drawn up after extensive public consultation and is valid for a six-year period. Currently in Finland, there are two completed periods. The first RBMP was adopted in 2009 including for the years 2009–2015 (Vuori et al., 2009) and the second in 2015 including the years 2016–2021 (Aroviita et al., 2012). The third period in 2022–2027 is ongoing.

One RBMP period is a cycle with at least four steps (Stankey et al., 2005; Higgins et al., 2021). The first step is the monitoring of the water systems. It should be followed by the assessment and classification of them to the ecological status classes. Based on the classification, the policy-makers determine the programs of management actions. The period should end with the implementation of the actions.

2.1 Monitoring data

It is said that Finland is the land of thousands of lakes. In fact, there are about 187,000 lakes in Finland, according to the Finnish Environment Institute (Heiskanen et al., 2017). The correct amount depends on how the lake is defined: it is affected by the area of the lake and the stability of water. The water management

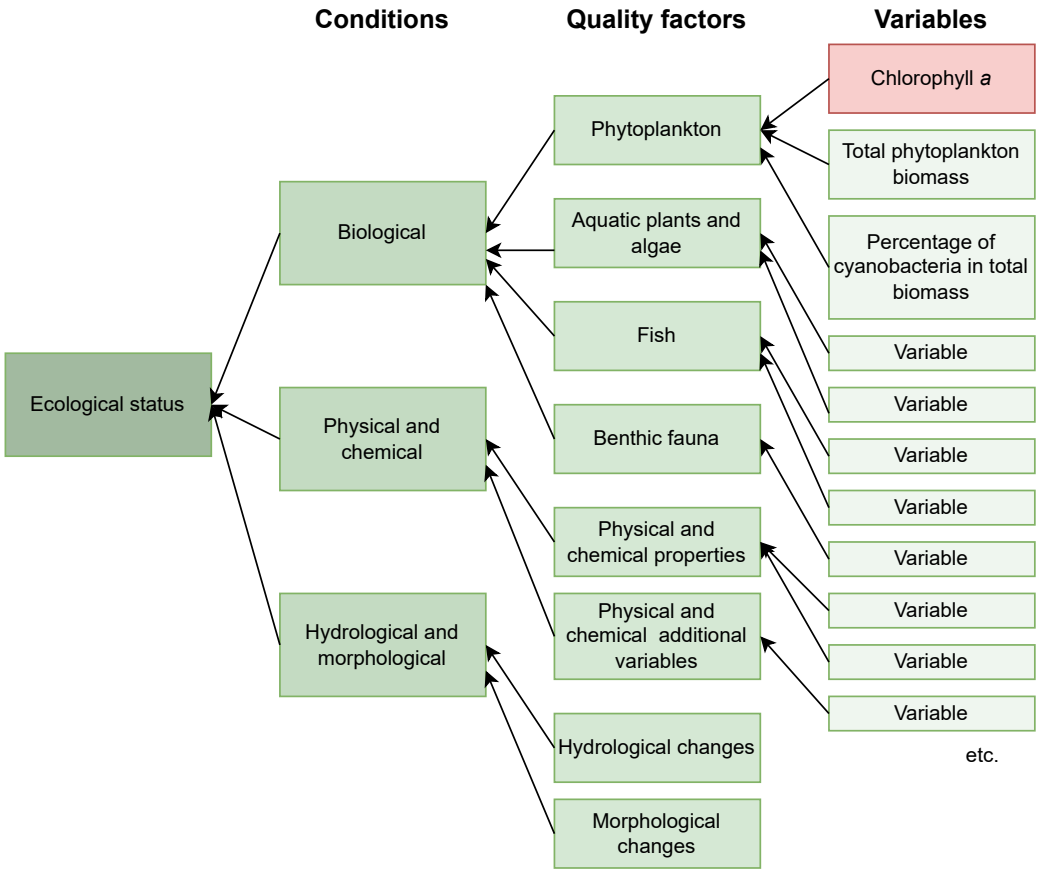


FIGURE 1 The ecological status of the lake is based on several indicator variables. We are interested in one of the most important indicators of eutrophication, chlorophyll-*a* concentration. The presentation follows the description presented by Aroviita et al. (2019).

in Finland is an extensive effort regulated by EU legislation.

In water management, the basic unit is a water body. It is a unit belonging to the same surface water type, whose status assessment and environmental goals can be unambiguously defined. Water management examines in more detail those waters that are designated as water bodies. Not all smaller waterways have been defined as water bodies and therefore do not fall within the scope of the classification (Aroviita et al., 2019). The water bodies have been published as open spatial data (Fig. 3, left). In this thesis, we are interested in the ecological status of Finnish lakes. A lake can form one water body or it may be divided into multiple water bodies if it is ecologically justified. Each lake has at least one sampling site, but the largest lakes may have multiple sites because they have different habitats and therefore, multiple water bodies. Currently, a total of 4,639 lakes have been defined as water bodies in Finland (Aroviita et al., 2019). Monitoring and management measures are being targeted at these lakes.

The ecological status data is a part of the official Finnish lake monitoring program. The data is produced and collected mainly by the organizations of environmental administration, especially Centers for Economic Development, Transport and the Environment (ELY Centers), and stored as an open source database by the Finnish Environment Institute (SYKE). In Articles II and III, we utilize the data from the latest ecological status classification based on the monitoring data collected during the years 2012–2017. We have the status classification from 4,360 water bodies (Fig. 3, center). Since the need for management actions is our main interest, we reclassified the water bodies based on it. Of the lakes, 3,616 lakes do not demand management actions (target ecological status, i.e. high or good status class) while 744 need them (non-target status). In addition, we use register-based data maintained by SYKE (https://www.syke.fi/en-US/Open_information/Open_web_services/Environmental_data_API). The lake register contains basic features of the Finnish lakes having area over one hectare, therefore, several of them are not defined as water bodies. We have the basic features from 58,707 lakes in Finland. The central basic features contained in the register are the information about a lake's location, such as the municipality, drainage basin and center latitude and longitude coordinates, as well as other information about a lake's features, such as waterbed area, length of shoreline, average and maximum depth, volume of water mass and altitude above sea level. In addition to basic variables listed above, we use an agricultural area by the municipality where the lake is located (Official Statistics of Finland, 2020).

In Article II, we use the data for the 4,360 lakes (Fig. 3, center), but we simulate a situation where the status is yet to be defined in our study. In turn, in Article III, when the status classification is available for 4,360 of the 58,707 lakes, we forecast the status for the remaining lakes, using the model trained on data from lakes.

As said, the total ecological status classification is based on data collected for several indicators (Fig. 1). In Articles I and IV, we limit the study to one indicator variable, phytoplankton, more specifically to chlorophyll-*a*. The chlorophyll-*a* concentration indicates well the human-induced eutrophication, which poses a

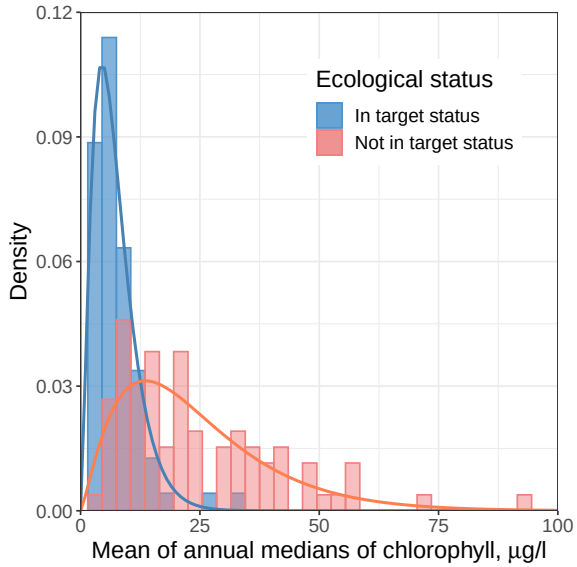


FIGURE 2 Figure from Article I. Histograms and fitted gamma distributions of chlorophyll-*a* concentration of 166 water bodies. The water bodies are categorized into two classes based on the need for management actions: either the lake is in need of them (red) or not (blue). The value on the horizontal axis is the aggregated value of annual and spatial chlorophyll-*a* samples in a water body. The data is used in Articles I and IV.

significant threat to freshwater ecosystems (Carpenter et al., 1998). The data is maintained by SYKE in an open source database (http://www.syke.fi/en-US/Open_information). The samples are gathered from the sites in summer, which means approximately the period from May to August. We have chlorophyll-*a* samples collected during the years 2006–2012 and from 144 lakes, 166 water bodies within them. We selected the water bodies from which at least three observations were taken during a year. Eventually, the data we use in our analysis consist of 6,742 observations from 166 water bodies. We compiled the observations from different years and locations into means of annual medians per a water body. This is the standard current approach for assessing the ecological status of water bodies (Aroviita et al., 2019). From the water bodies, 25 are classified as high ecological status, 54 as good, 61 as moderate, 25 as poor and 1 as bad according to the principles of the ecological status classification of the year 2014. We reclassified the water bodies based on whether they are in the need of management actions or not, resulting in 79 as the ones that are not in need (target ecological status) and 87 as the ones that need them (non-target status). Finally, we estimated the distribution of compiled values of chlorophyll-*a* over time and locations by fitting gamma distributions, separately for water bodies in both target status and non-target status (Fig. 2). The locations of the lakes from which the chlorophyll observations has been collected are presented in Figure 3 (right).

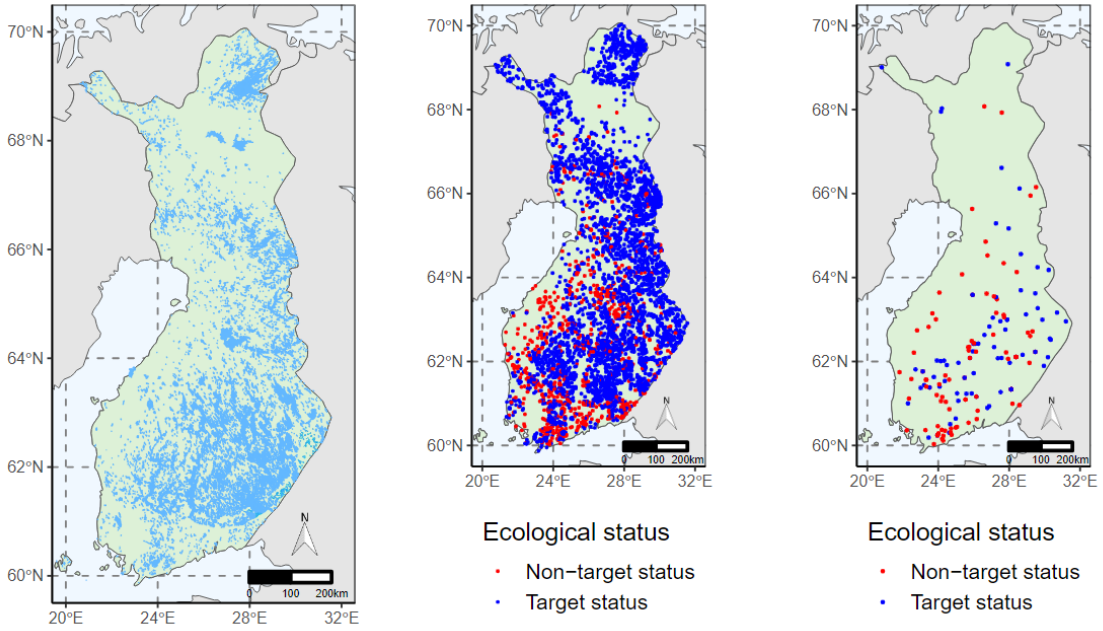


FIGURE 3 Left: Water bodies according to WFD (European Parliament, 2000) from SYKE's open environmental information system. The data is available at https://wwwd3.ymparisto.fi/d3/gis_data/spesific/VHSvesimuodostumat2016.zip. The map is adapted from Article III. Center: The ecological status classification of 4,360 water bodies used in Articles II and III. Right: The ecological status classification of 166 water bodies with chlorophyll-*a* data, which was used in Articles I and IV.

2.2 Monetary value of lakes

In Articles I, III and IV, we discuss the value of the monitoring data, for the calculation which is needed to address the monetary values of the lake. To assess the monetary values of the status of Finnish lakes, we applied the results of a valuation research by Ahtiainen (2008). The study evaluates the financial benefits of improving the condition of Lake Hiidenvesi in Southern Finland (area about 3000 hectares), by querying residents' readiness to pay for reducing the eutrophication of the lake. The status of the lake was moderate at the time of the study.

The query indicates that the estimated sum for willingness of properties to pay was between EUR 3 and 5.7 million over the management actions implementation period of five years. In Articles I, III and IV, we choose EUR 3 million per 3000 hectares = EUR 1000 per hectare, which is the most conservative value for the value of a water body with a high or good status. A lake in moderate, poor or bad status is valued at EUR 0 per hectare. The cost of the management is assumed to be EUR 200 per hectare (based on the Finnish Environmental Institute, personal communication). Therefore, we use EUR 1000 - EUR 200 = EUR 800 per hectare as the value of a water body when management actions are performed, regardless of the success of the management actions. These estimations are the basis of all VOI calculations in this thesis.

3 DECISION THEORY

Decision-making is choosing between alternatives, that are often mutually exclusive. Decision theory is a branch of applied probability theory concerning decision situations, where the consequences of the alternatives are uncertain. The basic ideas in decision theory are similar to the mathematical game theory (von Neumann and Morgenstern, 1947) with the difference that the other player in this game is the "nature" or the "state of the world". It is an interdisciplinary field, but the methods and used concepts are usually statistical and econometric.

Decision analysis is the study of applying the methods of the decision theory into practice, providing guidance in important real-life decision situations (Eidsvik et al., 2015). The aim is to help the decision-maker to make better decisions. The term was first used by Howard (1964).

In this chapter, we review general terminology and notations related to decision-making situations. The primary source of these sections is the book "Value of information in the earth sciences: Integrating spatial modeling and decision analysis" by Eidsvik et al. (2015). In particular, we present concepts that are needed to define the value of information in Section 4.1. In addition, we extend the examination from the risk neutral decision-maker to other risk preferences.

3.1 Basic notations

We start the chapter by stating the basic notations. The variables related to the decision-making situation are classified according to whether the decision-maker is able to affect the value of the variable. The variables that are under the decision-maker' control are referred as decisions, and the possibilities of the decisions are referred as alternatives or actions. We denote them a. The decision-maker can choose an alternative from the set of alternatives $\mathcal{A}$. The decision-maker cannot control the state of the world. We denote the data connected to the state of the world by $x \in \Omega$, with a probability $p(x) \geq 0$ such that $\int_\Omega p(x)\mathrm{d}x = 1$. Later, the probability $p(x)$ can also be denoted as $p(x|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ denotes the model

parameter vector, if it is justified (see Section 4.2). Note, that both variables a and x can be either discrete or continuous. In our application in Articles I–IV (see Chapter 2 for details), we have a finite set of available alternatives $\mathcal{A}$ and discrete sample space Ω. More specifically, in the lake management context, the alternative is either to implement the management actions ($a = 1$) or do nothing ($a = 0$) and the lake is either in the need of management actions ($x = 1$) or not ($x = 0$). However, here we have extended the notations to a continuous variable x for the sake of generalizability.

Once the decision a and the data x are decided, a scenario is obtained. In a decision situation, there are always a number of $|\mathcal{A}| \times |\Omega|$ scenarios. Each scenario is associated with a value function $v(\cdot)$, with some decision a and the data x. The value is the realized (usually monetary) outcome of a decision that the decision-maker gains, in the presence of x. However, when we want to account for a decision-maker's attitude towards risk (see Section 3.4), the value is needed to extend to utility. The utility function is introduced in Section 3.2.

The value is most commonly measured in monetary terms, but other measures could be used as well. In fact, Howard and Abbas (2015, p. 216) list several advantages of using money as a value measure. Most importantly, money is familiar. Also, according to Howard and Abbas, money is "fungible, meaning there is no preference for any unit of this measure over another unit", and it is "divisible, meaning it can be divided into smaller units as necessary". Other measures for value would be time or saved time, or in the environment conservation, the value of ecosystem services. However, in this thesis, those alternatives are not discussed further.

3.2 Utility

The consequences of the decisions are described by the utility function, denoted $u(\cdot)$ (von Neumann and Morgenstern, 1947), where $u(\cdot)$ is twice differentiable, $u(0) = 0$, and $u'(\cdot) > 0$. The utility u takes into account the costs (monetary or other) of the experimentation as well as the consequences (monetary or other) of the selected decision. It takes the units of value as an input and returns units of utility.

There are many elicitation methods to elicitate the decision-maker's utility. Usually the methods consist of a set of questions that determine the decision-maker's attitude to the decision-making situation, and the answers are used to estimate the curve. A review of different elicitation schemes is provided by Farquhar (1984).

The most common utility functions are listed below. Also, e.g., Gerber and Pafum (1998) present them and the applications in economy.

Linear utility function

For a risk neutral decision-maker, the utility function is linear. It can be expressed in a form $u(v) = a + bv$, where a and b are constants such that $b > 0$ and v is the (monetary) value (Eidsvik et al., 2015, p. 70).

Exponential utility function

Usually, an exponential utility function is used for a risk averse decision-maker. We express it in a form $u(v) = a + b\exp(-\gamma v)$, where a and b are constants and the parameter γ is referred to as the risk aversion coefficient (see more in Section 3.4). If $\gamma > 0$, then the decision-maker is risk averse and b must be negative whereas if $\gamma < 0$, then the decision-maker is risk seeking and b must be positive. If $\gamma = 0$, the decision-maker is risk neutral and the linear utility function should be used (Eidsvik et al., 2015, p. 71). Besides the linear utility function, the exponential utility is the only one that will satisfy what is known as a delta property (see Section 3.5) and thus, has a constant risk averse function. This feature also ensures that the value of information is easier to calculate, to which we return in Section 4.1.

Logarithmic utility function

An alternative for the utility of a risk averse decision-maker is a logarithmic utility function, which has a form $u(v) = \log(v)$ (see, e.g., Eeckhoudt and Godfroid (2000)). Unlike the exponential, the logarithmic utility cannot represent risk seeking behaviour. Moreover, for a logarithmic utility function, the decision-maker must determine their initial wealth, and this type of utility function does not satisfy the delta property. As a consequence, the VOI calculation becomes more complex.

Power utility function

For a risk averse decision-maker, a power utility function has a form $u(v) = v^c$, $0 < c < 1$ (see, e.g., Abbas et al. (2013)). The smaller the parameter c is, the more risk averse the decision-maker is. If $c = 1$, the decision-maker is risk neutral, in which case the linear utility function is used.

3.3 Certain equivalent

The certain equivalent (CE, also known as a certainty equivalent) is sometimes a more approachable measure for the decision-maker's risk preferences, since, unlike the utility function, it is reported in the same units as the value (Eidsvik et al., 2015, p. 71). Generally, CE means the minimum price at which the decision-maker should sell the uncertain decision situation. If the offered price is lower,

the decision-maker should keep the situation in order to benefit from it after making the decision a. Formally, the decision-maker's CE of a decision situation is defined as

$$\begin{aligned} \mathrm{CE} &= u^{-1}\left(\max_{a\in A}\{\mathbb{E}(u(v(x,a)+w))\}\right) - w \\ &= u^{-1}\left(\max_{a\in A}\left\{\int_{\Omega} u(v(x,a)+w)p(x)\mathrm{d}x\right\}\right) - w, \end{aligned} \tag{1}$$

where u is the utility function, u^{-1} is the inverse of that, $v(x,a)$ is value of each scenario and w is the initial wealth (Eidsvik et al., 2015).

3.4 Risk aversion measures

Roughly, decision-makers can be divided into three groups according to their risk attitude. A risk neutral decision-maker should make decisions by maximizing the expected value, and pay attention only to the averages of random variables. Thus, the utility is a linear function of value. A risk averse decision-maker prefers an alternative which has a low uncertainty compared to one with a high uncertainty, even if the outcome of the latter alternative has an equal or higher expected (monetary) value. For a risk averse decision-maker, the utility function is concave. The opposite of a risk-averse decision-maker is a risk seeking decision-maker, who prefers alternatives with a high uncertainty compared to one with a low uncertainty if the expected (monetary) value of the outcome of the high uncertainty alternative is higher. Then, the utility function is convex. Figure 4 sums up the relation of the utility and the value of decision-makers with varying risk preferences, presenting examples of three utility functions presented over the values from v^0 to $v^\star$.

The Arrow-Pratt measure of absolute risk aversion (ARA) (Arrow, 1965; Pratt, 1964) is a measure for risk used in economics. It is defined as

$$\gamma(v) = -\frac{u''(v)}{u'(v)}, \tag{2}$$

where u' and u'' are the first-order and the second-order derivatives of the utility function, respectively, and v is the (monetary) value. The idea is to measure risk aversion as the second derivative of the utility function and to normalize it by the first derivative, which takes into account the magnitude of the utility function (Nadiminti et al., 1996). The greater the value of $\gamma(v)$, the larger the risk aversion. For a risk neutral decision-maker, the measure is zero (Table 1). Its unit is the reciprocal of the unit of the value measure and the same as for the utility (Howard and Abbas, 2015).

The measure is called constant absolute risk aversion (CARA), when it is a constant over all v, and is then denoted as $\gamma(v) = \gamma$. The linear and exponential utility functions are the only utility functions to meet this condition.

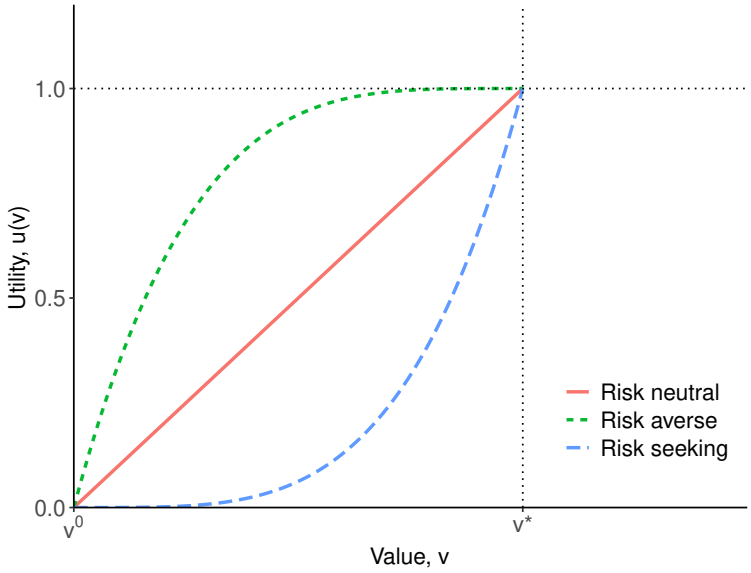


FIGURE 4 Examples of a utility function for a risk averse, a risk neutral and a risk seeking decision-maker. The presentation adapts the illustration presented by Eidsvik et al. (2015, p. 71).

If the decision-maker's risk attitude varies over v, i.e. the decision-maker changes from risk averse to risk seeking or vice versa, one should use a relative risk measure. The Arrow–Pratt measure of relative risk aversion (RRA) (Arrow, 1965; Pratt, 1964) is defined as

$$\Gamma(v) = v\gamma(v) = -\frac{vu''(v)}{u'(v)}. \tag{3}$$

Unlike the absolute risk aversion function $\gamma(v)$, the relative risk aversion function is a dimensionless quantity. Like for absolute risk aversion, the corresponding term constant relative risk aversion (CRRA) is used.

Sometimes, risk tolerance is also used (Howard and Abbas, 2015, p. 253). It means the level of risk the decision-maker is willing to take. The risk tolerance function is the inverse of the risk aversion function and is given in a form,

$$\rho(v) = \frac{1}{\gamma(v)} = -\frac{u'(v)}{u''(v)}. \tag{4}$$

The risk tolerance is expressed in the same (monetary) units as the value. For a decision-maker with the exponential risk attitude, the risk tolerance is again constant over the value v, so we denote $\rho = 1/\gamma$. Different attitudes towards risk using the risk aversion coefficient (γ) and the risk tolerance (ρ) are summarized in Table 1.

Risk-aversion function for the exponential and the logarithmic utility function

Next, we provide the computations for determining the risk-aversion function for both the exponential and logarithmic utility functions. The presentation follows the calculations outlined by Howard and Abbas (2015, p. 487). Consider an

TABLE 1 Risk relations according to (Howard and Abbas, 2015, p. 253).

	Risk preferring	Risk neutral	Risk averse
Risk aversion, γ	$\gamma < 0$	$\gamma = 0$	$\gamma > 0$
Risk tolerance, ρ	$\rho < 0$	$\rho = \infty$	$\rho > 0$

exponential utility function, $u(v) = -\exp(-\gamma v)$. We have $u'(v) = \gamma \exp(-\gamma v)$ and $u''(v) = -\gamma^2 \exp(-\gamma v)$. Then,

$$\gamma(v) = -\frac{u''(v)}{u'(v)} = -\frac{-\gamma^2 \exp(-\gamma v)}{\gamma \exp(-\gamma v)} = \gamma$$

and

$$\rho(v) = \frac{1}{\gamma(v)} = \frac{1}{\gamma}.$$

Therefore, both the risk-aversion function and the risk tolerance function are constant with respect to value v for an exponential utility function.

Next, consider a logarithmic utility function $u(v) = \log(v + w)$. For a logarithmic utility, we have,

$$u'(v) = \frac{1}{v+w} \quad \text{and} \quad u''(v) = -\frac{1}{(v+w)^2}.$$

These lead to

$$\gamma(v) = \frac{1}{v+w},$$

and

$$\rho(v) = \frac{1}{\gamma(v)} = v + w.$$

The risk-aversion function decreases linearly as the value v increases, and the risk tolerance increases linearly as the value v increases.

3.5 Delta property

In decision analysis, it is useful if the decision-maker's utility function satisfies the delta property, which happens when they have an exponential utility function. In fact, the delta property is routinely assumed both in the literature and in practice (Eidsvik et al., 2015, p. 75). This may seem restrictive at first, but in practice, the exponential utility function is usually sufficient to describe the decision-maker's risk attitude. In addition, an exponential utility function is accurate for many decision situations since it can effectively approximate many utility functions (Kirkwood, 2004). That is why, in this section, we explore the delta property in more detail.

In brief, by the delta property we mean that if a constant value Δ was added to all the prospects in a situation with an uncertain value, then the decision-maker's CE would increase by Δ (Eidsvik et al., 2015, p. 72). Sometimes in the literature, such a person is referred to as a deltaperson (Howard and Abbas, 2015). Formally, the utility function needs to satisfy

$$u^{-1}(\mathbb{E}(u(v(x,a)+\Delta))) = u^{-1}(\mathbb{E}(u(v(x,a)))) + \Delta, \tag{5}$$

for all values x and a, with any constant value Δ (Eidsvik et al., 2015). The delta property makes the decision-maker's CE independent of the person's initial wealth, which we denoted by w. It allows us to remove the initial wealth from the VOI calculations, because its actual value does not matter and it can be considered zero. We will return to this result in Section 4.1.

4 SUBSAMPLE SELECTION CRITERIA

In this thesis, the major goal was to discover, what kind of information and how much of it should be collected so that the resources invested in environmental management decision-making are optimally used. In fact, the decision-making situation in this context is two-phased: first, it is determined whether it is worth acquiring additional data, and then the actual decision leading to further measures is made, for example a decision on the restoration of the lake. With the help of statistical methods, we are able to measure the information contained in the data even before the data is collected. In this chapter, we present two alternative criteria that aid to make a decision on data acquisition. In Section 4.1, we consider the value of information, which is a concept of the decision-making theory. This section is also primarily based on the book by Eidsvik et al. (2015). Value of information is used in Articles I, IV, and III. In Section 4.2, we discuss the D-optimality criterion, which is a concept of the theory of optimal design and used particularly in the design of experiments. The D-optimality criterion is used in Article II in the case of cost-efficient planning of an observational study.

4.1 Value of information

Value of information (VOI) is a concept of decision theory to assess the value of additional information before it is actually collected (Eidsvik et al., 2015). Descriptively, VOI means the price at which the decision-maker is indifferent between purchasing additional information that helps in uncertain decision-making, and not purchasing additional information, which consequently means making the decision only with the information available at that moment. In other words, it is the maximum price, that the decision-maker should still pay for additional information to make a decision. The VOI is compared to the cost of the data: if the cost of data collection exceeds the VOI, the additional data should not be collected. Respectively, if the cost is lower than the VOI, the decision-maker should collect the additional data so that it would aid in the decision-making.

We continue with the notations introduced in Chapter 3, and compute the VOI in a decision situation with one variable $x \in \Omega$ and one decision $a \in \mathcal{A}$, and the utility $u(\cdot)$ describing the decision-maker's risk preferences. Two expected utilities, with and without the additional information, need to be equated. Based on the rules of the decision theory, the decision-maker's objective is to always choose the alternative that maximizes their expected utility. If the chosen alternative is the alternative a, then they gain the value $v(x,a) + w$, where w is the decision-maker's initial wealth. The maximum expected utility between the alternatives is

$$\max_{a \in A} \left\{ \int_\Omega u(v(x,a) + w) p(x) \mathrm{d}x \right\}, \tag{6}$$

where the variable x is observed with probability $p(x)$.

Next, we calculate the expected utility, when the information is available. The decision-maker pays the price $v^\star$ to get the information, thus, they gain the value $v(x,a) + w - v^\star$. However, the decision-maker does not know how the uncertainty will resolve until the decision is made. The expected utility when the information is available is then

$$\int_\Omega \max_{a \in A} \{u(v(x,a) + w - v^\star)\} p(x) \mathrm{d}x. \tag{7}$$

Formally, the value of (perfect) information is the price $v^\star$ at which the expected utilities in Equations (6) and (7) are equal:

$$\int_\Omega \max_{a \in A} \{u(v(x,a) + w - v^\star)\} p(x) \mathrm{d}x = \max_{a \in A} \left\{ \int_\Omega u(v(x,a) + w) p(x) \mathrm{d}x \right\}. \tag{8}$$

Equation (8) comes from (Eidsvik et al., 2015). VOI can be calculated from the equation by iteratively varying $v^\star$ until it is satisfied. A unique solution always exists because u is a strictly increasing function.

Equation (8) is a general definition, but it can also be presented in an easier form if the decision-maker's utility function satisfies the delta property. Taking the utility function's inverse and using Equation (5), Equation (8) becomes

$$\begin{aligned} u^{-1} \left(\int_\Omega \max_{a \in A} \{u(v(x,a))\} p(x) \mathrm{d}x \right) + w - v^\star \\ = u^{-1} \left(\max_{a \in A} \left\{ \int_\Omega u(v(x,a)) p(x) \mathrm{d}x \right\} \right) + w, \end{aligned}$$

where the initial wealth w can be omitted. Then, denoting $v^\star = \mathrm{VOI}(x)$, the value of (perfect) information can be expressed as a difference between two certain equivalents:

$$\begin{aligned} \mathrm{VOI}(x) = & u^{-1} \left(\int_\Omega \max_{a \in A} \{u(v(x,a))\} p(x) \mathrm{d}x \right) \\ & - u^{-1} \left(\max_{a \in A} \left\{ \int_\Omega u(v(x,a)) p(x) \mathrm{d}x \right\} \right). \end{aligned} \tag{9}$$

The first part of the sum can be understood as the decision-maker's CE of the situation with information when it is available for free, and the latter is the decision-maker's CE of the decision situation without information. The expression is now independent of the initial wealth w. Denoting $\mathrm{VOI}(x)$ implies that the VOI is calculated for perfect information x (Eidsvik et al., 2015). Since the delta assumption is so often made, the following notations are used for the certain equivalents in Equation (9):

$$\mathrm{VOI}(x) = \mathrm{PoV}(x) - \mathrm{PV}.$$

Above, PV (prior value) is a priori the maximum expected utility of all expected utilities, given all available alternatives, meanwhile, $\mathrm{PoV}(x)$ (posterior value) is the updated expected utility after additional information is gained.

In Equations (8) and (9), it is assumed that we obtain perfect knowledge about the state of x when gathering the data. However, in many cases that is not possible, but the decision-makers need to settle for imperfect data that indicates the state of x. Assume that we observe the value y of a continuous random variable with the density $p(y)$, which does not give a certain knowledge, but is only reflecting the state of x. In our application, it is the chlorophyll-*a* concentration of the lake. The value of imperfect information is the price $v^\star$, such that

$$\begin{aligned}\int_y \max_{a\in A}\left\{\int_\Omega u(v(x,a)+w-v^\star)p(x|y)\mathrm{d}x\right\}p(y)\mathrm{d}y \\ = \max_{a\in A}\left\{\int_\Omega u(v(x,a)+w)p(x)\mathrm{d}x\right\},\end{aligned} \tag{10}$$

where $p(x|y)$ is the posterior distribution of x given the uncertainty y (Eidsvik et al., 2015). Again, VOI can be calculated from the equation by iteratively varying the price until it is satisfied. Similarly as in the case of perfect information, if the decision-maker's utility function satisfies the delta property, VOI can be expressed more simply as a difference

$$\begin{aligned}\mathrm{VOI}(y) =& u^{-1}\left(\int_y \max_{a\in A}\left\{\int_\Omega u(v(x,a))p(x|y)\mathrm{d}x\right\}p(y)\mathrm{d}y\right) \\ &- u^{-1}\left(\max_{a\in A}\left\{\int_\Omega u(v(x,a))p(x)\mathrm{d}x\right\}\right).\end{aligned} \tag{11}$$

Note that the Equation (11) is again independent of the initial wealth w. Again, the following notations for the prior and posterior values are used:

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

Denoting $\mathrm{VOI}(y)$ implies that VOI is calculated for an imperfect information (Eidsvik et al., 2015).

In Articles I and IV, we calculate the VOI of a single lake, while in Article III, we want to calculate the VOI of a design D with multiple, spatially correlated observation points, which are in this case lakes. In Article III, the aim is to find a

design $D \in \mathcal{D}$, where $\mathcal{D}$ is defined as in Equation (22) in Section 5.1, that maximize VOI. Formally, this involves solving the following optimization problem:

$$D^\star = \arg\max_{D \in \mathcal{D}} \{\text{VOI}(D)\}, \quad \text{VOI}(D) > P(D), \tag{12}$$

where $\text{VOI}(D)$ is the VOI of the design D as in Equation (9) and $P(D)$ is the cost of gathering the data from the design D.

4.2 D-optimality

Optimal experimental designs are commonly derived using the alphabet criteria (Atkinson et al., 2007; Lawson, 2015). The goal is to find a design that maximizes (or minimizes) an optimality criterion that usually is a function of an information matrix (Fedorov, 1972; Pukelsheim and Torsney, 1991). The most popular criterion is D-optimality, which is equivalent to maximising the determinant of the information matrix.

Next, we will discuss the information matrices of generalized linear models to find a D-optimal design. The observed information matrix can be used to indicate how much information is contained in the data. It can be calculated, when the data is observed. Having a $(J \times 1)$ parameter vector $\boldsymbol{\theta} \in \Theta$, the observed information is a symmetric $(J \times J)$ matrix and for a generalized linear model, it is defined as minus the second derivative of the log-likelihood function $l(\boldsymbol{x}|\boldsymbol{\theta})$, where $\boldsymbol{x} = (x_1, \ldots, x_N)^\top$, or minus the slope of the score function $\boldsymbol{S}(\boldsymbol{\theta})$:

$$\mathcal{J}(\boldsymbol{\theta}) = -\sum_{i=1}^{N} \left(\frac{\partial^2 l_i(x_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right) = -\sum_{i=1}^{N} \left(\frac{\partial \boldsymbol{S}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right). \tag{13}$$

Before the data is gathered, the observed information cannot be obtained. However, we can calculate the expectation of the observed information. The expected information matrix (also called Fisher information matrix) informs, how much information the data is expected to contain. It is used when the data collection is still planned. It is also used in computational methods (e.g. Fisher scoring algorithm), instead of the observed information. Formally, when the parameter $\boldsymbol{\theta}$ is a $(J \times 1)$ vector, the expected information matrix of a generalized linear model is a symmetric $(J \times J)$ matrix:

$$\mathcal{I}(\boldsymbol{\theta}) = -\sum_{i=1}^{N} \mathbb{E} \left(\frac{\partial^2 l_i(x_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right). \tag{14}$$

It can be proved, that under regularity conditions, the following holds:

$$\begin{aligned} \mathcal{I}(\boldsymbol{\theta}) &= \sum_{i=1}^{N} \mathbb{E} \left(\frac{\partial l_i(x_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^2 = \sum_{i=1}^{N} \mathbb{E} \left(\frac{\partial l_i(x_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right) \left(\frac{\partial l_i(x_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^\top \\ &= -\sum_{i=1}^{N} \mathbb{E} \left(\frac{\partial^2 l_i(x_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right). \end{aligned}$$

The regularity conditions are satisfied by the exponential family or other commonly used models. See Davison (2003, sec. 4.4.2) for details. The above also implies that the expected information is the variance matrix of the score function:

$$\mathcal{I}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \mathrm{Cov}(\boldsymbol{S}_i(\boldsymbol{\theta})).$$

The most important difference between the expected and the observed information is that the expected information is the function of a parameter $\boldsymbol{\theta}$, as the observed information is the function of $\boldsymbol{\theta}$ and the observations x_i, $i = 1, \ldots, N$. However, the observed information is better understood as a single value or a single statistic, rather than as a function (Pawitan, 2001, p. 216). Both information matrices introduced above can be applied in the alphabet statistical criteria (Atkinson et al., 2007).

In Article II, the optimization problem corresponding to Equation (12) is formulated adapting the presentation of Chaloner and Verdinelli (1995). First, we recall the necessary notations. As in Section 3.1, data $\boldsymbol{x}$ from a sample space Ω will be observed once the design is selected. Based on the observed data, an alternative a must be chosen from all the possible alternatives $\mathcal{A}$. As in the Bayesian framework, a model $p(\boldsymbol{x}|D, \boldsymbol{\theta}) \geq 0$ is assumed for data, such that $\int_\Omega p(\boldsymbol{x}|D, \boldsymbol{\theta})\mathrm{d}x = 1$, where the model parameters $\boldsymbol{\theta}$ are defined in parameter space Θ. The prior distribution for the model parameters is denoted by $p(\boldsymbol{\theta})$. The posterior probability distribution $p(\boldsymbol{\theta}|\boldsymbol{x}, D)$ is proportional to the product $p(\boldsymbol{x}|D, \boldsymbol{\theta})p(\boldsymbol{\theta})$ and defines the current knowledge of the model parameters. The aim is to find a design $D^\star$ that maximizes the logarithm of the determinant of the information matrix:

$$\begin{aligned} D^\star &= \arg\max_{D \in \mathcal{D}} \int_\Omega \int_\Theta \log\det(\mathcal{I}(\boldsymbol{\theta}))p(\boldsymbol{\theta}, \boldsymbol{x}|D)\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{x} \\ &= \arg\max_{D \in \mathcal{D}} \int_\Omega \left[\int_\Theta \log\det(\mathcal{I}(\boldsymbol{\theta}))p(\boldsymbol{\theta}|\boldsymbol{x}, D)\mathrm{d}\boldsymbol{\theta} \right] p(\boldsymbol{x}|D)\mathrm{d}\boldsymbol{x}, \end{aligned} \tag{15}$$

where $\mathcal{I}(\cdot)$ is the Fisher information matrix as in Equation (14). The integrals average over what is unknown: data $\boldsymbol{x}$ have not yet been observed and for the model parameters $\boldsymbol{\theta}$ only a prior distribution is assumed.

Information matrices for a binary model

In the rest of this section, we derive the information matrices in the case of a binary regression model, as in Article II. First, assume that a binary response $x_i \in \{0, 1\}$, $i = 1, \ldots, N$, is distributed as

$$\begin{aligned} P(x_i = 1|\boldsymbol{z}_i) &= \pi_i, \\ P(x_i = 0|\boldsymbol{z}_i) &= 1 - \pi_i, \end{aligned} \tag{16}$$

where the parameter π_i is linked to the covariates $\boldsymbol{z}_i$ with a link function $g(\pi_i) = \eta_i$ as

$$\begin{aligned} g(\pi_i) = \text{logit}(\pi_i) &= \log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i \\ \pi_i &= \frac{\exp(\eta_i)}{1+\exp(\eta_i)} \\ (1-\pi_i) &= \frac{1}{1+\exp(\eta_i)}. \end{aligned} \tag{17}$$

The response x_i is modelled with a linear predictor $\eta_i = \boldsymbol{z}_i^\top \boldsymbol{\beta}$, where $\boldsymbol{z}_i$ is the ith row of a matrix $\boldsymbol{Z}$ including covariates. The parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)^\top$ is unknown and needed to be estimated. The expectation of the data is $\mathbb{E}(x_i) = \mu_i = \pi_i$ and the variance $\text{Var}(x_i) = \pi_i(1-\pi_i)$. Assuming conditional independence in Equation (17), the likelihood of data $\boldsymbol{x} = (x_1, \ldots, x_N)^\top$ is obtained by

$$p(\boldsymbol{x} \mid \boldsymbol{\beta}) = L(\boldsymbol{\beta}) = \prod_{i=1}^{N} L_i(\boldsymbol{\beta}) = \prod_{i=1}^{N} \pi_i^{x_i}(1-\pi_i)^{1-x_i} \tag{18}$$

and log-likelihood by

$$\begin{aligned} \log(p(\boldsymbol{x} \mid \boldsymbol{\beta})) = \log L(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) &= \sum_{i=1}^{N} l_i(\boldsymbol{\beta}) \\ &= \sum_{i=1}^{N} x_i \log(\pi_i) + (1-x_i)\log(1-\pi_i) \\ &= \sum_{i=1}^{N} x_i(\boldsymbol{z}_i^\top \boldsymbol{\beta}) - \log(1+\exp(\boldsymbol{z}_i^\top \boldsymbol{\beta})). \end{aligned} \tag{19}$$

A score function is defined as a gradient of log-likelihood. In the logistic regression situation, it is a $(J \times 1)$ vector obtained by

$$\begin{aligned} \boldsymbol{S}(\boldsymbol{\beta}) &= \sum_{i=1}^{N} \boldsymbol{S}_i(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{\partial l_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\ &= \sum_{i=1}^{N} \left(x_i - \frac{\exp(\eta_i)}{1+\exp(\eta_i)}\right) \boldsymbol{z}_i = \sum_{i=1}^{N} (x_i - \pi_i)\boldsymbol{z}_i. \end{aligned} \tag{20}$$

For a binary response and the logit link, based on the chain rule, the observed information matrix in Equation (13) can be expressed as

$$\begin{aligned} \mathcal{J}(\boldsymbol{\beta}) &= -\sum_{i=1}^{N} \left(\frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{S}_i(\boldsymbol{\beta})\right) = \sum_{i=1}^{N} \frac{\partial}{\partial \boldsymbol{\beta}} (\pi_i - x_i)\boldsymbol{z}_i \\ &= \sum_{i=1}^{N} \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{z}_i \pi_i = \sum_{i=1}^{N} \boldsymbol{z}_i \frac{\partial \pi_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \boldsymbol{\beta}}. \end{aligned}$$

We know, that

$$\frac{\partial \pi_i}{\partial \eta_i} = \frac{\partial}{\partial \eta_i} \frac{\exp(\eta_i)}{1+\exp(\eta_i)} = \frac{(1+\exp(\eta_i))\exp(\eta_i) - \exp(\eta_i)\exp(\eta_i)}{(1+\exp(\eta_i))^2} = \pi_i(1-\pi_i)$$

and

$$\frac{\partial \eta_i}{\partial \boldsymbol{\beta}} = \frac{\partial \boldsymbol{z}_i^\top \boldsymbol{\beta}}{\partial \boldsymbol{\beta}} = \boldsymbol{z}_i.$$

Thus, the (r, s) element of the observed information matrix takes the form

$$\begin{aligned} \frac{\partial^2 l_i}{\partial \beta_r \partial \beta_s} &= \sum_{i=1}^{N} [\pi_i(1-\pi_i)] z_{ir} z_{is} \\ &= \sum_{i=1}^{N} \left[\frac{\exp(\eta_i)}{1+\exp(\eta_i)} \left(1 - \frac{\exp(\eta_i)}{1+\exp(\eta_i)} \right) \right] z_{ir} z_{is}. \end{aligned} \tag{21}$$

Assuming a binary response and a logit link, taking the expectation of a parameter $\boldsymbol{\beta}$, the expression in Equation (21) remains unchanged. Thus, the expected information matrix is the same as the observed information matrix.

5 COMPUTATIONAL METHODS

In this section, we briefly discuss the main computational methods used in this thesis. The previous chapter introduced possible subsample selection criteria. In Section 5.1, we discuss the selection methods, i.e. methods to solve the optimization problem in Equations (12) and (15). In Section 5.2, we introduce Gaussian processes, to give the reader a basic understanding of them, as they are used in the VOI approximation discussed in Article III. Both topics being broad, we keep the presentations compact.

5.1 Selection of design

Unlike in experimental design, we assume that there is a finite set of available observation points. In this context, possible designs include an empty set, all single observation points, all sets of two, sets of three, etc., up to the design that includes all N available observation points. Formally, if we denote the N observation points by $s_1, \ldots, s_N$, the overall set of designs is denoted as $\mathcal{D} = \bigcup_{i=0}^{N} \mathcal{D}_i$, where

$$\begin{aligned} \mathcal{D}_0 &= \emptyset, \\ \mathcal{D}_1 &= \{(s_1), (s_2), \ldots, (s_N)\}, \\ \mathcal{D}_2 &= \{(s_1, s_2), (s_1, s_3), \ldots, (s_{N-1}, s_N)\}, \\ &\vdots \\ \mathcal{D}_N &= \{(s_1, s_2, \ldots, s_N)\}. \end{aligned} \tag{22}$$

If there are no constrains, there are 2^N possible designs to find an approximately optimal design D. In fact, the subsample selection problem is NP-hard (Welch, 1982), so heuristic optimizing methods are needed.

In Articles II and III, we use a greedy selection to find approximately optimal designs in an observational study. The greedy approach (also, a sequential search (Dykstra, 1971)), is familiar to mathematicians and computer scientists. In

general, a greedy method refers to a selection method in which sequential choices are made and the option that seems best at that moment is always chosen. The choices made cannot be canceled later. In subsample selection, the idea is to select design points one by one, always optimizing the selection criterion. Usually, a greedy strategy does not lead to the globally optimal solution, but it can produce a locally optimal solution that approximates the globally optimal solution (Yang et al., 2014). In addition, the approach is usually fast. Other methods, which we test in Article III, are a randomized exchange algorithm and an algorithm based on Bayesian optimization.

5.2 Gaussian processes

Gaussian processes (GP) are a flexible and powerful probabilistic framework commonly used in statistics for modeling and predicting complex, non-linear relationships in data (Rasmussen and Williams, 2006). A Gaussian process is an extension of the multivariate Gaussian distribution to an infinite dimension stochastic process (a collection of random variables indexed by time or space) for which any finite combination of dimensions will be a Gaussian distribution. As a Gaussian distribution is a distribution over a random variable that is fully defined in terms of its mean and covariance, a Gaussian process is a distribution over functions that is fully defined in terms of its mean function and covariance function (Brochu et al., 2010). The primary appeal of Gaussian processes lies in their nonparametric nature, which means they can adapt to data complexity without requiring a pre-specified functional form. This flexibility makes Gaussian processes an ideal choice for applications involving uncertain or sparse data, especially in environmental statistics.

The spatial regression model is based on Gaussian variables and linear relationships. It is likely the most widely used model in spatial statistics (see, e.g., Cressie (1993); Banerjee et al. (2004); Eidsvik et al. (2015)). In Article III, we model the data with a Bayesian latent spatial logistic model, which extends the standard regression model by accounting for the spatially correlated error terms. It is assumed that the values of response are spatially correlated, so ignoring a spatially structured error term gives biased estimates. Formally, the model for the binary response $x(s_i)$ at a site s_i is

$$\begin{gathered} P(x(s_i) = 1|\boldsymbol{z}(s_i)) = \pi(s_i), \quad P(x(s_i) = 0|\boldsymbol{z}(s_i)) = 1 - \pi(s_i), \\ \text{logit}(\pi(s_i)) = \boldsymbol{z}(s_i)^\top \boldsymbol{\beta} + w(s_i), \end{gathered} \tag{23}$$

where $\boldsymbol{z}(s_i)$, $i = 1, \ldots, N$, is a $(J \times 1)$ vector of known covariates at site s_i and the parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)^\top$ is unknown and needed to be estimated, as in Equation (17). In addition, the spatial effects $\boldsymbol{w} = (w(s_1), \ldots, w(s_N))^\top$ are represented by a Gaussian process model with a zero mean $\mathbb{E}(\boldsymbol{w}) = \boldsymbol{0}$ and a covariance $\text{Cov}(w(s_t), w(s_k)) = \sigma^2 \text{Corr}(w(s_t), w(s_k))$, with a variance $\text{Var}(w) = \sigma^2$. A popular choice for a correlation structure is a Matern correlation function

such that $\text{Corr}(w(s_t), w(s_k)) = (1 + \phi h_{tk}) \exp(-\phi h_{tk})$, where h_{tk} is the great-circle distance between two sites s_t and s_k. Other examples of common kernels are the exponential kernel of the form $\text{Corr}(w(s_t), w(s_k)) = \exp(-\phi h_{tk})$ and a Gaussian kernel of the form $\text{Corr}(w(s_t), w(s_k)) = \exp(-\phi^2 h_{tk}^2)$.

6 RESEARCH CONTRIBUTION

This chapter summarizes the research contribution of each article in order from Article I to Article IV.

In Article I, we use VOI analysis to assess the cost-efficiency of acquiring additional lake monitoring data. The article fills the gap in the literature regarding real-world applications of VOI analysis in environmental monitoring data and proposes a method to enhance its more frequent implementation. More specifically, we calculate the VOI in the case of two ecological status classes based on whether the lake needs management actions or not, and two decisions of whether or not to implement the management actions. We form an analytical solution for the value of perfect information, and propose an estimation for the value of imperfect information, which is based on a Monte Carlo type method using empirical data of chlorophyll-*a* concentration. A similar approach was later used in their research by Luhede et al. (2024). In addition, we evaluate the uncertainty of the value of imperfect information with confidence intervals, which we estimate with the parametric percentile bootstrap method (Efron and Tibshirani, 1993). The monetary values of the ecological status of the lake are derived from a valuation study by Ahtiainen (2008) (see Section 2.2). Since monetary values are the most uncertain assumption, we also conduct a sensitivity analysis to study the effect of different monetary values on value of perfect, as well as imperfect information. For additional comparison, three different priors for ecological status are also employed.

The results show that generally, the VOI exceeds the cost of the chlorophyll-*a* data gathering. When comparing the realized monitoring costs to the estimated VOI, the costs are significantly lower, which makes them profitable to invest. It is particularly profitable to monitor lakes that are assumed to be in good condition based on prior information, in order to avoid expensive and unnecessary management actions.

In Article II, we focus on a subsample selection problem and apply it to the Finnish lake management setting. Unlike in other articles, we do not address VOI, but use the terms of the theory of optimal design. The research question is to find a design with which the fitted Bayesian logistic regression model predicting

the status of the lake estimates the model parameters as accurately as possible. The most familiar approach for sample selection is random sampling. However, when the data collection is time-consuming and costly, tools for the optimal design of a data collection are required to produce a data set with higher expected information per unit than obtained from a random sample. Since the prior information is poor, the initial model parameters may not be sufficient to describe the phenomenon properly. Therefore, we apply a Bayesian two-stage selection strategy where the selection of lakes to be measured at the second stage depends on the selection performed at the first stage. This kind of two-stage selection strategy is a compromise that aims to repair from poor prior information while still keeping the strategy relatively straightforward, compared to a sequential approach. For subsample selection in both stages, we use a greedy selection. The proposed selection method is based on the Fisher information matrix presented in Section 4.2.

The results show that the two-stage strategy has a modest advantage over the single-stage strategy. There appear to be no substantial differences in model parameter estimates when comparing the two-stage strategy to the single-stage strategy.

In Article III, we again consider a subsample selection problem in lake management and return to VOI analysis. Compared to Article I, we are interested in the VOI of multiple lakes in a spatially correlated situation instead of one lake. Compared to Article II, the research question is similar, but the selection criterion to achieve an optimal subsample differs. The aim is to find subsamples with high VOI and compare them to the cost of collecting data from that subsample. To solve this optimization problem, we use various heuristic algorithms: a greedy forward algorithm as in Article II, a randomized exchange algorithm and an algorithm based on Bayesian optimization. VOI calculations apply closed-form approximations by Evangelou and Eidsvik (2017), which are based on hierarchical general linear models. This enable fast VOI evaluation for each design. We show how large designs and what kind of designs can be selected compared to the costs. Finally, we compare the selected samples to samples selected with simpler criteria, forgetting the statistical models and decision-analytic perspectives.

In general, VOI analysis suggests that it is profitable to collect data from the lakes to observe the ecological status. In terms of VOI, good subsamples usually consist of lakes whose status is difficult to determine based on prior knowledge. Also, good subsamples aim for geographic coverage. The subsamples achieved by forward selection give reasonably large VOI, but they can still be outperformed with the randomized exchange algorithm and the algorithm based on Bayesian optimization. Moreover, the designs found by statistical approaches have much higher VOI than that of simpler selection criteria. Therefore, the study suggests that policy-makers use statistical methods in the design selection.

In Article IV, we rely on the setting of Article I and study more general risk preferences of the decision-maker. The study discusses the effect of decision-maker's risk aversion on the value of information in the context of the lake management, which, as far as we know, is lacking in the previous literature. The risk

aversion is more discussed in Section 3.4. We calculate the value of perfect and imperfect information for a risk neutral and a risk averse decision-maker. Particularly, the value of imperfect information from the standpoint of a risk averse decision-maker is a topic that seems to be inadequately addressed in earlier research. The degree of risk aversion is displayed through utility function, which in our case are an exponential utility and a power utility function.

Our results give the evidence that VOI is strongly dependent on the degree of risk aversion. A risk averse decision-maker's VOI may be lower or higher than a risk neutral decision-maker's VOI, depending on the prior probability of the lake status and the cost of the lake management actions. This suggests that much of the analysis may be overlooked if a simple assumption about a risk neutral decision-maker is made. Trying two different utility functions yields comparable results. Compared to the results of Article I, these results give even more evidence that the lake monitoring is cost-effective.

7 CONCLUSION

This thesis consists of four articles, three of which discuss VOI (Articles I, III and IV) and two of which discuss the subsample selection techniques (Articles II and III). The two methods are applied to the real-life lake monitoring data. Firstly, we applied the concepts of the value of perfect as well as imperfect information, and calculated VOI for a single example lake, as well as a spatial design with multiple lakes on it. In addition, we considered the decision-maker's different risk attitudes in VOI analysis. Secondly, we demonstrated approximate optimal subsample selection methods in the context of lake management. We considered two selection criteria, VOI and D-optimality criteria. The VOI criterion assesses the profitability of designs, accounting for the costs and benefits of monitoring and management actions and the associated uncertainty, while the D-optimality criterion aims to find a design that estimates the parameters of a model predicting the ecological status of a lake as precisely as possible.

This thesis has answered the questions related to the lake monitoring presented in the Introduction. It has shown that VOI analysis framework can be successfully applied to the lake monitoring data to assess the value of environmental monitoring. The lake designs selected with statistical methods clearly outperform designs made based on simpler criteria. However, a two-stage selection strategy may have only a modest advantage in this context. From the point of view of environmental management, the main results indicate that the monitoring is cost-efficient, particularly when the ecological status is initially assumed to be excellent or good, and the value may even increase if the decision-maker is risk averse. Monitoring efforts should prioritize lakes that are presumed, based on prior information, to be in excellent or good status, as well as those with uncertain status near the boundary between good and moderate class. It is recommended to gather additional data to confirm the status before making restoration decisions, thereby avoiding potentially unnecessary and costly restoration actions. Conversely, lakes in a priori weak status should be directly targeted with management actions. In such cases, monitoring resources should be allocated to assessing the impacts of management actions rather than evaluating the current status.

In the earlier literature, VOI analysis has been rarely used to demonstrate the value of environmental monitoring, although there has been a need for such analysis. As far as we know, this thesis is one of the first attempts to apply VOI framework to the real environmental monitoring data. It provides a perspective to the new studies to utilize a similar framework and tools in similar applications.

One difficulty that prevents the analysis, is that there is a lot of uncertainty associated with the estimated monetary value of the ecological status. In this thesis, we build on a real valuation study of a Finnish lake. We recognize that the results depend strongly on this decision. Considering that, we also performed a sensitivity analysis to evaluate the effect of monetary value on VOI in Article I.

Another decision that restricts our results is, that we chose to use the chlorophyll-*a* data as an ecological indicator. In reality, the overall ecological status is determined by the information obtained from many different indicators, with one of the most important being chlorophyll-*a* concentration. The data collection costs that are utilized to be compared to VOI, are based on this selection. In reality, the ecological status is based on several indicator variables representing many quality factors describing the biotic structure of a lake, however, with chlorophyll being the most important factor.

Also, relatively little research exists on the application of methods developed for the optimal design of experiments to the design of observational studies. This thesis continues to develop and apply the existing subsample selecting methods in a context of a real-life practical application. The same methodologies can be used for other applications as well.

The present thesis was focused on analyzing traditional water sampling data. An interesting research direction in environmental monitoring would be analyzing the data from different sources, for instance, remote sensing, biomonitoring, continuous water quality sensors and large-scale wireless sensor networks (Kotamäki et al., 2009; Gong et al., 2022). We considered the spatial dimension of the sampling data, but an interesting direction would also be to take into account the temporal dimension (see, e.g., Vanhatalo et al. (2021)). In our VOI analysis, we considered monetary value as a criterion to be optimized. However, in the lake management situation, we could be interested in optimizing for both monetary and biodiversity criteria, which leads to a multiple criteria optimization problem (see, e.g., Eyvindson et al. (2019)).

REFERENCES

Abbas, A. E., Bakır, N. O., Klutke, G.-A., and Sun, Z. (2013). Effects of risk aversion on the value of information in two-action decision problems. *Decision Analysis*, 10(3):257–275. DOI: 10.1287/deca.2013.0275.

Ahtiainen, H. (2008). *Järven tilan parantamisen hyödyt. Esimerkkinä Hiidenvesi.* Finnish Environment Institute (SYKE), Helsinki. (In Finnish.) Available online at: https://helda.helsinki.fi/handle/10138/38353 (Accessed May 22th, 2019).

Aroviita, J., Hellsten, S., Jyväsjärvi, J., Järvenpää, L., Järvinen, M., Karjalainen, S. M., Kauppila, P., Keto, A., Kuoppala, M., Manni, K., Mannio, J., Mitikka, S., Olin, M., Perus, J., Pilke, A., Rask, M., Riihimäki, J., Ruuskanen, A., Siimes, K., Sutela, T., Vehanen, T., and Vuori, K.-M. (2012). *Ohje pintavesien ekologisen ja kemiallisen tilan luokitteluun vuosille 2012–2013 – päivitetyt arviointiperusteet ja niiden soveltaminen*. Suomen ympäristökeskus (SYKE), Helsinki. (In Finnish.) Available online at: http://hdl.handle.net/10138/41788 (Accessed August 16th, 2024).

Aroviita, J., Mitikka, S., and Vienonen, S. (2019). *Pintavesien tilan luokittelu ja arviointiperusteet vesienhoidon kolmannella kaudella.* Finnish Environment Institute (SYKE), Helsinki. (In Finnish.) Available online at: https://helda.helsinki.fi/handle/10138/306745 (Accessed February 17th, 2020).

Arrow, K. (1965). *Aspects of the theory of risk-bearing*. Yrjö Jahnssonin Säätiö. Reprinted in: Essays in the Theory of Risk Bearing, Markham Publ. Co., Chicago, 1971, 90–109.

Atkinson, A. C., Donev, A. N., and Tobias, R. D. (2007). *Optimum experimental design with SAS*. Oxford University Press. DOI: 10.1093/oso/9780199296590.003.0013.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Hierarchical modeling and analysis for spatial data*. Monographs on statistics and applied probability. Chapman & Hall, Boca Raton, FL.

Bickel, J. E. (2008). The relationship between perfect and imperfect information in a two-action risk-sensitive problem. *Decision Analysis*, 5(3):116–128. DOI: 10.1287/deca.1080.0118.

Bolam, F. C., Grainger, M. J., Mengersen, K. L., Stewart, G. B., Sutherland, W. J., Runge, M. C., and McGowan, P. J. K. (2019). Using the value of information to improve conservation decision making. *Biological Reviews*, 94(2):629–647. DOI: 10.1111/brv.12471.

Bouma, J., van der Woerd, H., and Kuik, O. (2009). Assessing the value of information for water quality management in the North Sea. *Journal of Environmental*

Management, 90(2):1280 – 1288. DOI: https://doi.org/10.1016/j.jenvman.2008.07.016.

Brochu, E., Cora, V. M., and de Freitas, N. (2010). A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. Preprint. Available at: https://arxiv.org/abs/1012.2599.

Brown, J., Heuvelink, G., and Refsgaard, J. (2005). An integrated methodology for recording uncertainties about environmental data. *Water science and technology: a journal of the International Association on Water Pollution Research*, 52:153–60. DOI: 10.2166/wst.2005.0163.

Canessa, S., Guillera-Arroita, G., Lahoz-Monfort, J. J., Southwell, D. M., Armstrong, D. P., Chadès, I., Lacy, R. C., and Converse, S. J. (2015). When do we need more data? A primer on calculating the value of information for applied ecologists. *Methods in Ecology and Evolution*, 6(10):1219–1228. DOI: https://doi.org/10.1111/2041-210X.12423.

Carpenter, S. R., Caraco, N. F., Correll, D. L., Howarth, R. W., Sharpley, A. N., and Smith, V. H. (1998). Nonpoint pollution of surface waters with phosphorus and nitrogen. *Ecological Applications*, 8(3):559–568. DOI: 10.1890/1051-0761(1998)008[0559:NPOSWW]2.0.CO;2.

Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, 10(3):273 – 304. DOI: 10.1214/ss/1177009939.

Colyvan, M. (2016). Value of information and monitoring in conservation biology. *Environment Systems and Decisions*, 36(3):302–309. DOI: 10.1007/s10669-016-9603-8.

Cressie, N. (1993). *Statistics for spatial data*. John Wiley & Sons, Incorporated.

Davies, G. B. and Satchell, S. E. (2007). The behavioural components of risk aversion. *Journal of Mathematical Psychology*, 51(1):1–13. DOI: https://doi.org/10.1016/j.jmp.2006.10.003.

Davison, A. C. (2003). *Statistical models*. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511815850.

Delquié, P. (2008). The value of information and intensity of preference. *Decision Analysis*, 5(3):129–139. DOI: 10.1287/deca.1080.0116.

Dykstra, O. (1971). The augmentation of experimental data to maximize [X'X]. *Technometrics*, 13(3):682–688. DOI: 10.1080/00401706.1971.10488830.

Eeckhoudt, L. and Godfroid, P. (2000). Risk aversion and the value of information. *The Journal of Economic Education*, 31(4):382–388. DOI: 10.1080/00220480009596456.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Number 57 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, Boca Raton, Florida, USA.

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

European Parliament (2000). Directive 2000/60/EC, of the European Parliament and Council of 23 October 2000 establishing a framework for Community action in the field of water policy. *Off. J. Eur. Commun.*, L327:72.

Evangelou, E. and Eidsvik, J. (2017). The value of information for correlated GLMs. *Journal of Statistical Planning and Inference*, 180:30–48. DOI: https://doi.org/10.1016/j.jspi.2016.08.005.

Eyvindson, K., Hakanen, J., Mönkkönen, M., Juutinen, A., and 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. DOI: https://doi.org/10.1007/s00477-019-01745-4.

Farquhar, P. H. (1984). State of the art – utility assessment methods. *Management Science*, 30(11):1283–1300. DOI: 10.1287/mnsc.30.11.1283.

Fedorov, V. V. (1972). *Theory of optimal experiments*. New York: Academic Press.

García-Ródenas, R., García-García, J. C., López-Fidalgo, J., Martín-Baos, J. Á., and Wong, W. K. (2020). A comparison of general-purpose optimization algorithms for finding optimal approximate experimental designs. *Computational Statistics & Data Analysis*, 144:106844. DOI: https://doi.org/10.1016/j.csda.2019.106844.

Gerber, H. U. and Pafum, G. (1998). Utility functions – from risk theory to finance. *North American Actuarial Journal*, 2(3):74–91. DOI: 10.1080/10920277.1998.10595728.

Gong, M., O'Donnell, R., Miller, C., Scott, M., Simis, S., Groom, S., Tyler, A., Hunter, P., Spyrakos, E., Merchant, C., Maberly, S., and Carvalho, L. (2022). Adaptive smoothing to identify spatial structure in global lake ecological processes using satellite remote sensing data. *Spatial Statistics*, 50:100615. Special Issue: The Impact of Spatial Statistics. DOI: https://doi.org/10.1016/j.spasta.2022.100615.

Harman, R., Filová, L., and Richtárik, P. (2020). A randomized exchange algorithm for computing optimal approximate designs of experiments. *Journal of the American Statistical Association*, 115(529):348–361. DOI: 10.1080/01621459.2018.1546588.

Heiskanen, A.-S., Hellsten, S., Vehviläinen, B., and Putkuri, E. (2017). *How well is water protected in the Land of a Thousand Lakes*. Finnish Environment Institute, Helsinki, Finland. Available online at: http://hdl.handle.net/10138/177570.

Higgins, J., Zablocki, J., Newsock, A., Krolopp, A., Tabas, P., and Salama, M. (2021). Durable freshwater protection: A framework for establishing and maintaining long-term protection for freshwater ecosystems and the values they sustain. *Sustainability*, 13(4). DOI: 10.3390/su13041950.

Hilton, R. W. (1981). The determinants of information value: Synthesizing some general results. *Management Science*, 27(1):57–64. DOI: 10.1287/mnsc.27.1.57.

Howard, R. (1964). Decision analysis: Applied decision theory. *Proceedings of the 4th International Conference on Operational Research, Wiley-Interscience*, pages 55–71.

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

Karvanen, J., Vanhatalo, J., Kulathinal, S., Auranen, K., and Mäntyniemi, S. (2017). Optimal design of observational studies: overview and synthesis. Preprint. Available at: arXiv:1609.08347.

Keefer, D. L. (1991). Resource allocation models with risk aversion and probabilistic dependence: Offshore oil and gas bidding. *Management Science*, 37(4):377–395. DOI: 10.1287/mnsc.37.4.377.

Kirkwood, C. W. (2004). Approximating risk aversion in decision analysis applications. *Decision Analysis*, 1(1):51–67. DOI: 10.1287/deca.1030.0007.

Kotamäki, N., Arhonditsis, G., Hjerppe, T., Hyytiäinen, K., Malve, O., Ovaskainen, O., Paloniitty, T., Similä, J., Soininen, N., Weigel, B., and Heiskanen, A.-S. (2024). Strategies for integrating scientific evidence in water policy and law in the face of uncertainty. *Science of The Total Environment*, 931:172855. DOI: https://doi.org/10.1016/j.scitotenv.2024.172855.

Kotamäki, N., Thessler, S., Koskiaho, J., Hannukkala, A. O., Huitu, H., Huttula, T., Havento, J., and Järvenpää, M. (2009). Wireless in-situ sensor network for agriculture and water monitoring on a river basin scale in southern Finland: Evaluation from a data user's perspective. *Sensors*, 9(4):2862–2883. DOI: 10.3390/s90402862.

Lawrence, D. B. (1999). *The economic value of information*. Springer, New York.

Lawson, J. (2015). *Design and analysis of experiments with R*. Boca Raton: CRC Press, Taylor & Francis Group.

Lovett, G., Burns, D., Driscoll, C., Jenkins, J., Mitchell, M., Rustad, L., Shanley, J., Likens, G., and Haeuber, R. (2007). Who needs environmental monitoring? *Frontiers in Ecology and the Environment*, 5:253–260. DOI: https://doi.org/10.1890/1540-9295(2007)5[253:WNEM]2.0.CO;2.

Luhede, A., Yaqine, H., Bahmanbijari, R., Römer, M., and Upmann, T. (2024). The value of information in water quality monitoring and management. *Ecological Economics*, 219:108128. DOI: https://doi.org/10.1016/j.ecolecon.2024.108128.

Marchese, D. C., Bates, M. E., Keisler, J. M., Alcaraz, M. L., Linkov, I., and Olivetti, E. A. (2018). Value of information analysis for life cycle assessment: Uncertain emissions in the green manufacturing of electronic tablets. *Journal of Cleaner Production*, 197:1540–1545. DOI: https://doi.org/10.1016/j.jclepro.2018.06.113.

Mehrez, A. (1985). The effect of risk aversion on the expected value of perfect information. *Operations Research*, 33(2):455–458.

Mäntyniemi, S., Kuikka, S., Rahikainen, M., Kell, L. T., and Kaitala, V. (2009). The value of information in fisheries management: North Sea herring as an example. *ICES Journal of Marine Science*, 66(10):2278–2283. DOI: 10.1093/icesjms/fsp206.

Nadiminti, R., Mukhopadhyay, T., and Kriebel, C. H. (1996). Risk aversion and the value of information. *Decision Support Systems*, 16(3):241–254. DOI: https://doi.org/10.1016/0167-9236(95)00023-2.

Nõges, P., van de Bund, W., Cardoso, A. C., Solimini, A. G., and Heiskanen, A.-S. (2009). Assessment of the ecological status of European surface waters: a work in progress. *Hydrobiologia*, 633:197–211. DOI: 10.1007/s10750-009-9883-9.

Nygård, H., Oinonen, S., Hällfors, H. A., Lehtiniemi, M., Rantajärvi, E., and Uusitalo, L. (2016). Price vs. value of marine monitoring. *Frontiers in Marine Science*, 3(205). DOI: 10.3389/fmars.2016.00205.

Official Statistics of Finland (2020). Utilised agricultural area [e-publication], Natural Resources Institute Finland, Helsinki. Access method: http://www.stat.fi/til/kaoma/index_en.html.

Paglia, J., Eidsvik, J., and Karvanen, J. (2022). Efficient spatial designs using Hausdorff distances and Bayesian optimization. *Scandinavian Journal of Statistics*, 49(3):1060–1084. DOI: https://doi.org/10.1111/sjos.12554.

Pawitan, Y. (2001). *In all likelihood : statistical modelling and inference using likelihood*. Clarendon Press, Oxford.

Pozzi, M. and Kiureghian, A. D. (2011). Assessing the value of information for long-term structural health monitoring. In Kundu, T., editor, *Health Monitoring of Structural and Biological Systems 2011*, volume 7984, page 79842W. SPIE. DOI: 10.1117/12.881918.

Pratt, J. W. (1964). Risk aversion in the small and in the large. *Econometrica*, 32(1/2):122–136.

Pukelsheim, F. and Torsney, B. (1991). Optimal weights for experimental designs on linearly independent support points. *The Annals of Statistics*, pages 1614–1625.

Raiffa, H. and Schlaifer, R. (1961). *Applied statistical decision theory*. Harvard University, Boston.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press, Cambridge, Massachusetts. DOI: https://doi.org/10.7551/mitpress/3206.001.0001.

Reinikainen, J. and Karvanen, J. (2022). Bayesian subcohort selection for longitudinal covariate measurements in follow-up studies. *Statistica Neerlandica*, 76(4):372–390. DOI: https://doi.org/10.1111/stan.12264.

Reinikainen, J., Karvanen, J., and Tolonen, H. (2016). Optimal selection of individuals for repeated covariate measurements in follow-up studies. *Statistical Methods in Medical Research*, 25(6):2420–2433. DOI: 10.1177/0962280214523952.

Ryan, E. G., Drovandi, C. C., McGree, J. M., and Pettitt, A. N. (2016). A Review of modern computational algorithms for Bayesian optimal design. *International Statistical Review*, 84(1):128–154. DOI: https://doi.org/10.1111/insr.12107.

Stankey, G., Clark, R., and Bormann, B. (2005). Adaptive management of natural resources: theory, concepts, and management institutions. *USDA Forest Service - General Technical Report PNW*. DOI: 10.2737/pnw-gtr-654.

Sun, Z. and Abbas, A. (2014). On the sensitivity of the value of information to risk aversion in two-action decision problems. *Environment Systems and Decisions*, 34:24–37. DOI: 10.1007/s10669-013-9477-y.

Vanhatalo, J., Foster, S. D., and Hosack, G. R. (2021). Spatiotemporal clustering using Gaussian processes embedded in a mixture model. *Environmetrics*, 32(7):e2681. DOI: https://doi.org/10.1002/env.2681.

Venus, T. E. and Sauer, J. (2022). Certainty pays off: The public's value of environmental monitoring. *Ecological Economics*, 191:107220. DOI: https://doi.org/10.1016/j.ecolecon.2021.107220.

von Neumann, J. and Morgenstern, O. (1947). *Theory of games and economic behavior*. Princeton University Press.

Vuori, K.-M., Mitikka, S., and Vuoristo, H. (2009). *Pintavesien ekologisen tilan luokittelu*. Suomen ympäristökeskus (SYKE), Helsinki. (In Finnish.) Available online at: http://hdl.handle.net/10138/41785 (Accessed August 29th, 2024).

Welch, W. (1982). Algorithmic complexity: Three NP-Hard problems in computational statistics. *Journal of Statistical Computation and Simulation*, 15:17–25. DOI: 10.1080/00949658208810560.

Williams, B. K., Eaton, M. J., and Breininger, D. R. (2011). Adaptive resource management and the value of information. *Ecological Modelling*, 222(18):3429–3436. DOI: https://doi.org/10.1016/j.ecolmodel.2011.07.003.

Wätzold, F. and Schwerdtner, K. (2005). Why be wasteful when preserving a valuable resource? A review article on the cost-effectiveness of European biodiversity conservation policy. *Biological Conservation*, 123(3):327–338. DOI: https://doi.org/10.1016/j.biocon.2004.12.001.

Yang, C., Yan, J., Long, B., and Liu, Z. (2014). A novel test optimizing algorithm for sequential fault diagnosis. *Microelectronics Journal*, 45(6):719–727. DOI: https://doi.org/10.1016/j.mejo.2014.03.005.

Yokota, F. and Thompson, K. M. (2004a). Value of information analysis in environmental health risk management decisions: Past, present, and future. *Risk Analysis*, 24(3):635–650. DOI: https://doi.org/10.1111/j.0272-4332.2004.00464.x.

Yokota, F. and Thompson, K. M. (2004b). Value of information literature analysis: A review of applications in health risk management. *Medical Decision Making*, 24(3):287–298. DOI: 10.1177/0272989X04263157.

Zonta, D., Glisic, B., and Adriaenssens, S. (2014). Value of information: impact of monitoring on decision-making. *Structural Control and Health Monitoring*, 21(7):1043–1056. DOI: https://doi.org/10.1002/stc.1631.

ORIGINAL PAPERS

I

THE VALUE OF PERFECT AND IMPERFECT INFORMATION IN LAKE MONITORING AND MANAGEMENT

by

Koski, V., Kotamäki, N., Hämäläinen, H., Meissner, K., Karvanen, J., & Kärkkäinen, S. 2020

Science of the Total Environment, 726, DOI: https://doi.org/10.1016/j.scitotenv.2020.138396

Science of the Total Environment 726 (2020) 138396

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

The value of perfect and imperfect information in lake monitoring and management

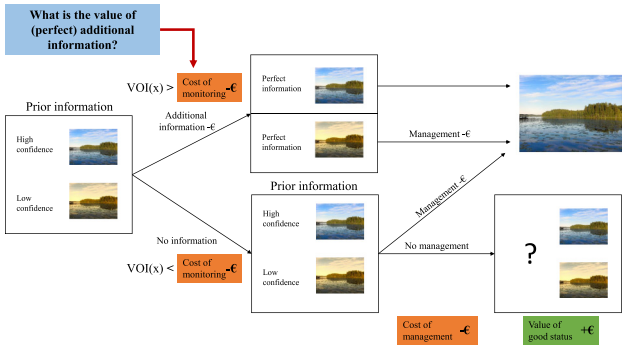
Vilja Koski [a,*], Niina Kotamäki [b], Heikki Hämäläinen [c], Kristian Meissner [d], Juha Karvanen [a], Salme Kärkkäinen [a]

[a] Department of Mathematics and Statistics, University of Jyvaskyla, P.O. Box 35, 40014 Jyväskylä, Finland
[b] Freshwater Centre, Finnish Environment Institute, Survontie 9 A, 40500 Jyväskylä, Finland
[c] Department of Biological and Environmental Sciences, University of Jyvaskyla, P.O. Box 35, 40014 Jyväskylä, Finland
[d] Programme for Environmental Information, Finnish Environment Institute, Survontie 9 A, 40500 Jyväskylä, Finland

HIGHLIGHTS

- Knowledge on the value of monitoring can assist decision-making in lake management.
- We calculate value of perfect information theoretically.
- We estimate value of imperfect information with Monte Carlo type of approach.
- Generally, monitoring is profitable to invest in if VOI exceeds the cost.
- Additional monitoring is profitable even if the lake is in good condition a priori.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 23 January 2020
Received in revised form 17 March 2020
Accepted 31 March 2020
Available online 4 April 2020

Editor: Fernando A.L. Pacheco

Keywords:
Decision making
Environmental management
Imperfect information
Lakes
Perfect information
Value of information

ABSTRACT

Uncertainty in the information obtained through monitoring complicates decision making about aquatic ecosystems management actions. We suggest the value of information (VOI) to assess the profitability of paying for additional monitoring information, when taking into account the costs and benefits of monitoring and management actions, as well as associated uncertainty. Estimating the monetary value of the ecosystem needed for deriving VOI is challenging. Therefore, instead of considering a single value, we evaluate the sensitivity of VOI to varying monetary value. We also extend the VOI analysis to the more realistic context where additional information does not result in perfect, but rather in imperfect information on the true state of the environment. Therefore, we analytically derive the value of perfect information in the case of two alternative decisions and two states of uncertainty. Second, we describe a Monte Carlo type of approach to evaluate the value of imperfect information about a continuous classification variable. Third, we determine confidence intervals for the VOI with a percentile bootstrap method. Results for our case study on 144 Finnish lakes suggest that generally, the value of monitoring exceeds the cost. It is particularly profitable to monitor lakes that meet the quality standards a priori, to ascertain that expensive and unnecessary management can be avoided. The VOI analysis provides a novel tool for lake and other environmental managers to estimate the value of additional monitoring data for a particular, single case, e.g. a lake, when an additional benefit is attainable through remedial management actions.

* Corresponding author.
E-mail address: vilja.a.koski@jyu.fi (V. Koski).

https://doi.org/10.1016/j.scitotenv.2020.138396
0048-9697/

1. Introduction

Human-induced stress and disturbances threaten inland and coastal waters more severely than many other ecosystem types (Sala et al., 2000). Therefore dedicated legislation, such as the Clean Water Act (CWA) in the U.S and the EU Water Framework Directive (WFD) (European Parliament, 2000) have been adopted to protect the ecological structure of inland and coastal aquatic ecosystems, to secure their functioning and provisioning of ecosystem services. In the European Union, the WFD aims at ensuring good status in rivers, lakes, coastal and ground waters by 2027. The WFD status classification of water bodies into five ecological status classes (high, good, moderate, poor and bad) is primarily based on regular and long-term monitoring data of parameters representing biotic structure, supported by the physical and chemical properties of water and hydrological and morphological features (European Communities, 2003). For each classification variable, the status class is assessed against the degree of deviance from the pre-determined reference conditions.

Under the WFD, assessing the ecological status that identifies possible management needs and subsequent restoration measures, requires extensive monitoring programs that produce reliable data for decision making. For cost-efficient decision making in water management, the use of relevant information is important. However, it is often challenging to know when these information criteria have been optimally met to achieve the most profitable outcome. In addition, in ecological monitoring the uncertainty is an inevitable part of the data (Carstensen and Lindegarth, 2016). The value of information (VOI) analysis can be a useful approach to control for that uncertainty and to assess concretely how much it is profitable to pay for monitoring.

The VOI is a concept of decision theory that assesses the value of additional information to solve a decision making problem. One of the earliest references is by Schlaifer and Raiffa (1961) and a modern presentation is given e.g. by Eidsvik et al. (2015) and Canessa et al. (2015). A tenet of the VOI is that while additional information can help reduce uncertainty, it is profitable to gather only if it affects the conclusion. More specifically, the VOI analysis aims to assess and compare the expected outcomes in the decision situation. Making a decision implies that one of all the possible alternatives must be chosen to achieve the specified objectives. The uncertainty in the decision situation affects the expected outcomes of each alternative. To calculate the VOI, one needs to specify the decision to make, the random variables that affect the decision situation, the scenarios formed from these decisions and random variables, and the monetary value of each scenario to the decision maker. The VOI is commonly divided into two categories:

1. The value of perfect information (also known as the expected value of perfect information, EVPI) is the value of data that provide exact information on the state of the system.
2. The value of imperfect information (also known as the expected value of sample information, EVSI) expresses the value of data providing less than perfect information.

In the literature, EVPI and EVSI are frequently used terms for the same concepts. However, we use the definitions of value of perfect and imperfect information according to Eidsvik et al. (2015).

The VOI analysis framework is already widely applied in the fields of economics, finance and medicine (Eidsvik et al., 2015) and the potential of the approach also in environmental and ecological decision making has been recognized and increasingly applied in recent years (Bolam et al., 2019; Eyvindson et al., 2019). Perhaps surprisingly, the VOI is still seldom applied to environmental monitoring data, despite the increasing demand (Colyvan, 2016) for such analysis. As far as we know, Nygård et al. (2016) is the first one to apply VOI analysis with perfect information to assess the value of marine monitoring data. They developed a conceptual model of the components that needed to be established when calculating the VOI of monitoring data. In the present study, we follow their model but as the major novelty, we extend the VOI analysis also to imperfect information in the context of surface water monitoring.

So why has the use of the VOI still remained limited with environmental monitoring? A major difficulty in applying the VOI approach to environmental management and monitoring is to define the monetary value of the present and targeted ecological status of the environment. Some economic evaluation studies for fresh waters (e.g. Atkins and Burdon (2006)) exist, but these estimates do not directly translate to our context. Here, we build on the valuation study by Ahtiainen (2008) who used the contingent valuation method (Carson et al., 2004) to study the economic benefits attributable to improvement of ecosystem status from moderate to good in the Finnish lake Hiidenvesi. Secondly, the high computational cost prevents the more common use of the VOI, especially for the value of imperfect information (Steuten et al., 2013).

In the present work, we want to fill the gap of missing real-life applications of VOI analysis concerning environmental monitoring data and propose a method to further the more frequent use of VOI. We aim to use the VOI analysis to assess the worth of the additional information needed to gain a more reliable estimate of the ecological status of a water body when there is already a preconception about its true status. We show that both perfect information as well as imperfect information approach can be used to evaluate the value of additional monitoring data. First, we aim to form an analytical solution for the value of perfect information in the case of two ecological status classes and two alternative decisions. Second, we propose how to calculate the value of imperfect information empirically using simulation methods. In addition, our aim is to evaluate the uncertainty of the value of imperfect information with confidence intervals. Third, we conduct a sensitivity analysis to study how different assumptions affect the VOI. The assumptions are related to: i) the monetary value of a lake meeting the quality requirements of good status, ii) the cost of the management action and iii) the outcome of the implemented management option, i.e. whether or not the target ecological status is achieved. Lastly, we compare the VOI to the realized costs of the monitoring data.

2. Materials and methods

2.1. The methodology for estimating the value of information

This section follows the concepts and notations by Eidsvik et al. (2015). All notations are summarized in Table 1. In a decision situation, there are two types of variables, 1) decisions and 2) variables with

Table 1
Definitions of used notations.

Notation	Definition
$x \in \Omega$	Discrete variable, direct measurement of status
$a \in A$	Alternative or action
(x, a)	Scenario
c	Cost of implementing alternative a
r	Ratio of target status obtained after actions
$v(x, a)$	(Monetary) value of the scenario (x, a)
y	Continuous variable, indirect measurement of x
$p(x)$	Prior knowledge of x
$p(y)$	Marginal density of y
$p(x \mid y)$	Posterior probability of x given y
PV	Prior value
$PoV(x)$	Posterior value of perfect information
$PoV(y)$	Posterior value of imperfect information
$VOI(x)$	Value of perfect information
$VOI(y)$	Value of imperfect information

uncertainty. If the decision maker can control the value of a variable, the variable is categorized as a decision. We refer to the values of a decision as alternatives or actions and denote the set of them by A. The decision maker can choose any alternative $a \in A$. Moreover, if the decision maker cannot control a variable value, it is classified as a variable with uncertainty. The value of a random variable is called a state or a realization and is denoted by x. A discrete random variable is defined based on its sample space Ω, with the probability $p(x) \geq 0$ of the state $x \in \Omega$ such that $\sum_{x\in\Omega} p(x) = 1$. For example, in an environmental framework we can have two alternatives: i) a management action to a water body ($a = a_1$) or ii) no action ($a = a_0$), while water bodies may have two states: i) a target status ($x = x_1$), and ii) a non-target status ($x = x_0$).

A scenario is an instantiation of every variable in the decision situation. The decision situation always involves a total of $|\Omega| \times |A|$ different scenarios. Each scenario with the decision a and the uncertainty x has an outcome with a value function $v(x, a)$ given by the decision maker. It is equal to the value of the realized outcome for the decision maker when also the costs of the action and the change of the value due to the action are taken into account. For example, we could have a cost for a management action ($c = c_1$) and for no actions ($c = c_0$). The effectiveness of an action may be specified by a parameter $r \in [0,1]$. It is the ratio from value $v(x, a)$ of how much an action can affect the monetary value compared to a situation where an action is not performed. The utility function $u(\cdot)$ is an extension of the value function that also measures the decision maker's ability to tolerate risk (von Neumann and Morgenstern, 1944). Risk seeking or risk averse decision makers could be taken into account by measuring the expected utility of outcomes instead of the expected value. In our set-up, we assume that the decision maker is risk neutral, so $u(v(x, a)) = v(x, a)$.

The value of information (VOI) is the price threshold at which the decision maker is indecisive about whether or not to acquire additional information to make a decision on an action, for example on a management action. In other words, the VOI is the maximum price, yet still profitable to invest into additional information. The decision making has two steps:

1. Make a decision about whether or not to obtain additional information.
2. Make an actual decision, either based on prior knowledge alone or on prior information and on the additional information.

The flowchart for decision making progress is shown in Fig. 1. The VOI is calculated in the first step.

The value of perfect information can be written as

$$VOI(x) = PoV(x) - PV, \tag{1}$$

where

$$\begin{aligned} PV &= \max_{a\in A} \{\mathrm{E}(v(x,a))\} \\ &= \max_{a\in A} \left\{ \sum_{x\in\Omega} v(x,a)p(x) \right\} \end{aligned} \tag{2}$$

and

$$\begin{aligned} PoV(x) &= \mathrm{E}\left(\max_{a\in A} \{v(x,a)\} \right) \\ &= \sum_{x\in\Omega} \max_{a\in A} \{v(x,a)\} p(x). \end{aligned} \tag{3}$$

Above, PV (prior value) is a priori the maximum expected benefit of all expected benefits, given all available alternatives. A rational decision maker should choose the alternative that maximises the average benefit. Secondly, $PoV(x)$ (posterior value) is the updated expected benefit after new information is gained, i.e. the average maximum benefit. The $VOI(x)$ is the difference between these benefits. If the $VOI(x)$ exceeds the price of the information, the decision maker should invest in collecting the data. The $VOI(x)$ is always non-negative, since the averaging of the maximum benefit of states is always at least as large as the maximum benefit of averaging over states.

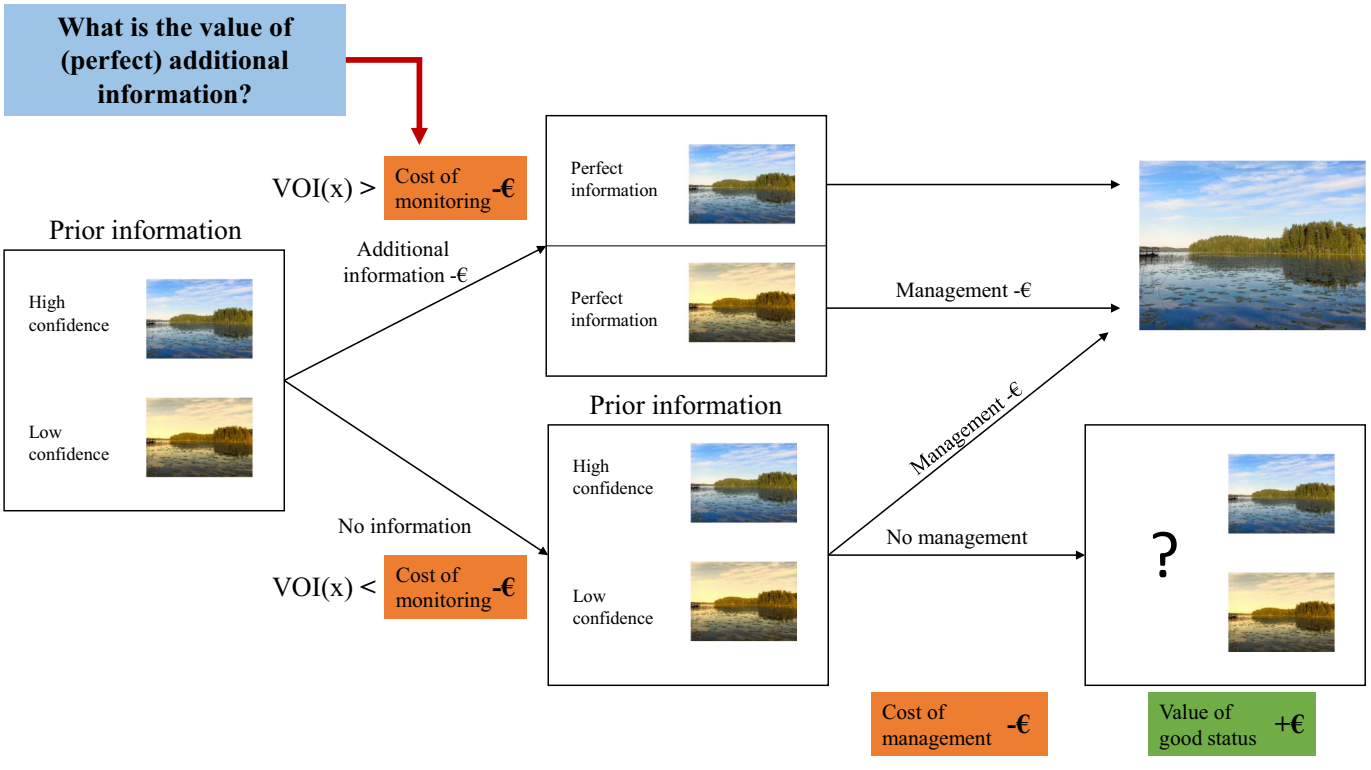


Fig. 1. The decision making progress for lake management.

In Eq. (3) it is assumed that the additional information is perfect, providing a certain knowledge of the state x. The $VOI(x)$ is then the absolute maximum, at which it is still profitable to pay for additional information. However, in many cases additional information cannot provide a completely accurate knowledge about the state x. Instead, we observe, for example, the value y of a continuous random variable with the density $p(y)$ reflecting, but imperfectly, the state x. Then, the observed information is referred to as imperfect information.

The value of imperfect information is given as

$$VOI(y) = PoV(y) - PV, \tag{4}$$

where PV is as in Eq. (2) and

$$\begin{aligned} PoV(y) &= \int_y \max_{a\in A} \{\mathrm{E}(v(x,a)|y)\}p(y)dy \\ &= \int_y \max_{a\in A} \left\{ \sum_{x\in\Omega} v(x,a)p(x|y) \right\} p(y)dy. \end{aligned}$$

The posterior distribution $p(x|y)$ above can be calculated with Bayes' rule, see Eq. (5).

A major source of complexity in the VOI problems is the need to model continuous probability distributions (Yokota and Thompson, 2004). To simplify the decision situation and problem solving, a continuous input is often categorized. Categorization of a continuous variable is normally a bad idea since it leads to loss of information. In addition, the results are depending on arbitrary cut points set by the decision maker (Royston et al., 2006). Our aim is to avoid categorization, but an analytic solution for $VOI(y)$ is rarely available because of a continuous sample space and hence, integration. To obtain an approximate solution, we utilize a Monte Carlo type of approach to the integration using empirical data (Robert and Casella, 2005). We approximated the posterior value for imperfect information by

$$\begin{aligned} \widehat{PoV}(y) &= \frac{1}{n}\sum_{i=1}^{n} \max_{a\in A} \{\mathrm{E}(v(x,a)|y_i)\} \\ &= \frac{1}{n}\sum_{i=1}^{n} \max_{a\in A} \left\{ \sum_{x\in\Omega} v(x,a)\hat{p}(x|y_i) \right\}, \end{aligned}$$

where n is the number of observations, and in our case, y_i, $i=1,\ldots,n$, are the values sampled from distribution $\hat{p}$ fitted to the data of chlorophyll concentration and x is the ecological status. Furthermore, the posterior distribution is given by Bayes' rule

$$\hat{p}(x|y_i) = \frac{\hat{p}(x)\hat{p}(y_i|x)}{\hat{p}(y_i)}, \tag{5}$$

where we estimated $\hat{p}(y_i \mid x)$ by gamma distribution (see an example in Fig. 2). The marginal distribution of y_i is defined for states of x as follows:

$$\hat{p}(y_i) = \sum_{x\in\Omega} \hat{p}(x)\hat{p}(y_i|x).$$

For example, for two environmental states $x_j, j = 0, 1$, in the following we use $\hat{p}(x_1) = 0.48$, which is the estimated proportion of water bodies in the target status.

We estimate confidence intervals of $\widehat{PoV}(y)$ using the parametric percentile bootstrap method (Efron and Tibshirani, 1993). The simulation is implemented as follows:

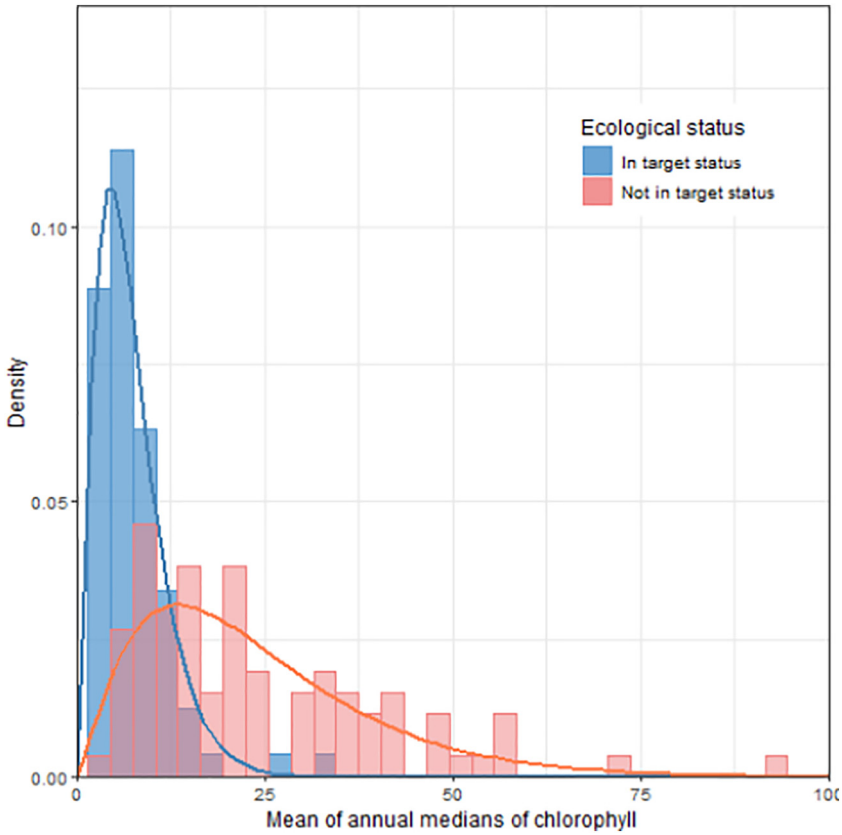


Fig. 2. Histograms and fitted gamma distributions of chlorophyll *a* concentration of 166 water bodies based on monitoring. The water bodies are categorized into two classes, in target status (blue) and not in target status (red). The value on the horizontal axis is the aggregated value of the chlorophyll *a* concentration over seven years (2006 to 2012) and monitoring locations in water bodies.

1. Random samples $y_1, \ldots, y_{n0}$ and $y_1, \ldots, y_{n1}$ are drawn from gamma distributions of target and non-target water bodies fitted to the original data, with respect to the original proportions.
2. Gamma distributions $Gamma(\alpha_0^*, \beta_0^*)$ and $Gamma(\alpha_1^*, \beta_1^*)$ are re-fitted to the random samples.
3. The posterior value $\widehat{PoV}(y^*)$ is calculated using the re-calculated fits.

We repeated this $B = 1000$ times in the spirit of bootstrap and obtained bootstrap replicates $\widehat{PoV}(y^{*(b)})$, $b = 1,\ldots,B$. Confidence intervals of $\widehat{VOI}(y)$ can be derived from these confidence intervals by substracting the prior value *PV*. All the calculations were implemented with R (R Core Team, 2018).

2.1.1. Conceptual model

Nygård et al. (2016) developed a conceptual model that sums up the components that are needed when evaluating the VOI of monitoring data. We applied and extended their model to inland water monitoring using imperfect information. After identifying the variables connected to monitoring, the following steps need to be performed:

1. List the alternative monitoring activities that could be carried out to gain additional information. The list can include several variables and several strategies. In our application, we considered two alternatives; either to implement monitoring of chlorophyll data or not to implement it, depending on the price of monitoring compared to VOI (see Section 3.1).
2. Estimate the costs of these monitoring alternatives, which can subsequently be compared to the VOI (comparison in Section 3.1).
3. Assess the status (prior information $\hat{p}(x)$) based on expert judgment. The existing past data or knowledge can be used to obtain the prior, the subjective probability. For several status classes, the relative likelihood approach, for example, could be used, see French et al. (2010). We defined the priors in our case study in Section 2.3.
4. Assess the status after the selected monitoring activity has been carried out. If imperfect information (y) is used, the assessment of the status can be based on the statistical classification. See Section 2 and Eq. (5).
5. List the alternative management actions (a) depending on the status of the system. We applied two alternatives; either implement (a_1) or do not implement actions (a_0) (Table 2).
6. Estimate the costs (c) of these management actions (examples of costs, c_1 and c_0, presented in Table 2).
7. Estimate the change in the state of the system if the management options are implemented. We used the ratio r for describing the degree of change (see Table 2).
8. Estimate the monetary values $v(x, a)$ of different states of the system. In our case, defining the monetary value of reaching a target ecological status in the water body was sufficient. (Section 2.4).

After implementing the steps 1–8, the formula (1) was applied to calculate the $VOI(x)$ and formula (4) for the $VOI(y)$.

2.2. Monitoring activity on imperfect information

According to the WFD, the overall ecological status assessment of a water body is based on data collected for several biological quality elements and indicators. However, we limit our VOI analysis to the biological quality element phytoplankton, or more specifically to chlorophyll *a*, one of the many indicator variables in lake status assessment. Chlorophyll *a* content is indicative of water body productivity and therefore generally correlates well with the ecological status of lentic water bodies mainly suffering from human-induced eutrophication. Lentic water bodies are also subject to other major anthropogenic stress, such as intense water level regulation, were not included in this analysis.

The data that we used in the analysis are produced by the official Finnish lake monitoring program and stored in the open source database of the Finnish Environment Institute (http://www.syke.fi/en-US/Open_information). We used chlorophyll *a* data from the second assessment period of the river basin management plans (years 2006–2012). Our data contain 144 lakes and 166 water bodies within them: a water body is either a whole lake or more rarely, a limited, homogeneous part of a lake. In the following, we refer to water bodies sometimes simply as lakes for brevity. We selected the most frequently sampled water bodies with at least 3 summertime observations per year. Overall we included 6742 observations from 166 water bodies. We aggregated annual and local observations into means of annual medians per water body. This is the standard current approach in ecological status assessment of water bodies (Aroviita et al., 2019). The lakes are divided into 14 lake types with 1–31 lakes per type. We accounted for lake type in status classification, as types naturally differ in chlorophyll concentration. We note that joining all lake types in our analysis may overestimate the uncertainty of chlorophyll as an indicator of status resulting in a more inaccurate VOI. However, the sample size is insufficient to allow for a closer study by lake type.

Since the overall ecological status of a water body determines the need for management actions, we categorized water bodies into those that either met the target status during the second classification period, or those that did not. Of the total of 166 water bodies, 79 (48%) met high or good status while 87 (52%) did not, i.e. belonged to either the moderate, poor or bad status class. We assume that the chlorophyll content indicates the status of a water body and represents the value y. Thus, the $VOI(y)$ is estimated by using the empirical distribution of chlorophyll. Fitting gamma distributions, we estimated the distribution of aggregated values of chlorophyll over time and monitoring locations separately for water bodies in both good and in less than good status (Fig. 2).

2.3. Priors

We used three distinct priors for the distribution of the ecological status, to illustrate how the prior knowledge about the status affects the VOI. Without more detailed knowledge about the ecological status of a given water body, a prior was estimated from the data: the proportion of water bodies in target status to the total number, i.e. $\hat{p}(x_1) = 79/166 = 0.48$. We also have more detailed prior information on some lakes, for example, as in the case of lake Hiidenvesi, which is currently

Table 2
A summary of costs c and the monetary values $v(x, a)$ for an example lake Hiidenvesi, where the value of the target ecological status equals EUR 1000 per ha (Ahtiainen, 2008). The costs of the management action were set to EUR 200 per ha. The monetary value takes the cost of a management alternative into account. See text for more details.

		Monetary value (EUR/ha)	
		Ecological status x	
Alternative a	Cost c of alternative (EUR/ha)	x_0: non-target status	x_1: target status
a_0: no actions	$c_0 = 0$	0	1000
a_1: actions	$c_1 = 200$	1000–200 = 800	1000–200 = 800

assessed to be in moderate status with high degree of certainty. So, we set the prior to $\hat{p}(x_1) = 0.20$ and also to $\hat{p}(x_1) = 0.80$.

2.4. Monetary value of lakes

Evaluation of the monetary values of lakes $v(x, a)$ is challenging because a valuation of the environment is not straightforward. Reynaud and Lanzanova (2017) conducted a meta-analysis on the economic value of ecosystem services delivered by lakes and their value to the private properties located next to lakes. According to Reynaud and Lanzanova, the mean value of a lake to an adjacent property in Finland was USD 265.9 (in 2010) per property per year. However, these estimates do not directly translate into the benefit achieved by management, i.e. the monetary value between the two status categories treated here.

We used a valuation study by Ahtiainen (2008) that studied the economic benefits attributable to the improvement of the status of a single lake, Hiidenvesi, with an area about 3000 hectares (ha), currently in moderate condition. She assessed the willingness of residents to pay for management actions to reduce the eutrophication of the lake. As the described target status in the poll broadly corresponds to the definition of good status under the WFD, her results are appropriate for our purposes. Based on the poll, the mean sum residents were willing to pay ranged between EUR 4.08–54.48 per property per year. Furthermore, the overall estimated willingness of properties to pay ranged between EUR 3 and 5.7 million over the course of the five-year management implementation period. For generality, we used the VOI of a water body per hectares to combine it with the estimated average cost c_1 of a management action cost of EUR 200 per ha (source: Finnish Environmental Institute). In doing so, our results apply to water bodies of all sizes.

We first chose to use the most conservative value of EUR 3 million per 3000 ha = EUR 1000 per ha as the value $v(x_1, a_0)$ for a water body in target status with no performed actions. When constructing the value for the scenarios, summarized in Table 2, the monetary value of ecological status was estimated by subtracting the cost c of each management option from a value of a scenario by each row. Therefore, if a water body indeed needs and receives management (average EUR 200 per ha) and its status also improves to the target status, the value of the water body increases to EUR 1000 per ha. However, the value of a water body is only EUR 800 per ha after taking into account the cost of management. For simplicity, we first assumed that the target status is achieved as a result of management actions. Later, we also released this assumption so that the target status is not always reached after implementing a management action (Table 4). Here, the VOI is insensitive to the absolute monetary value; only the increase in monetary value when the status of a lake increases, is significant.

In the preceding example, the value $v(x, a)$ of a water body in target status ($x = x_1$) with no restoration ($a = a_0$) is fixed to EUR 1000 per ha. As this estimated value is uncertain and may vary substantially among lakes, we also performed a sensitivity analysis, to examine the effect of a variable monetary value on the VOI by varying the value of $v(x, a)$ from EUR 200 to 2000 per ha.

3. Results

First, using four monetary values $v(x, a)$ given in Table 2 and when a priori we are more certain that a lake is indeed in target status ($\hat{p}(x_1) = 0.8$), the expected values of two alternative actions a_0 and a_1, $E(v(x, a_0))$ and $E(v(x, a_1))$, are equal. Thus, maximum expected value (PV, Eq. (2)) has the same value, EUR 800 per ha (Table 3, first row). However, additional information is profitable to gather and worth paying for up to a maximum of EUR 160 per ha for perfect information (Eq. (1)). For imperfect information, it is worth to pay up to EUR 100 with 95% CI (85.7, 115.1) per ha (Eq. (4)) to ascertain the ecological status of the lake.

If in turn we are a priori more certain of the water body to be in non-target status ($\hat{p}(x_1) = 0.2$), i.e. it likely needs restoration, then it is profitable to implement the restoration to achieve the expected value of EUR 800 per ha (Table 3, second row). Furthermore, it is worth paying a maximum of EUR 40 per ha for perfect information and EUR 0 with 95% CI (0, 3.5) per ha for imperfect information.

When the proportion of lakes in target status, as estimated from the data, is used as the prior, i.e. $\hat{p}(x_1) = 0.48$, the highest expected return is obtained by management: the expected value is then EUR 800 per ha (Table 3, third row). Moreover, it is worth paying EUR 95 per ha for perfect information and EUR 15 with 95% CI (0, 33.2) per ha for imperfect information to ascertain the true status of the lake.

3.1. Monitoring costs

If the VOI exceeds the price paid for gathering the information, the additional information is profitable for decision making. We compared the obtained VOI to actual monitoring costs, based on the information from the Finnish Environmental Institute (personal communication). One sample of chlorophyll a, from collection to analysis, currently costs EUR 138. Thus, costs for the entire data, i.e. the 6742 samples equals EUR 930 396. In our data, 107 chlorophyll a observations were taken from Hiidenvesi over the years 2006–2012, which equals EUR 14 766.

Depending on the prior knowledge presented, $VOI(x)$ ranges between EUR 40–160 per ha and $VOI(y)$ between EUR 0–100 per ha on average, when the monetary value of target status was fixed to EUR 1000 per ha. If we assume that the ecological status meets the target ($\hat{p}(x_1) = 0.8$) and that additional information provides imperfect knowledge about the status, the $VOI(y)$ equals EUR 100 per ha. Hiidenvesi has an area of 3030 ha and thus a $VOI(y)$ of EUR 303000. If we assume that the ecological status does not meet the target ($\hat{p}(x_1) = 0.2$), $VOI(y)$ for Hiidenvesi equals EUR 0. If we assume the prior $\hat{p}(x_1) = 0.48$, $VOI(y)$ for Hiidenvesi equals EUR 45450. When the prior $\hat{p}(x_1)$ equals either

Table 3
VOI analysis for an example lake, Hiidenvesi, when the monetary value of the target status is EUR 1000 per ha. The prior value is based on the maximizing alternatives, i.e. implement management actions (a_1) or not (a_0). VOI should be compared with the monitoring cost of EUR 4.9 per ha obtained by dividing the monitoring cost EUR 14766 by the area of Hiidenvesi 3030 ha.

	Prior $\hat{p}(x)$		Prior value,	Perfect information		Imperfect information	
	Not in target status	In target status	PV (€/ha)	$PoV(x)$ (€/ha)	$VOI(x)$ (€/ha)	$PoV(y)$ (€/ha)	$VOI(y)$ (€/ha) (95% CI)
Prior given by the manager	0.2	0.8	800 (a_0/a_1)	960	160	900	100 (85.7, 115.1)
	0.8	0.2	800 (a_1)	840	40	800	0 (0, 3.5)
Prior estimated from data	0.52	0.48	800 (a_1)	895	95	815	15 (0, 33.2)

0.8 or 0.48, the realized monitoring cost are significantly smaller than the estimated $VOI(y)$ for Hiidenvesi. Hence, in both cases, it would be profitable to gather additional information. Similar calculation can be performed for any prior.

3.2. Sensitivity analysis

Since the most uncertain assumption is the monetary value of the ecological status of the lake, we modelled the effect of different monetary values on VOI of both perfect and imperfect information. Instead of using a fixed value of EUR 1000 we varied the value of $v(x_1, a_0)$ from EUR 200 to 2000 per ha. Table 4 presents a generalization of Table 2 for the purpose of sensitivity analysis. For further comparison, three different priors for ecological target status x_1 were used as earlier: $\hat{p}(x_1) \in \{0.80, 0.20, 0.48\}$, where the first two were provided by an expert and the third was estimated from data. In addition, the cost of management c_1 was either EUR 100 or EUR 200 per ha and a value of non-target status lake after choosing management alternative was reduced with ratio $r \in \{0.70,1\}$. Any other proper values for the prior, cost and ratio could be used as well.

The $VOI(x)$ can be shown to be a piecewise-defined function of monetary value $v = v(x_1, a_0)$ that consists of three different functions. If $v \le c_1/r$, gathering additional information would be useless because management activities are too expensive to implement, thus $VOI(x)$ equals zero. If $v > c_1/r$, it is useful to calculate VOI. If $c_1/r < v < c_1/(r - rp)$ with $p = \hat{p}(x_1)$, $VOI(x)$ is an increasing function of v. After the change point $v = c_1/(r - rp)$, $VOI(x)$ is a positive constant. The derivation of these results is presented in Appendix A. According to Fig. 3, also the $VOI(y)$ is an increasing function of value v until the same change point. After the change point, $VOI(y)$ starts to approach zero. If the monetary value is large compared to the costs, it is always profitable to implement management actions to ascertain good ecological status, and any additional information is then unprofitable. For the expectation that a water body does not need management ($\hat{p}(x_1) = 0.80$), the cost equals EUR $c_1 = 200$ per ha and the ratio $r = 1$ (Fig. 3, top left panel), $VOI(x)$ starts to increase from zero when the monetary value of the target status of a water body equals EUR 200 per ha. The maximum of VOI is reached when $v = 1000$. Then, the $VOI(x)$ is EUR 160 per ha and that of imperfect information EUR 121 per ha, respectively. The same pattern for $VOI(x)$ and $VOI(y)$ is repeated for other assumptions of prior $\hat{p}(x_1)$, cost c_1 and r.

The maximum value of $VOI(x)$ increases on average when it is increasingly certain that the lake is in the target status, i.e. the value of $\hat{p}(x_1)$ increases (Fig. 3). In turn, the more certain it is that the lake is in non-target status, the faster $VOI(x)$ reaches its maximum value. The $VOI(x)$ is the absolute maximum worth paying for additional information. The $VOI(y)$ depends on the priors and the data, but it is always less than the $VOI(x)$.

4. Discussion

According to our knowledge, this study is the first attempt to implement $VOI(y)$ to lake monitoring data. From a methodological point of view, the main results are shown in Fig. 3, where the VOI is presented as a function of monetary value. The results for perfect information are derived theoretically while the results for imperfect information are based on simulations. $VOI(x)$ naturally exceeds $VOI(y)$ for all monetary values, and the change point seems to be the same for perfect and imperfect information. In the case of perfect information, $VOI(x)$ first increases linearly until the monetary value reaches the change point, and then remains constant. Thus, in this setup if the monetary value is known to exceed the change point, it is not necessary to fix the monetary value more exactly. In contrast, for imperfect information, $VOI(y)$ first increases linearly until the monetary value reaches the change point and then decreases. Thus, the situation differs essentially from the case of perfect information because the exact determination of the monetary value is always needed to calculate $VOI(y)$.

From the environmental management point of view, the main result is that the monitoring is most often cost-efficient. When comparing the realized monitoring costs and the estimated VOI, costs are significantly smaller and thus still profitable to invest in. Interestingly, even with a good a priori understanding of the ecological status of the lake, it may still be profitable to gather additional information. We found, perhaps somewhat counter-intuitively, that the VOI is highest when the ecological status is expected to meet the target, and the decision maker is fairly certain that there is no need for management actions. In this case, it is worth gathering additional information to unequivocally confirm that the lake meets the quality standards, in order to avoid unnecessary and expensive management actions, while minimizing any risks of losing the expected benefits of good ecological status. Indeed our results suggest that while river basin management strives to be more cost-efficient (Carvalho et al., 2019), the monetary investment in the current lake monitoring is often actually profitable.

We related the benefit of additional information to chlorophyll *a* in this work. However, a one year intensive sampling of a lake using all required biological quality elements includes also 5 annual physico-chemical samplings, and the sampling of phytoplankton on three occasions. Moreover, in fully compliant WFD assessments, littoral and/or profundal macroinvertebrates should be sampled twice and a single fish and macrophyte survey should be conducted in the course of each river basin management period of six years. Based on the information from Finnish Environmental Institute the estimated cost for all the aforementioned is around EUR 6000 per year per lake (personal communication). But even using EUR 6000 per year as the true monitoring cost per lake, our calculations suggest that monitoring is financially profitable for lakes within the size criteria monitored under the WFD.

We recognize that much uncertainty is associated with the estimated monetary value of status improvement. The economic value of lakes has been studied quite intensively (Reynaud and Lanzanova, 2017), but the results are difficult to generalize especially for our specific purposes. Also, results of valuation studies are context specific: Hjerppe et al. (2017) recently estimated that the recreational value of the Finnish lake Pien-Saimaa in its present moderate ecological status is EUR 21 100 000 per year. However, the comparable value for our purposes is the difference in the recreational value between the lake in current moderate status and good status (EUR 21 560 000) and it is only EUR 38 per ha per year, the area of Pien-Saimaa being 120 km^2. This would mean only EUR 190 per ha for a 5 year period, which is much smaller than the wholesale value of EUR 1000 per ha provided by

Table 4
The monetary values $v(x, a)$ in a four scenario management decision-making situation, i.e. with two possible decision alternatives and two states for the uncertainty. The cost c of a management alternative is taken into account, as well as the possibility that implementing a management option does not necessarily help to reach the target status. This is implemented with the ratio r.

		Monetary value $v(x, a)$	
		Ecological status x	
Alternative a	Cost c of alternative	x_0: non-target status	x_1: target status
a_0: no actions	$c_0 = 0$	$v(x_0, a_0)$	$v(x_1, a_0)$
a_1: actions	c_1	$v(x_0, a_1) = r \cdot v(x_1, a_0) - c_1$	$v(x_1, a_1) = v(x_1, a_0) - c_1$

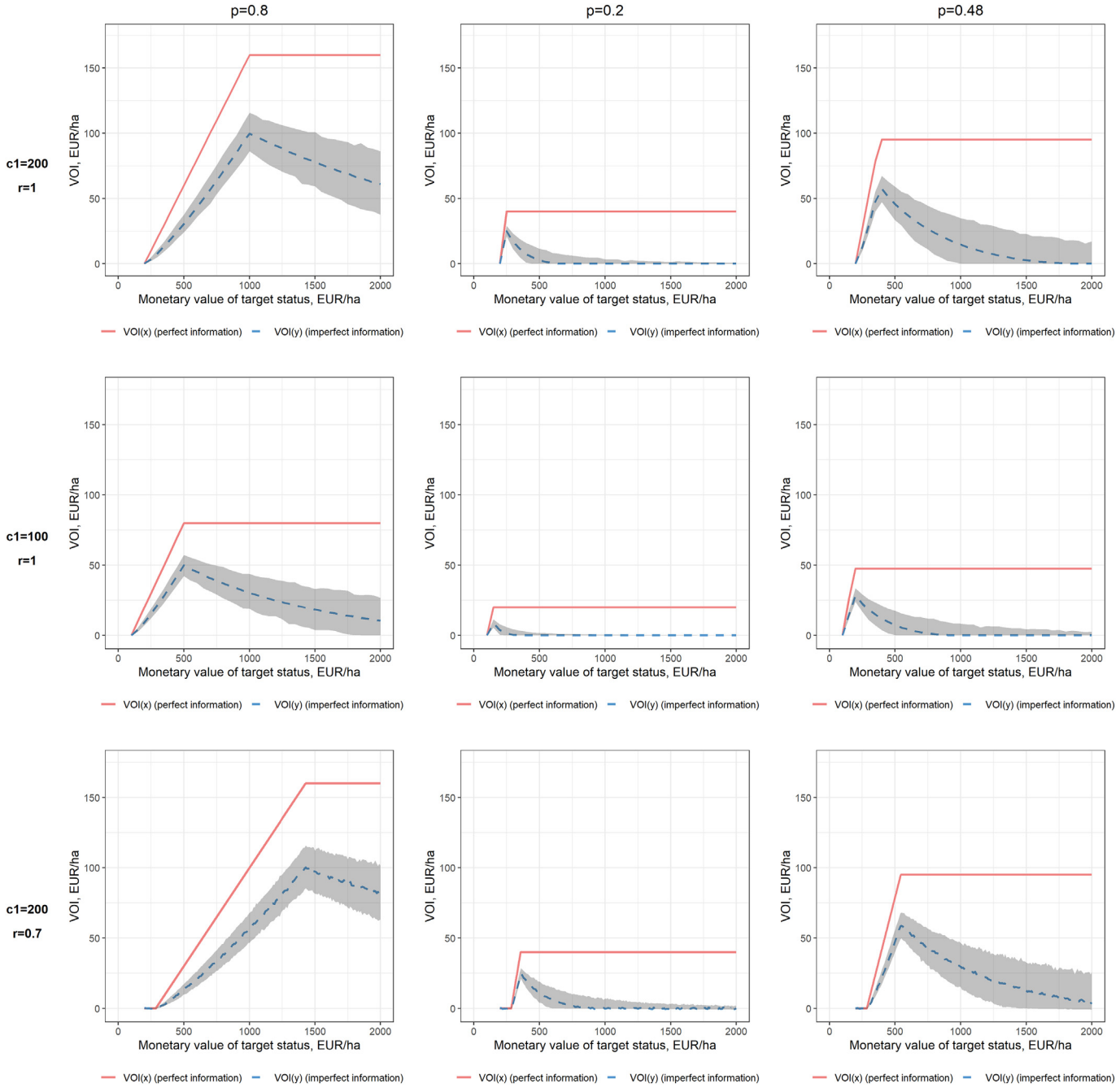


Fig. 3. The effect of the value of a water body in target status on the VOI. $VOI(x)$ is the value of perfect information and the maximum value of VOI. $VOI(y)$ with 95% confidence intervals is the value for imperfect information. The prior is fixed from left column to right: $\hat{p}(x_1) = 0.8$, $\hat{p}(x_1) = 0.2$ and $\hat{p}(x_1) = 0.48$, respectively. The cost of management is fixed from top row to bottom: $c_1 = 200$, $c_1 = 100$ and $c_1 = 200$. The ratio r is also fixed from top row to bottom: $r = 1$, $r = 1$ and $r = 0.7$.

Ahtiainen (2008), that we used in our analyses. Interestingly, if indeed the monetary value in EUR/ha is as small as Hjerppe et al. (2017) suggest our study indicates that in most scenarios, collecting any additional information on the status of the lake would be useless from the decision-making point of view (Fig. 3). Another study by Artell and Huhtala (2017) estimated that owners of a lakefront property were willing to spend up to EUR 5400 for the improvement of the lake status from moderate to good. Economic value of a lake being this inconsistent, we performed a sensitivity analysis to evaluate the influence of changes in monetary value on the VOI.

In this work, we scaled the value to the size of a lake in an attempt to generalize results to different size water bodies, thus implicitly assuming that the costs and value of status improvement per unit area remain constant. The results on the effects of size on the lake value are inconsistent, but the recent meta-analysis by Reynaud and Lanzanova (2017) suggests a positive relationship between the value per property and lake area. Larger lakes might be more highly valued than smaller lakes, since they might underpin a wider range of ecological functions (Brander et al., 2006) and perhaps a greater variety of valued water uses. We do not have any data on the dependency of per unit area management costs on the lake size. However, in practise relative monitoring costs per unit area are smaller for large water bodies, where fewer samples per area are taken for status assessment. Therefore, in reality the cost of one sample is greater in smaller water bodies.

Lastly we see great potential in the use of the VOI in environmental management and guidance of when to commit more resources to monitoring and when not to. The decision about whether to monitor or not is particularly applicable in the context of adaptive management of

natural resources (e.g. Canessa et al. (2015), Williams et al. (2011)). Adaptive management is an iterative process where uncertainties can be reduced and management improved by monitoring the management outcomes and learning from them (Holling, 1978). A future challenge will be to extend the VOI analysis to other environmental monitoring alongside growing support for a wider adoption of the concept of adaptive management.

5. Conclusion

The main aim of this paper was to demonstrate that the concept of VOI analysis can be successfully applied to monitoring in a lake management decision making context. To do so, we applied VOI analysis to lake monitoring data on chlorophyll *a* concentrations. As a baseline for the analysis, we first proposed the analytical formulas for the value of perfect information in the case of two ecological status classes and two alternatives. Second, we proposed how to calculate the value of imperfect information from the monitoring data by using a Monte Carlo type of simulation method and how to evaluate the uncertainty with confidence intervals based on the percentile bootstrap method. Third, we implemented a sensitivity analysis to study how the monetary value of a water body in target status affects the VOI in the case of perfect and imperfect information. The main restrictions we needed to take into account were choosing one ecological indicator, aggregating sampling data over seven years, assessing the effect of the monetary value on the calculations and scaling the monetary value to the size of a lake. From an environmental management point of view, the main results are that the monitoring is cost-effective especially when the lake is a priori in target status.

The VOI analysis provides a novel tool for lake and other environmental managers to estimate the value of additional monitoring data for a particular, single case, e.g. a lake, when an additional benefit is attainable through remedial management actions. In such a case, decision makers should have a prior knowledge about the present status of e.g. a lake and about the value of the desired outcome, e.g. good ecological status. Further, knowledge on the $VOI(y)$ in management scenarios is useful and can be extended also to other environmental contexts thus expanding the work of e.g. Nygård et al. (2016).

While we gained important insights, in our study we focused on traditional water sampling data. However, there are emerging techniques of collecting environmental data (e.g. remote sensing) which have been hailed as potential alternatives for future monitoring. Assessing the VOI of these alternative data sources is important and would allow to identify the most effective and cost-efficient ways to monitor and assess the state of European inland and coastal waters.

CRediT authorship contribution statement

Vilja Koski: Methodology, Software, Formal analysis, Investigation, Visualization, Writing - original draft. **Niina Kotamäki:** Conceptualization, Validation, Investigation, Resources, Visualization, Writing - review & editing. **Heikki Hämäläinen:** Validation, Investigation, Writing - review & editing. **Kristian Meissner:** Validation, Resources, Writing - review & editing. **Juha Karvanen:** Conceptualization, Validation, Formal analysis, Visualization, Writing - review & editing. **Salme Kärkkäinen:** Conceptualization, Methodology, Supervision, Writing - original draft, Project administration.

Acknowledgements

Vilja Koski and Salme Kärkkäinen were supported by the Academy of Finland (grant number 289076). Kristian Meissner was supported by BONUS FUMARI: BONUS (art. 185), which is jointly funded by the EU, the Academy of Finland and Swedish Research Council Formas. Niina Kotamäki was supported by Strategic Research Council of Academy of Finland (grant number 312650). The work is related to the thematic research area "Decision analytics utilizing causal models and multiobjective optimization" (DEMO) of University of Jyväskylä supported by Academy of Finland (grant number 311877).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. The change point at value of perfect information with respect to the monitoring value

Let a_0 and a_1 be two alternatives of a decision situation and the uncertainty $x \in \{x_0, x_1\}$ be discrete, $p = p(x_1)$ is constant. Values of scenarios can be chosen for example as in Table 4: $v = v(x_1, a_0) \geq 0$ is the value of lake in target ecological status set by decision maker, $r \in [0,1]$ is the ratio from value v of how much a management improves a status of lake not in target status and $c_1 \geq 0$ is the (constant) cost of implementing alternative a_1. In some situations, r could be alternatively interpreted as the probability of achieving the target status after the management. The cost of implementing alternative a_0 is $c_0 = 0$. In this case, the posterior value $PoV(x)$ is an increasing function of v.

According to the Eq. (3), the posterior value in our situation is

$$\begin{aligned} PoV(x) &= \max\{0, rv - c_1\} \cdot (1-p) + \max\{v, v - c_1\} \cdot p \\ &= \begin{cases} 0 + pv, & \text{if } rv - c_1 < 0 \Leftrightarrow v < \frac{c_1}{r} \\ (rv - c_1)(1-p) + pv, & \text{otherwise.} \end{cases} \end{aligned} \tag{6}$$

Furthermore, the prior value PV is a piecewise-defined linear function, as follows. According to the Eq. (2), the expected values (prior values) of the two alternatives are

$$\begin{aligned} \mathrm{E}(v(x, a_0)) &= (1-p) \cdot 0 + p \cdot v \\ &= pv \end{aligned} \tag{7}$$

and

$$\begin{aligned} \mathrm{E}(v(x, a_1)) &= (1-p) \cdot (rv - c_1) + p \cdot (v - c_1) \\ &= (r - rp + p)v - c_1. \end{aligned} \tag{8}$$

The prior value is the maximum of these two expectations:

$$\begin{aligned} PV &= \max_{a \in A} \{\mathrm{E}(v(x, a_0)), \mathrm{E}(v(x, a_1))\} \\ &= \max_{a \in A} \{pv, (r - rp + p)v - c_1\} \\ &= \begin{cases} pv, & \text{if } v \leq \frac{c_1}{r - rp} \\ (r - rp + p)v - c_1, & \text{otherwise.} \end{cases} \end{aligned} \tag{9}$$

Thus, the value of perfect information is

$$\begin{aligned} VOI(x) &= PoV(x) - PV \\ &= \begin{cases} 0, & \text{if } v \leq \frac{c_1}{r} \\ (r - rp + p)v - (1-p)c_1, & \text{if } \frac{c_1}{r} < v \leq \frac{c_1}{r - rp} \\ pc_1, & \text{if } v > \frac{c_1}{r - rp} \end{cases} \end{aligned} \tag{10}$$

The change point of $VOI(x)$ can be found at the change point of the piecewise-defined function of PV, and it is $v = c_1/(r - rp)$.

References

Ahtiainen, H., 2008. Järven tilan parantamisen hyödyt. Esimerkkinä Hiidenvesi. Finnish Environment Institute (SYKE), Helsinki. ISBN: 978-952-11-3284-1 In Finnish. Available online at. https://helda.helsinki.fi/handle/10138/38353, Accessed date: 22 May 2019.

Aroviita, J., Mitikka, S., Vienonen, S., 2019. Pintavesien tilan luokittelu ja arviointiperusteet vesienhoidon kolmannella kaudella. Finnish Environment Institute (SYKE), Helsinki ISBN 978-952-11-5074-6. (In Finnish.) Available online at. https://helda.helsinki.fi/handle/10138/306745, Accessed date: 17 February 2020.

Artell, J., Huhtala, A., 2017. What are the benefits of the water framework directive? Lessons learned for policy design from preference revelation. Environ. Resour. Econ. 68 (4), 847–873 ISSN 1573-1502. doi: 10.1007/s10640-016-0049-8. URL. Dec. https://doi.org/10.1007/s10640-016-0049-8.

Atkins, J.P., Burdon, D., 2006. An initial economic evaluation of water quality improvements in the Randers Fjord, Denmark. Mar. Pollut. Bull. 53 (1), 195–204. https://doi.org/10.1016/j.marpolbul.2005.09.024 ISSN 0025-326X. URL. http://www.sciencedirect.com/science/article/pii/S0025326X05004108 Recent Developments in Estuarine Ecology and Management.

Bolam, F.C., Grainger, M.J., Mengersen, K.L., Stewart, G.B., Sutherland, W.J., Runge, M.C., McGowan, P.J.K., 2019. Using the value of information to improve conservation decision making. Biol. Rev. 94 (2), 629–647. https://doi.org/10.1111/brv.12471. https://onlinelibrary.wiley.com/doi/abs/10.1111/brv.12471.

Brander, L.M., Florax, R.J.G.M., Vermaat, J.E., 2006. The empirics of wetland valuation: a comprehensive summary and a meta-analysis of the literature. Environ. Resour. Econ. 33 (2), 223–250 ISSN 1573-1502. doi: 10.1007/s10640-005-3104-4. URL doi. Feb. https://doi.org/10.1007/s10640-005-3104-4.

Canessa, S., Guillera-Arroita, G., Lahoz-Monfort, J.J., Southwell, D.M., Armstrong, D.P., Chadés, I., Lacy, R.C., Converse, S.J., 2015. When do we need more data? A primer on calculating the value of information for applied ecologists. Methods Ecol. Evol. 6 (10), 1219–1228. https://doi.org/10.1111/2041-210X.12423. https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12423.

Carson, R.T., Conaway, M.B., Hanemann, M.W., Krosnick, J.A., Mitchell, R.C., Presser, S., 2004. Valuing Oil Spill Prevention: A Case Study of California's Central Coast. Kluwer Academic Publishers, Boston (ISBN 978-0-7923-6497-9).

Carstensen, J., Lindegarth, M., 2016. Confidence in ecological indicators: a framework for quantifying uncertainty components from monitoring data. Ecol. Indic. (ISSN: 1470-160X) 67, 306–317. https://doi.org/10.1016/j.ecolind.2016.03.002 URL. http://www.sciencedirect.com/science/article/pii/S1470160X16301066.

Carvalho, L., Mackay, E.B., Cardoso, A.C., Baattrup-Pedersen, A., Birk, S., Blackstock, K.L., Borics, G., Borja, A., Feld, C.K., Ferreira, M.T., Globevnik, L., Grizzetti, B., Hendry, S., Hering, D., Kelly, M., Langaas, S., Meissner, K., Panagopoulos, Y., Penning, E., Rouillard, J., Sabater, S., Schmedtje, U., Spears, B.M., Venohr, M., van de Bund, W., Solheim, A.L., 2019. Protecting and restoring Europe's waters: an analysis of the future development needs of the water framework directive. Sci. Total Environ. 658, 1228–1238. https://doi.org/10.1016/j.scitotenv.2018.12.255 ISSN 0048-9697. URL. http://www.sciencedirect.com/science/article/pii/S004896971835126X.

Colyvan, M., Sep 2016. Value of information and monitoring in conservation biology. Environ. Syst. Decis. (ISSN: 2194-5411) 36 (3), 302–309. https://doi.org/10.1007/s10669-016-9603-8 (URL doi:10.1007/s10669-016-9603-8).

Efron, B., Tibshirani, R.J., 1993. An introduction to the bootstrap. Number 57 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, Boca Raton, Florida, USA.

Eidsvik, J., Mukerji, T., Bhattacharjya, D., 2015. Value of Information in the Earth Sciences: Integrating Spatial Modeling and Decision Analysis. Cambridge University Press, Cambridge.

European Communities, 2003. Common Strategy on the Implementation of the Water Framework Directive (2000/60), Guidance Document No. 13, Overall Approach to the Classification of Ecological Status and Ecological Potential.

European Parliament, 2000. Directive 2000/60/EC, of the European Parliament and council of 23 October 2000 establishing a framework for community action in the field of water policy. URL. http://eur-lex.europa.eu/resource.html?uri=cellar:5c835afb-2ec6-4577-bdf8-756d3d694eeb.0004.02/DOC_1 format=PDF.

Eyvindson, K., Hakanen, J., Mönkkönen, M., Juutinen, A., Karvanen, J., 2019. Value of information in multiple criteria decision making: an application to forest conservation. Stoch. Env. Res. Risk A. 33 (11−12), 2007–2018. https://doi.org/10.1007/s00477-019-01745-4.

French, S., Maule, J., Papamichail, N., 2010. Decision Behaviour, Analysis and Support. Cambridge University Press, Cambridge 9780511609947, p. 02. https://doi.org/10.1017/CBO9780511609947.

Hjerppe, T., Seppälä, E., Väisänen, S., Marttunen, M., 2017. Monetary assessment of the recreational benefits of improved water quality – description of a new model and a case study. J. Environ. Plan. Manag. 60 (11), 1944–1966. https://doi.org/10.1080/09640568.2016.1268108.

Holling, C., 1978. Adaptive Environmental Assessment and Management. 01. John Wiley & Sons Ltd, New York.

Neumann, J. von, Morgenstern, O., 1944. Theory of Games and Economic Behavior. Princeton University Press, New Jersey 9780691130613 http://www.jstor.org/stable/j.ctt1r2gkx.

Nygård, H., Oinonen, S., Hällfors, H.A., Lehtiniemi, M., Rantajärvi, E., Uusitalo, L., 2016. Price vs. value of marine monitoring. Front. Mar. Sci. 3 (205) doi: 10.3389.

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria https://www.R-project.org/.

Reynaud, A., Lanzanova, D., 2017. A global meta-analysis of the value of ecosystem services provided by lakes. Ecol. Econ. 137, 184–194. https://doi.org/10.1016/j.ecolecon.2017.03.001 ISSN 0921-8009. http://www.sciencedirect.com/science/article/pii/S0921800916309168.

Robert, C.P., Casella, G., 2005. Monte Carlo statistical methods. Springer Texts in Statistics, 2nd ed. Springer, Berlin URL. https://cds.cern.ch/record/1187871.

Royston, P., Altman, D.G., Sauerbrei, W., 2006. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat. Med. 25 (1), 127–141. https://doi.org/10.1002/sim.2331. https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.2331.

Sala, O.E., Chapin III, S.F., Armesto, J.J., Berlow, E., Bloomfield, J., Dirzo, R., Huber-Sanwald, E., Huenneke, L.F., Jackson, R.B., Kinzig, A., Leemans, R., Lodge, D.M., Mooney, H.A., Oesterheld, M., Poff, N.L., Sykes, M.T., Walker, B.H., Walker, M., Wall, D.H., 2000. Global biodiversity scenarios for the year 2100. Science (ISSN: 0036-8075) 287 (5459), 1770–1774. https://doi.org/10.1126/science.287.5459.1770. http://science.sciencemag.org/content/287/5459/1770.

Schlaifer, R., Raiffa, H., 1961. Applied Statistical Decision Theory. Harvard University, Boston.

Steuten, L., van de Wetering, G., Groothuis-Oudshoorn, C., Retèl, V., 2013. A systematic and critical review of the evolving methods and applications of value of information in academia and practice. PharmacoEconomics 31 (01), 25–48. https://doi.org/10.1007/s40273-012-0008-3.

Williams, B., Eaton, M.J., Breininger, D., 2011. Adaptive resource management and the value of information. Ecol. Model. 222, 3429–3436 09. https://doi.org/10.1016/j.ecolmodel.2011.07.003.

Yokota, F., Thompson, K.M., 2004. Value of information analysis in environmental health risk management decisions: past, present, and future. Risk Anal. 24 (3), 635–650 ISSN 1539-6924. doi: 10.1111/j.0272-4332.2004.00464.x. URL. 6. https://doi.org/10.1111/j.0272-4332.2004.00464.x.

II

SUBSAMPLE SELECTION METHODS IN THE LAKE MANAGEMENT

by

Koski, V., Kärkkäinen, S. & Karvanen, J. 2024

JABES, DOI: https://doi.org/10.1007/s13253-024-00630-0

Subsample Selection Methods in the Lake Management

Vilja KOSKI, Salme KÄRKKÄINEN, and Juha KARVANEN

The problem of subsample selection among an enormous number of combinations arises when some covariates are available for all units, but the response can be measured only for a subset of them. When estimating a Bayesian prediction model, optimized selections can be more efficient than random sampling. The work is motivated by environmental management of aquatic systems. We consider data on 4360 Finnish lakes and aim to find an approximately optimal subsample of lakes in the sense of Bayesian D-optimality. We study Bayesian two-stage selection where the choice of lakes to be measured at the second stage depends on the measurements carried out at the first stage. The results indicate that the two-stage approach has a modest advantage compared to the single-stage approach.

Key Words: Approximate design; Bayesian design; Information matrix; Optimal design; Optimality criteria; Utility function.

1. INTRODUCTION

The problem of subsample selection is always present when it is not possible to measure the whole population of interest. The most common approach to subsample selection is a simple random sampling. However, when the aim is to estimate a prediction model fitted to the subsample as precisely as possible, a carefully selected nonrandom sample may provide higher expected information per unit than a random sample.

The practical motivation for this work arises from the environmental management of aquatic systems. We consider the data on 4360 Finnish lakes containing variables from monitoring data and register-based data. This dataset represents large data from which a subsample will be selected. Based on the monitoring data gathered, each lake either has a healthy water quality status or is in the need of management actions to improve the status. Our interest is to model the relationship between the hard-to-measure quality status and the easily available register-based lake features using a Bayesian logistic regression model. The

V. Koski (✉)· S. Kärkkäinen · J. Karvanen
Department of Mathematics and Statistics, University of Jyvaskyla, P.O. Box 35, 40014 Jyvaskyla, Finland
(E-mail: *vilja.a.koski@jyu.fi*).

Journal of Agricultural, Biological, and Environmental Statistics
https://doi.org/10.1007/s13253-024-00630-0

Published online: 04 June 2024

aim is to select such a set of lakes that we can estimate the model parameters as accurately and precisely as possible. However, the data acquisition in the current monitoring program is considered being costly, and that is why the selection of lakes must be made wisely. The question is how to select the lakes to be monitored, if only a subset of them can be measured.

In general, the selection problem considered here is NP-hard (Welch 1982) because of the massive number of combinations. When the number of lakes is n, there are 2^n possible subsets if no constraints are applied. A truly optimal selection would require analysing all those available combinations, which understandably is not possible and heuristic selection algorithms are needed.

Optimal subsample selection is related to optimal experimental design (Ryan et al. 2016; Atkinson et al. 2007; Chaloner and Verdinelli 1995) although there are some substantial differences. In both problems, the goal is to find a design that maximizes (or minimizes) an optimality criterion that usually is a function of the information matrix. A popular and well-known criterion is D-optimality, which is equivalent to maximising the determinant of the information matrix. In the case of a nonlinear model (Pronzato and Pázman 2013), the optimal design depends on the unknown model parameters in both problems. Wynn (1982) and Fedorov (1989) consider optimal design when there are restrictions for the density of the design measure. Pronzato and Wang (2021) propose a sequential subsampling method.

Optimal experimental designs are often found using a point-exchange algorithm (Fedorov 1972) or a coordinate-exchange algorithm (Meyer and Nachtsheim 1995). In subsample selection, candidate points are restricted to those in the data and replicates are not available, as in optimal experimental design. This implies that coordinate-exchange algorithms are not applicable, while point-exchange algorithms could be used. In earlier works, approximately optimal subsamples have been found for instance by the greedy method (Reinikainen et al. 2016; Reinikainen and Karvanen 2022), a randomized search (Paglia et al. 2022) or a two-step algorithm (Zuo et al. 2021). We focus on methods that are scalable in sense that the computational load remains bearable in the Bayesian setting even if the size of the population and the size of the subsample increase, and use therefore the greedy method.

In this paper, we study a subsample selection in Finnish lake monitoring setting with the aim of estimating the parameters of a Bayesian logistic regression model predicting the ecological status of a lake as precisely as possible. As the model is nonlinear, the optimal selection depends on the model parameters via the information matrix. When prior information is scarce, the initial model parameters may not be good enough to describe the phenomenon correctly. Updating the model sequentially (Pronzato 2006) after each measurement solves the problem theoretically better but is not feasible in lake management because of the time needed to analyse water samples. A two-stage strategy (Sitter and Forbes 1997; Montepiedra and Yeh 1998; Guillera-Arroita et al. 2014; Pronzato and Pázman 2013) is a practical compromise that helps to recover from poor prior information but keeps the process relatively straightforward.

Plenty of research is available on two-stage methods in optimal experimental design. Ruggoo and Vandebroek (2004) implement a two-stage procedure to improve the model estimates in Bayesian optimal experimental design. Karvanen (2009) considers the cost of the experiment as a design criterion in a sequential design selection. Reilly (1996) studies a general two-stage design in epidemiology, where the response and some easily obtained

covariates are available on the first stage, while the more expensive covariates are ascertained only for the subsample of second stage subjects.

In the two-stage selection, we use prior data to select lakes to be measured and update the model with these measurements (first stage). Then we use the updated model to select more lakes to be measured (second stage) and combine all measurements to obtain the final estimates. By updating the model once, we assume to reach a better result than if we had selected the lakes using only the prior information. We are interested in learning how the accuracy of the final estimates behaves as a function of the number of lakes selected at the first stage when the total sample size is fixed.

The paper is organized as follows: Sect. 2 introduces a compound dataset of lake monitoring data and register-based data containing lake features. In Sect. 3, we present notations and derive the D-optimal selection criterion for a logistic regression model. In Sect. 4, we introduce a greedy forward selection algorithm. In Sect. 5, the greedy forward algorithm is applied to the real dataset. Section 6 concludes with discussion about the results and future directions.

2. DATA

As a motivating real-life example, we consider an optimal design problem connected to lake monitoring data in Finland. Thanks to the Water Framework Directive (WFD) of European Union (European Parliament 2000), the monitoring program is implemented to improve and to secure the quality of inland waters in EU. Regular and long-term monitoring data of parameters representing biotic structure, supported by the physical and chemical properties of water and hydrological and morphological features, are used to classify the waters into ecological status classes (European Communities 2003). For each classification variable, the status class is assessed against the degree of deviance from the pre-determined reference conditions (Aroviita et al. 2019), and also the expert's opinions about the status of a lake affects the classification. According to the directive, management actions are needed to implement to improve the ecological status if the lake is in moderate status or less. The collecting and analysing the monitoring data, however, has been considered to be expensive. That is why our goal is selecting lakes to be monitored to save resources reserved to data acquisition. In the value of information context, the lake monitoring has been considered by Koski et al. (2020) for one lake and by Koski and Eidsvik (2024) for several lakes.

We use data about the latest ecological status classification of lakes in Finland based on the monitoring data from the years 2012–2017. The classification is available as an open-source data maintained by Finnish Environment Institute (http://www.syke.fi/en-US/Open_information). Since the demand of management actions is our main interest, we are interested in the status class based on the need of management actions (Fig. 1, left). In addition, we utilise data from other free sources to support the knowledge about the lakes and their characteristics when modelling the ecological status.

According to Finnish Environment Institute, there are about 187,000 lakes in Finland (Heiskanen et al. 2017). The total number is depending on the definition of a lake: the lake area and the stability of water. We have the status classification for $N = 4360$ lakes. Of

the lakes, 174 are located in Helsinki-Uusimaa region, 445 located in Southern Finland, 656 in Western Finland and 3085 in Northern and Eastern Finland. For this study, we excluded the lakes from Åland due to small sample sizes and from other regions the lakes that comprises many water bodies, have large area or are important from other reasons, for example Finland's largest lake Saimaa. The latter ones are anyway monitored frequently. We have the status classification for these $N = 4360$ lakes already available, but in our example we simulate a situation where the status is yet to be determined.

In the current study, the ecological status of the lake is modelled by using lake features which are easily available from data sources. Basic features of the lakes having area over one hectare can be found from the interface maintained by Finnish Environment Institute (https://www.syke.fi/en-US/Open_information/Open_web_services/Environmental_data_API). From this interface, we uploaded for each lake the information about the location (the municipality, drainage basin and centre latitude and longitude coordinates) and basic information about lake features, such as waterbed area (hectares) length of shoreline (kilometres), average and maximum depth (meters), volume of water mass (1000 cubic meters) and altitude above sea level (meters). The circularity of the lake was calculated as the ratio of the circumference (of a circular lake) to the length of shoreline. In addition to these variables, we have an agricultural area (Official Statistics of Finland 2020a) and a number of free-time residences in the municipality where each lake is located (Official Statistics of Finland 2020b). We divided the agricultural area and the number of free-time residences of municipalities by the area of the municipality to obtain the percentage of agricultural area in each municipality and the density of free-time residences (Fig. 1, right). As the result of the model selection, we chose as covariates the waterbed area and the agricultural land in the municipality the lake is located for the model used in subsample selection.

3. BAYESIAN TWO-STAGE SELECTION FOR LOGISTIC REGRESSION

3.1. TWO-STAGE DESIGNS

A general two-stage procedure has the following steps:

1. Based on the initial data, the initial parameter estimates are estimated. Based on them, choose an optimal design (first stage selection).
2. Collect the data according the first stage design and update the initial parameter estimates (first stage analysis).
3. Based on the initial data and the data collected on the first stage, choose an optimal design for additional data collection (second stage selection).
4. Collect the data according the second stage design and analyse the full dataset, from both stages and the initial data, to obtain the final estimates (second stage analysis).

In next sections, we will describe a Bayesian implementation of this procedure.

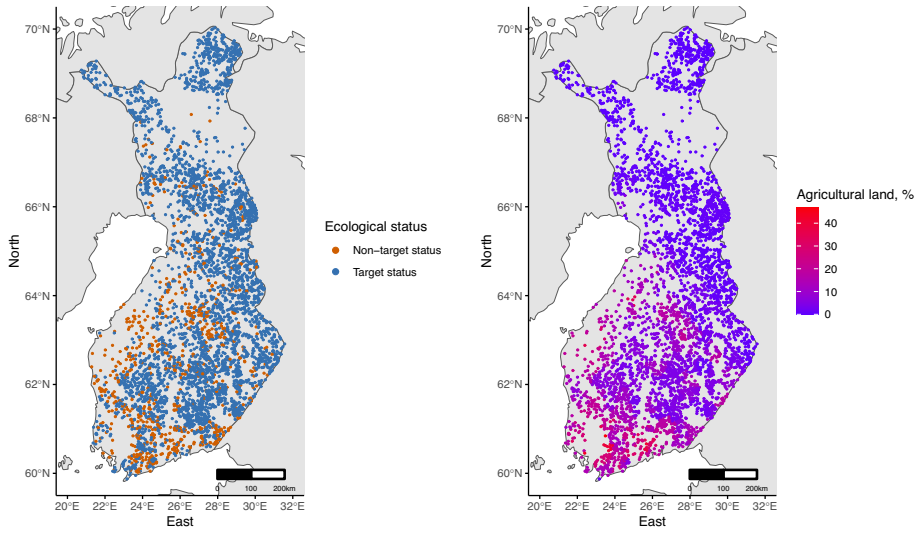


Figure 1. Left: The ecological status classification on 4360 lakes on the map of Finland where the original five classes are reduced into binary based on the demand of management actions: nontarget status replacing status classes moderate, poor or bad and target status replacing classes high or good. Right: Data on 4360 lakes on the map of Finland by the percentual amount of agricultural land in the municipality where the lake is located.

3.2. BAYESIAN SUBSAMPLE SELECTION

Let N be the number of units, for example lakes as in our case, in the population of interest and let a set of covariates to be available for the whole population. First, let us assume that we have already measured a $(n_0 \times 1)$ response vector $\boldsymbol{y}_0$ (the ecological status) for a small subset S_0 of size n_0. Furthermore, let $\boldsymbol{x}_0 = \{x_{0ij}\}$, $j = 1, \ldots, J$, and $i = 1, \ldots, n_0$, be the $(n_0 \times J)$ covariate matrix representing the J covariates.

In the Bayesian framework, we have a likelihood $p(\boldsymbol{y}_0|\boldsymbol{x}_0, \boldsymbol{\beta})$ where the model parameters $\boldsymbol{\beta}$ are defined in parameter space Θ. The prior distribution is denoted by $p(\boldsymbol{\beta})$, and it is obtained from the initial knowledge. The posterior probability distribution $p(\boldsymbol{\beta}|\boldsymbol{y}_0, \boldsymbol{x}_0)$ is proportional to the product $p(\boldsymbol{y}_0|\boldsymbol{x}_0, \boldsymbol{\beta})p(\boldsymbol{\beta})$ and describes our current knowledge of the model parameters. In the later analysis, $p(\boldsymbol{\beta}|\boldsymbol{y}_0, \boldsymbol{x}_0)$ is considered as a prior distribution.

First, as the knowledge about the model parameters $\boldsymbol{\beta}$ is still poor based on the prior, we are interested in selecting a subset from the $n = N - n_0$ units with unknown response to improve our knowledge. We have the $(n \times J)$ covariate matrix $\boldsymbol{x}$ that we have observed, but the $(n \times 1)$ response vector $\boldsymbol{y}$ in space $\mathcal{Y}$ is still remaining unmeasured. In order to increase the accuracy of the modelling, we aim to select a subset S_1 of size $n_1 < n$ from the population from which we measure the $(n_1 \times 1)$ vector $\boldsymbol{y}_1$ and thus have a set $S_0 \cup S_1$ measured. This is called the first stage selection. Using the prior and the data collected at the first stage, we obtain the posterior $p(\boldsymbol{\beta}|\boldsymbol{y}_0, \boldsymbol{y}_1, \boldsymbol{x}_0, \boldsymbol{x}_1) \propto p(\boldsymbol{y}_0, \boldsymbol{y}_1|\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{\beta})p(\boldsymbol{\beta})$ that will be applied as a prior at the second stage.

At the second stage selection, we are interested in selecting a subset from the $N - n_0 - n_1$ units with unknown response. We aim to select a subset S_2 of size $n_2 < n - n_1$ from the

population that is still remaining unmeasured. Thus, we want to have a set $S_0 \cup S_1 \cup S_2$ measured in total.

The selected subsamples are specified by a measurement plan $\boldsymbol{r} = (r_1, r_2, \ldots, r_n)^\mathsf{T}$, where $r_i = 1$ if the response y_i is planned to be measured for a unit i and $r_i = 0$ otherwise. In other words, $S_1 \cup S_2$ is a set of units $i = 1, \ldots, n$ with $r_i = 1$. With the general notation for missing data, we may write

$$y(r_i) = \begin{cases} y_i, & \text{if } r_i = 1 \\ \text{NA}, & \text{if } r_i = 0. \end{cases}$$

The predictive distribution of the response $\boldsymbol{y}(\boldsymbol{r})$ before the data are collected according to a measurement plan can be written at the first stage as

$$p(\boldsymbol{y}(\boldsymbol{r})|\boldsymbol{x}, \boldsymbol{y}_0, \boldsymbol{x}_0) = \int_\Theta p(\boldsymbol{y}(\boldsymbol{r})|\boldsymbol{x}, \boldsymbol{\beta}) p(\boldsymbol{\beta}|\boldsymbol{y}_0, \boldsymbol{x}_0) \mathrm{d}\boldsymbol{\beta}, \tag{1}$$

and at the second stage as

$$p(\boldsymbol{y}(\boldsymbol{r})|\boldsymbol{x}, \boldsymbol{y}_0, \boldsymbol{y}_1, \boldsymbol{x}_0, \boldsymbol{x}_1) = \int_\Theta p(\boldsymbol{y}(\boldsymbol{r})|\boldsymbol{x}, \boldsymbol{\beta}) p(\boldsymbol{\beta}|\boldsymbol{y}_0, \boldsymbol{y}_1, \boldsymbol{x}_0, \boldsymbol{x}_1) \mathrm{d}\boldsymbol{\beta}, \tag{2}$$

where $p(\boldsymbol{y}(\boldsymbol{r})|\boldsymbol{x}, \boldsymbol{\beta})$ is the model for the new data.

Our aim is to select the subsamples S_1 and S_2 in an optimal way. The optimality is defined in terms of a utility function, which we denote in the general case as $U(\beta, \boldsymbol{y}(\boldsymbol{r}), \boldsymbol{x})$. Given n_1, the optimal subsample is found at the first stage when we find a measurement plan $\boldsymbol{r}$ that maximises

$$\bar{U}_1(\boldsymbol{r}) = \int_\mathcal{Y} \left[\int_\Theta U(\boldsymbol{\beta}, \boldsymbol{y}(\boldsymbol{r}), \boldsymbol{x}) p(\boldsymbol{\beta}|\boldsymbol{y}(\boldsymbol{r}), \boldsymbol{x}) \mathrm{d}\boldsymbol{\beta} \right] p(\boldsymbol{y}(\boldsymbol{r})|\boldsymbol{x}, \boldsymbol{y}_0, \boldsymbol{x}_0) \mathrm{d}\boldsymbol{y} \tag{3}$$

given the constraint $\sum_{i=1}^n r_i = n_1$. Then, given n_2, the optimal subsample is found at the second stage when we find a measurement plan $\boldsymbol{r}$ that maximises

$$\bar{U}_2(\boldsymbol{r}) = \int_\mathcal{Y} \left[\int_\Theta U(\boldsymbol{\beta}, \boldsymbol{y}(\boldsymbol{r}), \boldsymbol{x}) p(\boldsymbol{\beta}|\boldsymbol{y}(\boldsymbol{r}), \boldsymbol{x}) \mathrm{d}\boldsymbol{\beta} \right] p(\boldsymbol{y}(\boldsymbol{r})|\boldsymbol{x}, \boldsymbol{y}_0, \boldsymbol{y}_1, \boldsymbol{x}_0, \boldsymbol{x}_1) \mathrm{d}\boldsymbol{y} \tag{4}$$

given the constraint $\sum_{i=1}^n r_i = n_1 + n_2$ and $r_i = 1$ for units already selected at the first stage. The posterior probability distribution $p(\boldsymbol{\beta}|\boldsymbol{y}(\boldsymbol{r}), \boldsymbol{x})$ for the model parameters $\boldsymbol{\beta}$ could be estimated for example with importance sampling.

In our analysis, we use an approximation of Eq. (4) (Chaloner and Verdinelli 1995)

$$\bar{U}_2(\boldsymbol{r}) \approx \int_\mathcal{Y} \left[\int_\Theta U(\boldsymbol{\beta}, \boldsymbol{y}(\boldsymbol{r}), \boldsymbol{x}) p(\boldsymbol{\beta}|\boldsymbol{y}_0, \boldsymbol{y}_1, \boldsymbol{x}_0, \boldsymbol{x}_1) \mathrm{d}\boldsymbol{\beta} \right] p(\boldsymbol{y}(\boldsymbol{r})|\boldsymbol{x}, \boldsymbol{y}_0, \boldsymbol{y}_1, \boldsymbol{x}_0, \boldsymbol{x}_1) \mathrm{d}\boldsymbol{y}, \tag{5}$$

where the posterior $p(\boldsymbol{\beta}|\boldsymbol{y}(\boldsymbol{r}), \boldsymbol{x})$ is replaced with the prior probability distribution $p(\boldsymbol{\beta}|\boldsymbol{y}_0, \boldsymbol{y}_1, \boldsymbol{x}_0, \boldsymbol{x}_1)$ for the model parameters $\boldsymbol{\beta}$. Equation (3) is approximated in a similar way.

In our case, the utility function measures the precision for the parameters of interest as a selection criterion. Thus, as an D-optimality criterion, the selection criterion maximizes the logarithm of the determinant of the expected information matrix

$$U(\boldsymbol{\beta}, \boldsymbol{y}(\boldsymbol{r}), \boldsymbol{x}) = \log\det(\mathcal{I}(\boldsymbol{\beta}|\boldsymbol{x}, \boldsymbol{r})) = \log\det\left(\sum_{i=1}^{n} \mathcal{I}(\boldsymbol{\beta}|x_i, r_i)\right), \tag{6}$$

where

$$\mathcal{I}(\boldsymbol{\beta}|x_i, r_i) = \begin{cases} \mathcal{I}(\boldsymbol{\beta}|x_i), & r_i = 1 \\ 0, & r_i = 0 \end{cases}$$

is the expected information (Chaloner and Verdinelli 1995). In the frequentist approach, the distribution $p(\boldsymbol{\beta}|\boldsymbol{y}_0, \boldsymbol{y}_1, \boldsymbol{x}_0, \boldsymbol{x}_1)$ would not be used, but the value of $\boldsymbol{\beta}$ is fixed to its current maximum likelihood estimate $\hat{\boldsymbol{\beta}}$ and $U(\hat{\boldsymbol{\beta}}, \boldsymbol{y}(\boldsymbol{r}), \boldsymbol{x})$ measures the utility of measurement plan $\boldsymbol{r}$.

The `brms`-package (Bürkner 2017) is used to implement the Bayesian model fitting, which provides an `R` (R Core Team 2023) interface to `Stan` (Stan Development Team 2022). The package uses Markov chain Monte Carlo (MCMC) algorithms to draw random samples from the posterior to estimate the model parameters $\boldsymbol{\beta}$. The sampling is implemented via adaptive Hamiltonian Monte Carlo (Hoffman and Gelman 2014).

3.3. EXPECTED INFORMATION MATRIX FOR A LOGISTIC MODEL

The subset selection with the criterion of Eq. (6) is studied further in the context of generalised linear models. We assume a binary response $y_i, i = 1, \ldots n$, that is distributed as

$$\begin{aligned} P(y_i = 1 \mid \boldsymbol{x}_i) &= \pi_i, \\ P(y_i = 0 \mid \boldsymbol{x}_i) &= 1 - \pi_i, \end{aligned}$$

where the parameter π_i is linked to the set of covariates with a link function $g(\pi_i) = \eta_i$ as follows:

$$\begin{aligned} g(\pi_i) &= \text{logit}(\pi_i) = \eta_i, \\ \pi_i &= \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}. \end{aligned}$$

Here, the response y_i is modelled with a linear predictor $\eta_i = \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}$ and $\boldsymbol{\beta} = \{\beta_j\}$, $j = 1, \ldots, J$, is an unknown parameter vector that is needed to estimate. The log-likelihood of the data $\{y_i, \boldsymbol{x}_i^{\mathsf{T}}\}$ is obtained as

$$\log p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \log p(y_i|x_i, \boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \log\left(\frac{\exp(\eta_i)}{1 + \exp(\eta_i)}\right)$$

$$+ (1 - y_i) \log \left(1 - \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}\right)$$
$$= \sum_{i=1}^{n} y_i \boldsymbol{x}_i^\mathsf{T} \boldsymbol{\beta} - \log(1 + \exp(\boldsymbol{x}_i^\mathsf{T} \boldsymbol{\beta})).$$

Next, we give the form for the expected information matrix. The expected information matrix (the Fisher information matrix) tells how much information the data are expected to contain. It is used when the collection of the data is still being planned. Formally, the expected information matrix of a generalized linear model is the expected value of the observed information:

$$\mathcal{I}(\boldsymbol{\beta}|\boldsymbol{x}, \boldsymbol{r}) = -\sum_{i=1}^{n} r_i \, \mathbb{E}\left(\frac{\partial^2 \log p(y_i|x_i, \boldsymbol{\beta})}{\partial \beta \partial \beta^\mathsf{T}}\right), \tag{7}$$

where the expectation is taken with respect to response y_i. In our case, for a binary response and the logit link, the (j, k) element of the matrix in Eq. (7), also in Eq. (6), takes the form

$$\mathcal{I}(\beta|x_i, r_i) = -r_i \, \mathbb{E}\left[\frac{\partial^2 \log p(y_i|x_i, \boldsymbol{\beta})}{\partial \beta_j \partial \beta_k^\mathsf{T}}\right] = r_i \left[\frac{\exp(\eta_i)}{1 + \exp(\eta_i)}\left(1 - \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}\right)\right] x_{ij} x_{ik}. \tag{8}$$

4. ALGORITHMS FOR FINDING D-OPTIMAL SELECTION

Since our selection problem has a massive number of combinations and all of these combinations are not possible to analyse, we need heuristic algorithms to find approximately optimal designs. Here, we chose to use the greedy forward selection in the two-stage approach to find D-optimal designs. The greedy approach, also known as a sequential search (Dykstra 1971), is well-known by mathematicians and computer scientists. The idea of the method is to sequentially add units to the design, optimizing the selection criterion only for the selection of a single unit at each round until the desired size n_1 is reached. In many problems, a greedy strategy does not produce an optimal solution, but a greedy heuristic can yield locally optimal solutions that approximate the globally optimal solution and can be found in a reasonable running time. The forward selection (Algorithm 1) starts with no additional units selected and adds the most promising unit until desired size n_1 is reached. If two units give the same criterion value U, the selection between them is performed randomly. If $n_1 = n$, the procedure will sort all units into the preferred selection order. The algorithm was implemented in R R Core Team (2023).

5. APPLICATION TO FINNISH LAKE DATA

We first introduce the setting to study the performance of the two-stage design selection. Next, we present results on the planning of the two-stage selection in the Bayesian framework. We report the model parameter estimates for the best scenario for two-stage design and compare them to the single-stage situation.

Algorithm 1: Greedy forward selection.

Input : The number of units already selected $k' = 0$, the total number of units k, the number of units to be selected k_1, model parameters $\boldsymbol{\beta}$, the measurement plan $r_i = 0$, $i = 1, \ldots, k$, the candidate measurement plan $r_i^* = 0$, $i = 1, \ldots, k$

Output: The measurement plan $\boldsymbol{r} = (r_1, \ldots, r_k)^\intercal$ with $\sum_{i=1}^{k} r_i = k_1$

```
while k' ≤ k_1 do
    for i = 1, ..., k do
        if r_i = 0 then
            r_i^* = 1 ;
            Calculate I_i(β|x, r) = Σ_{j:r_j=1} I(β|x, r_j) + I(β|x, r_i^*) ;
            Calculate U_i = U(β, y(r), x) as in Eq. (6);
            r_i^* = 0
        end if
    end for
    Find the unit i that maximises the utility U_i and then set r_i = 1, k' = k' + 1.
end while
return r
```

5.1. STUDY SETTING

We applied the two-stage selection described in Sect. 3 and the greedy forward selection algorithm described in Sect. 4 to Finnish lake management problem introduced in Sect. 2. We aimed to imitate the real-life decision-making process of monitoring data gathering and fixed the initial data. It is realistic to assume that if any initial measurements are already available, they are available for the largest lakes. However, this implies that some bias may be unavoidable because the initial datasets have been not selected randomly.

The datasets used in the application were formulated as follows. We first sorted the lake data according to the area of the lake from the largest to the smallest and separated the $\max(n_0) = 200$ largest lakes. Three initial datasets containing $n_0 = 25$, $n_0 = 50$ and $n_0 = 200$ largest lakes were extracted from this subset. The three choices considered for n_0 represent situations where a small amount of the prior data ($n_0 = 25$), a moderate amount of the prior data ($n_0 = 50$) and a large amount of the prior data ($n_0 = 200$) are available. Next, we used the remaining dataset of size $n = 4360 - 200 = 4160$ for the subsample selection. Figure 2 summarizes the study setting.

In the logistic regression model, the binary response was the status of the lake (the target status being 1 and the nontarget status being 0) and the waterbed area and the agricultural land in the municipality where the lake is located were covariates. We aimed to keep the model simple since the optimization is challenging in a high-dimensional model where the size of the information matrix is large (García-Ródenas et al. 2020). The covariates were centred and standardized by dividing by the standard deviation.

As a preparation for the subsample selection, we first fitted a Bayesian logistic regression model to the initial data of size n_0 using a noninformative prior in model fitting implemented in the `brms`-package (Bürkner 2017). Student's t-distribution was used as a prior for the intercept, and uniform distributions were used for regression coefficients. The model fitted

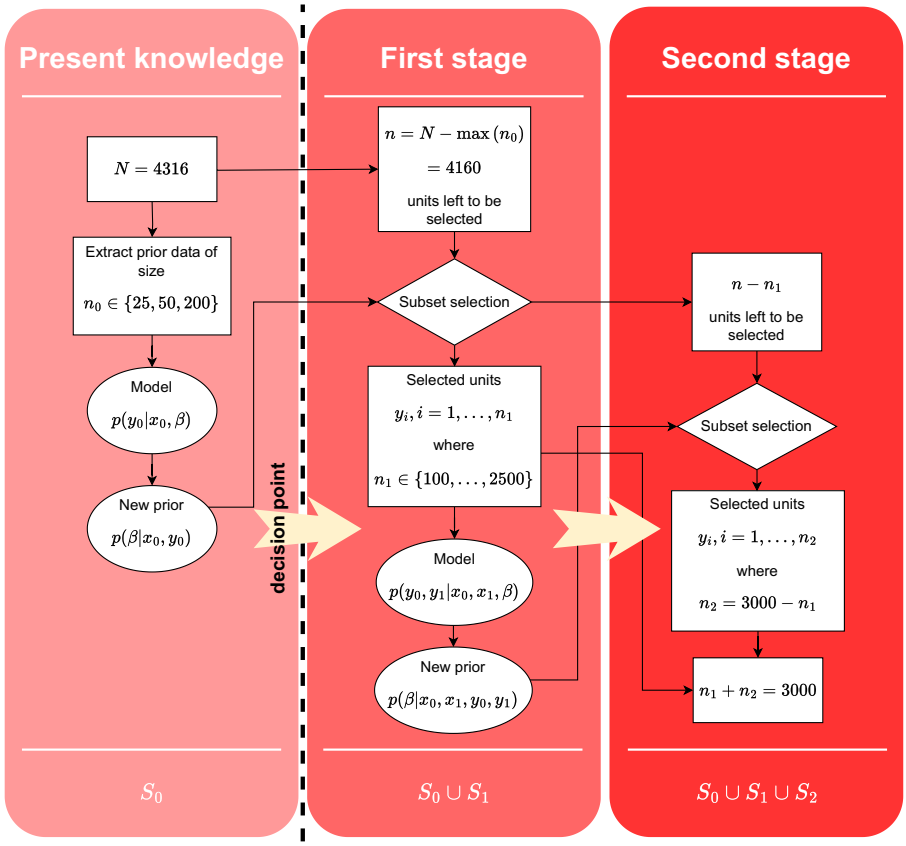


Figure 2. Flow chart of the two-staged study setting using the notations and sample sizes of the Finnish lake data application. The decision-maker is assumed to have all the knowledge in "Present knowledge," while the first stage and the second stage are to be planned .

to the initial data represents the knowledge that is available when the subsample selection starts and the decision on the measurement plan for the first stage is to be made.

In the subsample selection, we implemented the first stage selection and selected n_1 additional units using Algorithm 1. Here, we use the model fitted to the initial data as the prior and apply an MCMC algorithm to draw posterior samples to estimate the model parameters. The first stage size was varied and took values $n_1 \in \{100, 700, 1300, 1900, 2500\}$. For the selected n_1 lakes, the ecological status was "measured", i.e. obtained from the lake data. We added the selected subset to the initial data to obtain $n_0 + n_1$ units in total and used this to form an informative prior on the second stage.

On the second stage, we selected n_2 units using the same procedure as at the first stage but with the informative prior estimated from $n_0 + n_1$ units. We fixed the second stage design size n_2 so that at the second stage, the total number of selected lakes was $n_1 + n_2 = 3000$. This makes the selections with a different number of first stage design sizes comparable with each other. Thus, in total, the number of lakes to be measured was $n_0 + n_1 + n_2$. The D-criteria after this second stage selection are reported in the next section. We repeated the

Table 1. Parameter estimates and their standard deviations of a Bayesian logistic regression model fitted in the initial lake data of size $n_0 \in \{25, 50, 200\}$ and in the full lake data of size $N = 4360$

	$n_0 = 25$ $U = 7.649$		$n_0 = 50$ $U = 9.791$		$n_0 = 200$ $U = 12.570$		$N = 4360$ $U = 20.069$	
	$\hat{\beta}_j$	SD $\left(\hat{\beta}_j\right)$	$\hat{\beta}_j$	SD $\left(\hat{\beta}_j\right)$	$\hat{\beta}_j$	SD $\left(\hat{\beta}_j\right)$	$\hat{\beta}_j$	SD $\left(\hat{\beta}_j\right)$
Intercept	−0.002	0.849	1.094	0.478	1.559	0.227	1.854	0.049
Waterbed area	0.081	0.094	0.004	0.059	−0.028	0.048	−0.069	0.039
Agricultural land	−0.026	0.484	−0.137	0.388	−1.045	0.201	−0.994	0.042

optimal selection 100 times, generating new initial model parameter values at each time and studied the range of the quantities over these 100 repetitions.

For comparison, we also considered a case where $n_1 + n_2 = 500$ and the first stage size took values $n_1 \in \{10, 120, 230, 340, 450\}$. All other details were similar to the case with $n_1 + n_2 = 3000$.

5.2. RESULTS ON BAYESIAN TWO-STAGE SELECTION

Table 1 shows the model parameter estimates and the standard deviations when a small amount of the prior data ($n_0 = 25$), a moderate amount of the prior data ($n_0 = 50$) and a large amount of the prior data ($n_0 = 200$) are available. We use these estimates as informative priors in the subsample selection. Naturally, these estimates have larger standard deviations compared to the model fitted in the full data of size $N = 4360$ for which the D-criterion (Eq. 6) equals $U(\boldsymbol{\beta}, \boldsymbol{y}(\boldsymbol{r}), \boldsymbol{x}) = 20.07$. Not surprisingly, the estimates seem to be biased for $n_0 = 25$ and $n_0 = 50$ because the initial data contain only the largest lakes. With $n_0 = 200$, on the other hand, the model manages to estimate the effects of the covariates in the same way as for the full data $N = 4360$. The D-criteria (Eq. 6) of these initial models are, 7.649, 9.791 and 12.570, respectively.

Next, we implemented the two-stage selection. The top row in Fig. 3 shows the mean D-criterion over 100 repetitions after the second stage selection with varying first stage design sizes n_1 when the Bayesian two-stage approach is used with different sizes of prior data $n_0 \in \{25, 50, 200\}$. The total number of selected lakes is $n_0 + n_1 + n_2 = 3025$, $n_0 + n_1 + n_2 = 3050$ and $n_0 + n_1 + n_2 = 3200$, respectively. Comparing the results with the three different values of the amount of the prior data, naturally, the general level of D-criterion increases, when the total number of observations increases; see the definition of U in (6).

When $n_0 = 25$, the highest second stage D-criterion value is obtained when approximately 1900 lakes are selected to be measured at the first stage. If the first stage design size is higher than 1900, it seems that the second stage D-criterion will be lower. However, the increase in the D-criterion value starts to flatten out already when $n_1 = 700$. Thus, by selecting more than 700 lakes at the first stage does not clearly improve the final result given that $n_1 + n_2 = 3000$. The same pattern can be seen with $n_0 = 50$.

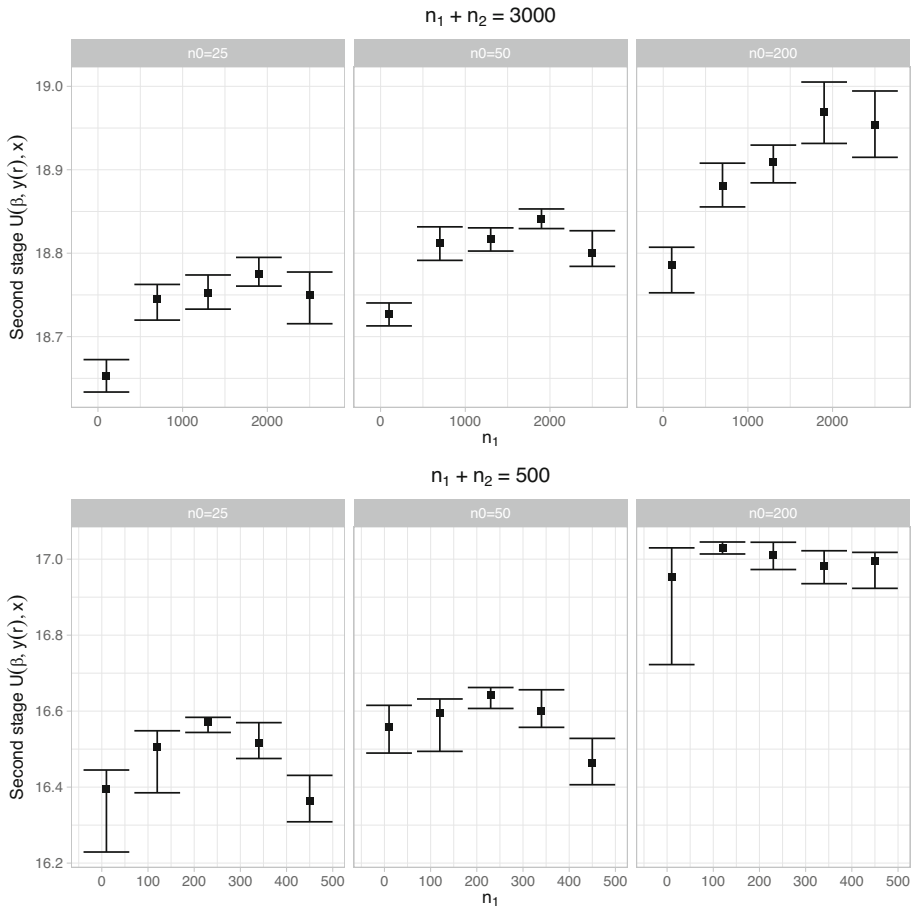


Figure 3. The second stage D-criterion as a function of the size of the first stage in Bayesian subsample selection. The squares and error bars represent the mean, maximum and minimum D-criteria of 100 independent selections. The total number of lakes selected at the first and second stage is $n_1 + n_2 = 3000$ (top row) or $n_1 + n_2 = 500$ (bottom row) .

When $n_0 = 200$, no flattening out after $n_1 = 700$ can be seen, but the second stage D-criterion increases strongly when the value of n_1 increases. Intuitively, a large enough amount of prior data makes the two-stage selection less beneficial. More prior data gathered at the first stage simply improve the final result. However, similarly to a small and a moderate amount of the prior data, the maximum can still be found at $n_1 = 1900$.

Based on Fig. 3 (top row), selecting $n_1 = 1900$ at the first stage and $n_2 = 1100$ is the best strategy for all the scenarios of the amount of the prior data. Table 2 shows the parameter estimates and their standard deviations for this strategy as means over 100 independent iterations. The mean D-criteria in these cases are 18.776, 18.841 and 18.969, respectively. In addition, Table 2 shows the model parameters and the standard deviations for the single-stage approach where the sizes are $n_1 = 3000$ and $n_2 = 0$. Assuming the three different prior data scenarios, the total data size used for model fitting is $n_0 + n_1 + n_2 \in \{3025, 3050, 3200\}$. The mean D-criteria in these situations are 18.737, 18.802 and 18.886, respectively. Comparing

Table 2. Parameter estimates and their standard deviations of a Bayesian logistic regression model fitted in the selected additional data

	$n_0+n_1+n_2=3025$ $U=18.776$		$n_0+n_1+n_2=3050$ $U=18.841$		$n_0+n_1+n_2=3200$ $U=18.969$	
	$\hat{\beta}_j$	SD $\left(\hat{\beta}_j\right)$	$\hat{\beta}_j$	SD $\left(\hat{\beta}_j\right)$	$\hat{\beta}_j$	SD $\left(\hat{\beta}_j\right)$
Intercept	1.871	0.065	1.874	0.065	1.820	0.061
Waterbed area	−0.065	0.027	−0.061	0.028	−0.066	0.026
Agricultural land	−0.985	0.044	−0.981	0.044	−0.968	0.043
	$n_0+n_1=3025$ $U=18.737$		$n_0+n_1=3050$ $U=18.802$		$n_0+n_1=3200$ $U=18.886$	
	$\hat{\beta}_j$	SD $\left(\hat{\beta}_j\right)$	$\hat{\beta}_j$	SD $\left(\hat{\beta}_j\right)$	$\hat{\beta}_j$	SD $\left(\hat{\beta}_j\right)$
Intercept	1.899	0.066	1.901	0.066	1.862	0.063
Waterbed area	−0.066	0.027	−0.063	0.028	−0.070	0.025
Agricultural land	−0.995	0.044	−0.992	0.044	−0.992	0.044

In the upper part of the table, the first stage size is $n_1 = 1900$ and the second stage size is $n_2 = 1100$, with the initial lake data of size $n_0 \in \{25, 50, 200\}$, the total data size used for model fitting being $n_0 + n_1 + n_2 \in \{3025, 3050, 3200\}$. In the lower part of the table, the selection is made with the single-stage approach. The values are means over 100 iterations

the two-stage approach to the single-stage approach, the two-stage approach works slightly better, but the differences are minor. Once the selection is made, the analysis could be extended and a model made using the other variables as covariates introduced in Sect. 2.

The top row of Fig. 3 shows the result when the total amount of selected lakes is quite large, $n_1 + n_2 = 3000$. We now present a situation where the total amount is smaller. The bottom row of Fig. 3 shows the results for the case where $n_1 + n_2 = 500$ and the first stage design size varies to be $n_1 \in \{10, 120, 230, 340, 450\}$. For $n_0 = 25$, $n_0 = 50$ and $n_0 = 200$, the total number of selected lakes are $n_0 + n_1 + n_2 = 525$, $n_0 + n_1 + n_2 = 550$ and $n_0 + n_1 + n_2 = 700$, respectively. When $n_0 = 25$ and $n_0 = 50$, the highest second stage D-criterion value is obtained when approximately 230 lakes are selected to be measured at the first stage. If the first stage design size is higher than 230, the D-criterion value decreases fast. This suggests that the model should be updated early enough. Otherwise, the lakes selected based on insufficient prior information at the first stage constitute a too large a proportion of the all lakes selected. With $n_0 = 200$, the pattern seems different: the D-criterion values increase when n_1 increases until approximately 120 lakes are selected at the first stage, and then decreases moderately. It seems that the prior information is sufficient to estimate the model, and thus, no major differences are observed between the first stage design sizes. When $n_1 = 230$ and $n_2 = 270$, the mean D-criteria for $n_0 = 25$, $n_0 = 50$ and $n_0 = 200$ are 16.571, 16.643 and 17.011, respectively. The corresponding mean D-criteria values for one-stage sampling are 16.239, 16.343 and 16.898. The benefits of two-stage sampling seem to be slightly larger when $n_1 + n_2 = 500$ than when $n_1 + n_2 = 3000$.

6. DISCUSSION

We have considered the problem of subsample selection in the context of lake management. The aim was to select lakes that are maximally informative in the sense of Bayesian D-optimality for the prediction of water quality status. We compared two-stage selection to single-stage selection and found that two-stage selection may have only a modest advantage.

Subsample selection has similarities with experimental design. In both, the aim is to maximize D-optimality or another function of information matrix. The main differences are that in subsample selection the set of units is fixed and each unit (lake) can be chosen only once. Algorithms developed for experimental design are often useful also in subsample selection but require modifications. Our choice of using a Bayesian approach together with the greedy forward selection takes into account the uncertainty in the model parameters and keeps the computational time reasonable.

Sequential or multi-stage approaches are never theoretically worse than single-stage approaches. In our application of lake management, the sequential strategy is not feasible because of the time needed to analyse water samples. The two-stage approach is a compromise that is possible to implement. Figure 3 shows that the D-criterion obtained in two-stage selection depends on the size of the first stage subsample in a nontrivial way. In practice, however, the benefit of the two-stage selection remained modest in the lake management data, as shown in Table 2.

The prior distributions of the model parameters were estimated from the initial data on the lakes with the largest surface area. It is realistic to assume that the initial data are collected for lake management purposes and therefore the lakes are not selected randomly. Using only the largest lakes obviously causes bias in the priors, which leads to suboptimal selection and may cause some bias even in the final estimates, especially if the size of the selected data is small.

The lake data had the status classification measured for 4360 lakes, but in the analysis we simulated the situation where the status is yet to be determined. In reality, there are actually 58,707 lakes that can be found from the database maintained by the Finnish Environment Institute. Since the status classification is already available for 4360 lakes of those 58,707 lakes with basic characteristics available, the classification is still missing for 54,347 lakes. The current work can be utilized in the planning of the data collection strategy for these lakes.

The presented methods may be applicable also in other environmental monitoring problems where the quantity of interest is expensive or difficult to measure. In addition to D-optimality, it is possible to consider value of information (Eidsvik et al. 2015; Koski et al. 2020; Koski and Eidsvik 2024) and other criteria that link directly to decision making.

ACKNOWLEDGEMENTS

Corresponding author acknowledges the support by the Emil Aaltonen Foundation and Kone foundation. CSC–IT Center for Science, Finland, is acknowledged for computational resources.

Funding Open Access funding provided by University of Jyväskylä (JYU).

Declarations
Conflict of interest The authors have no conflict of interest to declare.

[Received September 2022. Revised April 2024. Accepted May 2024.]

REFERENCES

Aroviita J, Mitikka S, Vienonen S (2019) Pintavesien tilan luokittelu ja arviointiperusteet vesienhoidon kolmannella kaudella. Finnish Environment Institute (SYKE), Helsinki

Atkinson A, Donev A, Tobias R (2007) Optimum experimental designs, with SAS. Oxford University Press, Oxford

Bürkner P-C (2017) brms: an R package for Bayesian multilevel models using Stan. J Stat Softw 80(1):1–28. https://doi.org/10.18637/jss.v080.i01

Chaloner K, Verdinelli I (1995) Bayesian experimental design: a review. Stat Sci 10(3):273–304. https://doi.org/10.1214/ss/1177009939

Dykstra O (1971) The augmentation of experimental data to maximize [X'X]. Technometrics 13(3):682–688. https://doi.org/10.1080/00401706.1971.10488830

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

European Communities (2003) Common strategy on the implementation of the water framework directive (2000/60), guidance document no. 13, overall approach to the classification of ecological status and ecological potential

European Parliament (2000) Directive 2000/60/EC, of the European parliament and council of 23 October 2000 establishing a framework for community action in the field of water policy. http://eur-lex.europa.eu/resource.html?uri=cellar:5c835afb-2ec6-4577-bdf8-756d3d694eeb.0004.02/DOC_1&format=PDF

Fedorov V (1989) Optimal design with bounded density: optimization algorithms of the exchange type. J Stat Plan Inference 22(1):1–13. https://doi.org/10.1016/0378-3758(89)90060-8

Fedorov VV (1972) Theory of optimal experiments. Academic Press, New York

García-Ródenas R, García-García JC, López-Fidalgo J, Martín-Baos JÁ, Wong WK (2020) A comparison of general-purpose optimization algorithms for finding optimal approximate experimental designs. Comput Stat Data Anal 144:106844. https://doi.org/10.1016/j.csda.2019.106844

Guillera-Arroita G, Ridout M, Morgan B (2014) Two-stage Bayesian study design for species occupancy estimation. JABES 19:278–291. https://doi.org/10.1007/s13253-014-0171-4

Heiskanen A-S, Hellsten S, Vehviläinen B, Putkuri E (2017) How well is water protected in the land of a thousand lakes. Finnish Environment Institute, Helsinki, Finland

Hoffman MD, Gelman A (2014) The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. J Mach Learn Res 15(1):1593–1623

Karvanen J (2009) Approximate cost-efficient sequential designs for binary response models with application to switching measurements. Comput Stat Data Anal 53(4):1167–1176

Koski V, Eidsvik J (2024) Sampling design methods for making improved lake management decisions. Environmetrics. https://doi.org/10.1002/env.2842

Koski V, Kotamäki N, Hämäläinen H, Meissner K, Karvanen J, Kärkkäinen S (2020) The value of perfect and imperfect information in lake monitoring and management. Sci Total Environ 726:138396. https://doi.org/10.1016/j.scitotenv.2020.138396

Meyer RK, Nachtsheim CJ (1995) The coordinate-exchange algorithm for constructing exact optimal experimental designs. Technometrics 37(1):60–69

Montepiedra G, Yeh AB (1998) A two-stage strategy for the construction of D-optimal experimental designs. Commun Stat Simul Comput 27(2):377–401

Official Statistics of Finland (2020) Utilised agricultural area [e-publication]. Natural Resources Institute Finland, Helsinki

Official Statistics of Finland (2020) Buildings and free-time residences [e-publication]. Statistics Finland, Helsinki

Paglia J, Eidsvik J, Karvanen J (2022) Efficient spatial designs using Hausdorff distances and Bayesian optimization. Scand J Stat 49(3):1060–1084. https://doi.org/10.1111/sjos.12554

Pronzato L (2006) On the sequential construction of optimum bounded designs. J Stat Plan Inference 136(8):2783–2804. https://doi.org/10.1016/j.jspi.2004.10.020

Pronzato L, Pázman A (2013) Design of experiments in nonlinear models, vol 212. Lecture notes in statistics. Springer, Cham

Pronzato L, Wang H (2021) Sequential online subsampling for thinning experimental designs. J Stat Plan Inference 212:169–193

R Core Team (2023) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Reilly M (1996) Optimal sampling strategies for two-stage studies. Am J Epidemiol 143(1):92–100

Reinikainen J, Karvanen J (2022) Bayesian subcohort selection for longitudinal covariate measurements in follow-up studies. Stat Neerl 76(4):372–390. https://doi.org/10.1111/stan.12264

Reinikainen J, Karvanen J, Tolonen H (2016) Optimal selection of individuals for repeated covariate measurements in follow-up studies. Stat Methods Med Res 25(6):2420–2433. https://doi.org/10.1177/0962280214523952

Ruggoo A, Vandebroek M (2004) Bayesian sequential D-D optimal model-robust designs. Comput Stat Data Anal 47(4):655–673

Ryan EG, Drovandi CC, McGree JM, Pettitt AN (2016) A review of modern computational algorithms for Bayesian optimal design. Int Stat Rev 84(1):128–154. https://doi.org/10.1111/insr.12107

Sitter RR, Forbes B (1997) Optimal two-stage designs for binary response experiments. Stat Sin 7(4):941–955

Stan Development Team (2022) Stan modeling language users guide and reference manual version 2.18.0. http://mc-stan.org/

Welch W (1982) Algorithmic complexity: three NP-hard problems in computational statistics. J Stat Comput Simul 15:17–25. https://doi.org/10.1080/00949658208810560

Wynn H (1982) Optimum submeasures with application to finite population sampling. In: Gupta SS, Berger JO (eds) Statistical decision theory and related topics III. Academic Press, New York, pp 485–495

Zuo L, Zhang H, Wang H, Sun L (2021) Optimal subsample selection for massive logistic regression with distributed data. Comput Stat 36:1–28. https://doi.org/10.1007/s00180-021-01089-0

III

SAMPLING DESIGN METHODS FOR MAKING IMPROVED LAKE MANAGEMENT DECISIONS

by

Koski, V., & Eidsvik, J. 2024

Environmetrics, e2842. DOI: https://doi.org/10.1002/env.2842

Received: 20 March 2023 | Revised: 9 November 2023 | Accepted: 18 January 2024

DOI: 10.1002/env.2842

RESEARCH ARTICLE

WILEY

Sampling design methods for making improved lake management decisions

Vilja Koski[1] | **Jo Eidsvik**[2]

[1]Department of Mathematics and Statistics, University of Jyväskylä, Jyväskylä, Finland

[2]Department of Mathematical Sciences, NTNU, Trondheim, Norway

Correspondence
Vilja Koski, Department of Mathematics and Statistics, University of Jyväskylä, P.O Box 35, Jyväskylä 40014, Finland.
Email: vilja.a.koski@jyu.fi

Funding information
Koneen Säätiö; Norges Forskningsråd, Grant/Award Numbers: 305445, 309960; Emil Aaltosen Säätiö

Abstract

The ecological status of lakes is important for understanding an ecosystem's biodiversity as well as for service water quality and policies related to land use and agricultural run-off. If the status is weak, then decisions about management alternatives need to be made. We assess the value of information of lake monitoring in Finland, where lakes are abundant. With reasonable ecological values and restoration costs, the value of information analysis can be compared with the survey's costs. Data are worth gathering if the expected value from the data exceeds the costs. From existing data, we specify a hierarchical Bayesian spatial logistic regression model for the ecological status of lakes. We then rely on functional approximations and Laplace approximations to get closed-form expressions for the value of information of a sampling design. The case study contains thousands of lakes. The combinatorially difficult design problem is to wisely pick the right subset of lakes for data gathering. To solve this optimization problem, we study the performance of various heuristics: greedy forward algorithms, exchange algorithms and Bayesian optimization approaches. The value of information increases quickly when adding lakes to a small design but then flattens out. Good designs are usually composed of lakes that are difficult to manage, while also balancing a variety of covariates and geographic coverage. The designs achieved by forward selection are reasonably good, but we can outperform them with the more nuanced search algorithms. Statistical designs clearly outperform other designs selected according to simpler criteria.

KEYWORDS

data collection, decision-making, environmental monitoring, optimal design, value of information

1 | INTRODUCTION

We consider a survey design problem connected to environmental monitoring. The inspiration for this study comes from the real-life challenge of lake monitoring in Finland, where lakes are abundant. Inland waters and freshwater biodiversity constitute a valuable natural resource in economic, cultural, aesthetic, scientific and educational terms and need to be protected (Dudgeon et al., 2006). As a result of the Water Framework Directive (WFD) of the European Union (European Parliament, 2000), Finland has implemented a water monitoring program for improving and securing the

quality of its inland waters. In the current program, lakes are classified into five ecological status classes (high, good, moderate, poor and bad) according to several variables representing biotic structure, supported by the physical and chemical properties of water, and hydrological as well as morphological features. Existing data on these variables are used to determine the reference conditions for each status class. In addition, according to the directive, some management alternatives must be implemented to improve the ecological status if the status of the water system is classified as moderate or lower. Though biologically principled, the current monitoring program has been considered to be very expensive, and the question is if the efforts are worth it. How should decision-makers wisely allocate monitoring resources at a subset of lakes to significantly aid the decisions about the management alternatives?

A critical question is then to find the optimal sampling design under some information criterion. Regarding the ecological status of Finnish lakes, there are relatively clear management alternatives and rather specific monetary values associated with the various alternatives. Hence, it makes sense to phrase the design criterion according to the notion of decision theory (Abbas & Howard, 2015). In particular, we value information that can improve management decisions via the expected posterior value (PoV) as compared with the prior value (PV) using only the currently available knowledge (Eidsvik et al., 2015). An integral part of this criterion is reduced uncertainty in the statistical model for lake status because it enters into the expected values used in the decision rule. The goal is to find the sampling design which gives the largest value of information (VOI) compared with the cost of data acquisition and processing.

In this paper, the VOI is calculated assuming a Bayesian spatial logistic regression model for the ecological status data. Statistical model parameters are specified from existing data gathered in Finnish lakes. Our large-scale VOI calculations rely on closed-form approximations for hierarchical general linear models (Evangelou & Eidsvik, 2017), which enable fast evaluation of the VOI for each design.

Generally, the problem of selecting an optimal design under some criterion is a central research question in the planning of survey data. However, there are several thousand lakes in Finland, and to find a truly optimal design one would have to evaluate all the available designs. This becomes a combinatorial challenge which is infeasible for our case. One can only evaluate a subset of the designs and we need heuristic algorithms to search for promising subsets. A straightforward heuristic which is easy to implement is the greedy method. It is well-known by mathematicians and computer scientists, and in statistics it is often referred to as a sequential search method (Dykstra, 1971). More nuanced heuristics can naturally build on the result obtained by this approach.

Fortunately, due to the traditional role of statistics in environmental planning, there already exists a significant amount of literature on effective data designs. Other design studies include Jauslin et al. (2022), who consider sequential balanced designs with inclusion probabilities and illustrate this on a data set of species of amphibians; Prentius and Grafström (2022), who compare efficiencies of two-phase methods for adaptive cluster sampling in environmental settings; Foss et al. (2022), who construct dynamic monitoring designs for characterizing the concentration of mine tailings using a spatio-temporal model; and Thilan et al. (2023), who propose adaptive spatio-temporal designs for evaluating trends in coral cover. Nguyen et al. (2018) provide a review of adaptive sampling designs in environmental monitoring. Recent studies concerning the evaluation of information in ecology include the VOI tutorial by Canessa et al. (2015) and its applications in species management, and Williams and Brown (2020), who use scenarios to split settings of pre-selected designs and alternatives that adapt to information. Reich et al. (2018) suggest minimizing the expected misclassification rate of occupancy maps in an ecological application with citizen science data. Our study is different in how we approach the spatial decision situations and in the methods connecting this to a logistic regression model with spatially correlated latent variables.

Several statistical researchers have focused on common optimality criteria of experimental spatial designs, such as D- and A-optimality. Woods et al. (2017) present several approaches for Bayesian design of experiments in logistic regression models with non-spatial applications. Hays et al. (2021) propose a method that links linear integer programming to optimality measures of covariance matrices resulting from mixed models, and as in this work, results are presented on data from freshwater sites. Integer programming has been a popular method to solve the subset selection problem (see, e.g., Arthur et al., 1997). More similar to our work, Paglia et al. (2022) study the VOI computation tasks and propose a Bayesian optimization technique to find approximately optimal spatial designs. We test this method for our case which is of much larger size and involves a different model concerning the hierarchical logistic regression model.

The article is organized as follows: Section 2 provides the background for the case on lake monitoring and the associated sampling design problems, along with a suggested workflow. Section 3 presents the decision situation and the Bayesian spatial logistic regression model for lake status variability, as well as the computational approaches for conducting VOI analysis and heuristic search algorithms used to find good designs. Section 4 explores the results of implementing

the algorithms on the lake example from Finland along with sensitivity analysis. Section 5 contains interpretations of the results. Section 6 concludes and presents future work.

2 | BACKGROUND

2.1 | Monitoring the ecological status of lakes

The aim of the WFD is to prevent the deterioration of the ecological status of water systems, with the aim of having at least good ecological and chemical status class. In order to put the legislation into practice in Finland (Figure 1), River Basin Management Planning (RBMP) is implemented in six-years cycles (Aroviita et al., 2019). In brief, the essential parts of one RBMP period are (Higgins et al., 2021; Stankey et al., 2005)

1. the monitoring of the water systems,
2. the assessment and classification of the water systems into status classes,
3. the planning of management alternatives based on the classification, and
4. the implementation of the alternatives.

In the first step, monitoring includes data acquisition of several parameters indicative of the water quality. Biological factors such as phytoplankton, chlorophyll-*a* content in algae, benthic fauna and aquatic plants are monitored at observation sites every 1 to 6 years, depending on the factors. Physical and chemical parameters, such as temperature, phosphorus, nitrogen and oxygen content are gathered from water samples at the lakes at regular intervals, either annually or every few years. Within a year, samples are usually taken about 2 to 12 times. Here, we are mainly interested in the biological quality of chlorophyll-*a* samples, which are collected from the observation sites in summertime (approximately from May to August). Chlorophyll-*a* content is indicative of water body productivity and therefore generally correlates well with the ecological status of lentic water bodies suffering from human-induced eutrophication.

In the second step, one defines the status class of each lake. The conditions are assessed on the basis of the intensity of the ecological changes caused by human activity (Nõges et al., 2009). Thus, the classification is based on several indicators

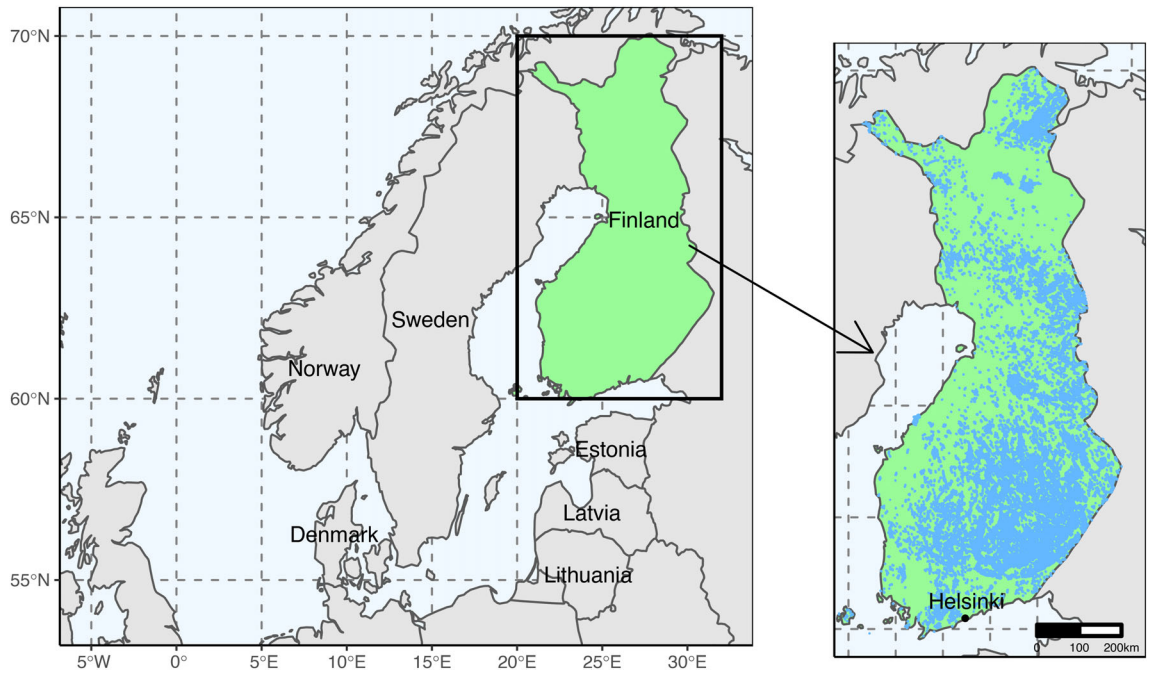


FIGURE 1 Lakes are abundant in Finland. The map shows the lakes which are defined as water bodies in Finland according to the Water Framework Directive (European Parliament, 2000). For the monitoring of ecological status, a subset of lakes must be chosen for sampling.

mentioned above. The classification provides information on the water systems that need measures to achieve or maintain good status.

In the third step, the decisions about the restoration and management alternatives are made based on the classification. These alternatives vary depending on the problems a lake might have. For instance, if the problem is eutrophication, then the management alternative should start with preventing nutrient discharge to the water system. As this is not always possible, the next steps may be dredging to decrease the amounts of aquatic plants and fishing (Søndergaard et al., 2007). Other attempts of lake restoration include raising the water level of the lake or biomanipulation (Jeppesen et al., 2017).

The fourth step is implementing the management alternatives. After the restoration, the effect of the alternatives must again be evaluated via monitoring, and this produces the new status assessment, which returns to the first step.

The basic unit in water management is a water body. It is a separate and significant part of surface water, such as a lake, a creek or a river. In this study, we are only interested in the status of lakes. A lake may form a single water body or it may be divided into several water bodies if it is justified from an ecological point of view. Each lake has at least one sampling site, but the largest lakes might have several sites since they have various habitats and thus several water bodies. Currently, not all smaller lakes (area less than 1 km^2) are defined as water bodies, and they are hence not included. However, smaller water bodies may also be included in the classification at a later stage if they are considered to be significant.

The classification of waters has been conducted three times in Finland. In this study, we use the third ecological status classification, and it is based on the monitoring data gathered during the third RBMP period from 2012 to 2017. The classification is available via open source data maintained by the Finnish Environment Institute (http://www.syke.fi/en-US/Open_information). Since the demand of management alternatives is our main interest, we have narrowed our inspection to the binary ecological classification of lakes, based on whether a lake needs management alternatives (bad, poor or moderate) or not (good or high).

The aim is to predict the ecological status of lakes, and then to use these predictions to make decisions about management alternatives. For the purposes of prediction, we use publicly available information on Finnish lakes. The basic features of 58,707 lakes in Finland can be found from the database maintained by the Finnish Environment Institute (https://www.syke.fi/en-US/Open_information/Open_web_services/Environmental_data_API). Each lake has characteristics such as location (the municipality, drainage basin, center latitude and longitude coordinates, altitude), waterbed area (hectares), length of shoreline (kilometers), average and maximum depth (meters) and volume of water mass (1000 cubic meters). There are also covariates for the agricultural area of municipalities where each lake is located (Official Statistics of Finland, 2020b) and the number of free-time residences in the municipality where each lake is located (Official Statistics of Finland, 2020a). To remove the effect of the municipality area, we divided the agricultural area of municipalities by the area of the municipality to obtain the percentage of agricultural area in each municipality (Figure 2, left).

Status classification is already available for 4360 of the 58,707 lakes. For the remaining lakes, we can predict the status class using the model trained on data from lakes with both status and covariates. Using a logistic regression model, we get the probabilities of ecological status displayed in Figure 2 (right). Here, the most important covariate appears to be agricultural land (left display). We point out that lakes that have a very high probability (near 1) of being in the ecological target status need no remediation. Further, lakes that have a very low probability (near 0) of being in the target status, clearly need to be addressed. Lakes in these two groups are hence not important or worthwhile to monitor because one already knows what to decide. However, there are plenty of lakes for which it is very difficult to make a management decision, and for these it can be very valuable to get information about the status class. But this kind of information comes with a cost, and the dilemma is which lakes should be sampled to make informed decisions on management alternatives for all the lakes.

2.2 | Workflow

In this section, we present the steps that sum up the process of sampling design selection. After framing the decision situation as described in Section 3.1, the following steps are performed:

1. Model fitting from existing data: Construct a statistical model for ecological status based on the 4360 lakes having both status classification and covariates (see Section 4.1).
2. Limit the scope to relevant lakes: Identify lakes with large uncertainty about ecological status classification that could be important to sample (see Section 4.1).

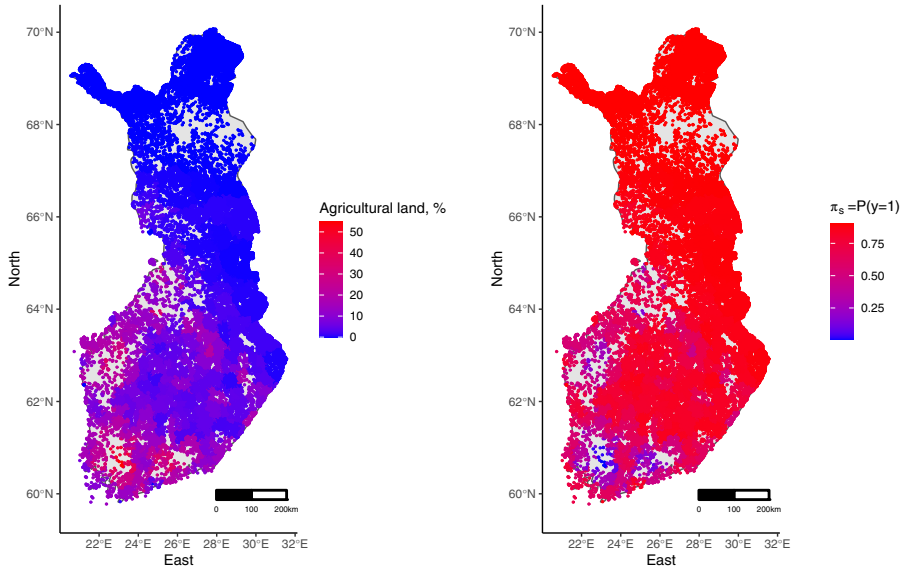


FIGURE 2 There are 54,347 lakes with all covariates available but with missing ecological status. Left: The lakes are color-coded according to the amount of agricultural land (in percentage) in the municipality where the lake is located. Right: The probability of a lake being in the target ecological status ($y_s = 1$), that is, in high or good status, based on a logistic regression model.

3. Sequential selection: Use a greedy forward selection algorithm to find sample designs with large VOI (see Algorithm 1 in Section 3.3.2).
4. Heuristic design search: Conduct nuanced exchange algorithms or Bayesian optimization to search for designs with larger VOI (see Section 3.3.2).

High-quality designs are characterized by large VOI. The results are compared with the cost of gathering data according to the respective sampling designs. In the search algorithms the goal is to optimize the VOI, but one could of course also be motivated by other criteria when selecting designs. We compare and discuss the value of other designs based on various criteria in Section 5.

3 | STATISTICAL FRAMEWORK

3.1 | Framing the decision and sampling problems

The notation connected to the sampling design problem is presented in Table 1. One can choose to leave a lake s untreated ($a_s = 0$) or act to bring the lake to a satisfying condition ($a_s = 1$). We adopt the monetary units associated with these alternatives from Koski et al. (2020). The value of a lake in good condition is set to $R =$ EUR 1000 per hectare, while a lake in poor condition is valued at EUR 0 per hectare. Under the management alternative, it costs EUR 200 per hectare to bring the lake to a sufficiently good condition. No matter the final condition of the lake, the resulting value is $C =$ EUR 1000 − EUR 200 = EUR 800 per hectare. Because of the uncertainty in determining the ecological condition of a lake, it is difficult for managers to make decisions about lake management. For a risk neutral decision maker, the PV is the maximum expected value over the two management alternatives. For a particular lake s with area A_s hectare, this can be written as

$$\mathrm{PV}_s = \max\{R_s \cdot \mathbb{E}(\pi_s),\ C_s\} = C_s + \max\{R_s \cdot \mathbb{E}(\pi_s) - C_s,\ 0\},$$

where for alternative $a_s = 1$ the monetary amount is $C_s = CA_s$, while for alternative $a_s = 0$ there is revenue $R_s = RA_s$ and $\mathbb{E}(\pi_s)$ denotes the expected condition of lake $s = 1, \ldots, N$. We will later model this via a latent logistic regression model for variable x_s, where $\pi_s = \frac{e^{x_s}}{1+e^{x_s}}$.

TABLE 1 Summary of the notation used in the article.

Notation	Definition
$s \in \{1, \dots, N\}$	Index for lakes
$\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_N\}$	Locations of all lakes
$D \in \mathcal{D}$	Design in all possible design sets
π_s	Probability for lake s not needing management alternatives
x_s	Latent random variable at lake s
$\boldsymbol{f}_s$	Covariates at lake s
$\boldsymbol{x}_D$	Latent length-$\lvert D\rvert$ random vector of design D
$\boldsymbol{y}_D$	Prospective data vector of length-$\lvert D\rvert$, gathered in design D
$\boldsymbol{F}_D$	Covariate matrix of design D
$a_s \in \{0, 1\}$	Management alternative to choose for lake s
A_s	Area of lake s
R_s	Revenue of alternative for lake s
C_s	Cost of alternative for lake s
PV	Prior value
PoV(D)	Posterior value of design D
VOI(D)	Value of information of design D
VOI$_s$(D)	Lake s effect value of information of design D
$P(D)$	Price or cost of data of design D

We assume that managers are free to select the best alternative for every lake (Eidsvik et al., 2015). This means that the total PV decouples to a sum over all lakes and we have

$$\mathrm{PV} = \sum_{s=1}^{N} \mathrm{PV}_s = \sum_{s=1}^{N} \left[C_s + \max\{R_s \cdot \mathbb{E}(\pi_s) - C_s,\ 0\} \right]. \tag{1}$$

Additional data can assist the decision-makers in choosing among the difficult lake management alternatives. In particular, the VOI is positive when various data outcomes lead to different alternatives being chosen, because this gives added value from that of the PV. Still, this gain must be compared with the cost of collecting and processing the data. Moreover, there are several possibilities for the design of gathering spatial data used in determining the ecological status of lakes.

Assume that one wants to select a subset of lakes to observe their ecological status. We denote such a subset by design D of size $|D|$. Collecting data for all lakes would be too expensive, and one can only afford to measure a subset. We denote the (latitude, longitude) positions of the N lakes of interest by $\boldsymbol{u}_1, \dots, \boldsymbol{u}_N$. Possible spatial survey designs contain no sites, single sites, couples, triplets, and so on, up to the design where all N sites are included, and the entire set of designs is denoted $\mathcal{D} = \bigcup_{i=0}^{N} \mathcal{D}_i$, where

$$\begin{aligned}
\mathcal{D}_0 &= \emptyset, \\
\mathcal{D}_1 &= \{(\boldsymbol{u}_1), (\boldsymbol{u}_2), \dots, (\boldsymbol{u}_N)\}, \\
\mathcal{D}_2 &= \{(\boldsymbol{u}_1, \boldsymbol{u}_2), (\boldsymbol{u}_1, \boldsymbol{u}_3), \dots, (\boldsymbol{u}_{N-1}, \boldsymbol{u}_N)\}, \\
&\vdots \\
\mathcal{D}_N &= \{(\boldsymbol{u}_1, \boldsymbol{u}_2, \dots, \boldsymbol{u}_N)\}.
\end{aligned} \tag{2}$$

Finding the optimal design is extremely difficult because there are 2^N possible designs. One often resorts to heuristics approaches to find several useful designs that provide a basis for decision support about information gathering.

Denote the prospective data to be measured in a design D by $\boldsymbol{y}_D = (y_{D,1}, \dots, y_{D,|D|})$. The associated covariates are denoted by matrix $\boldsymbol{F}_D$, which has one row for each design location. For this decision situation, the PoV of data $\boldsymbol{y}_D$ is defined by

$$\text{PoV}(D) = \sum_{\boldsymbol{y}_D} \sum_{s=1}^{N} \left[C_s + \max\{R_s \cdot \mathbb{E}(\pi_s|\boldsymbol{y}_D) - C_s,\ 0\}\right] p(\boldsymbol{y}_D), \tag{3}$$

where $\mathbb{E}(\pi_s|\boldsymbol{y}_D)$ is the conditional expected lake status, given observations $\boldsymbol{y}_D$ distributed according to the probability mass function $p(\boldsymbol{y}_D)$. Here, the sums over the data outcomes and lakes can be interchanged, so that $\text{PoV}(D) = \sum_{s=1}^{N} \text{PoV}_s(D)$, where the entries in the sum are the expected value contribution from the data $\boldsymbol{y}_D$ to the decision at lake s.

Under the assumptions of a risk neutral decision-maker (Eidsvik et al., 2015), the VOI equals the difference in PoV and PV, so that

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

Note that the fixed C_s part is the same for both Equations (1) and (3). Further, the decoupling over lake decisions means that the VOI is the additive contributions from the VOI at each lake s, that is,

$$\begin{aligned} \text{VOI}_s(D) &= \sum_{\boldsymbol{y}_D} \max\{R_s \cdot \mathbb{E}(\pi_s|\boldsymbol{y}_D) - C_s,\ 0\} p(\boldsymbol{y}_D) - \max\{R_s \cdot \mathbb{E}(\pi_s) - C_s,\ 0\}, \\ \text{VOI}(D) &= \sum_{s=1}^{N} \text{VOI}_s(D). \end{aligned} \tag{5}$$

The goal is to choose a sampling design D that is expected to provide data that substantially affect the decisions made, especially at those lakes where it is difficult to choose between the management alternatives. It is common to choose the design with the largest VOI(D) compared with the data gathering cost $P(D)$, as managers are interested in making the best out of their data acquisition and processing expenses. Alternatively, one can also have a budget for the data gathering, and the goal is to find the largest VOI among all designs D having costs not exceeding this budget. Overall, the VOI results for various designs will support difficult decisions related to data gathering.

3.2 | Binary regression

3.2.1 | Logistic model

Assume that a binary response $y_s \in \{0, 1\}$ at lake $s = 1, \dots, N$ is distributed as

$$\begin{aligned} P(y_s = 1|x_s) = \pi_s, \quad P(y_s = 0|x_s) = 1 - \pi_s, \\ \text{logit}(\pi_s) = x_s = \boldsymbol{f}_s'\boldsymbol{\beta} + w_s, \quad \pi_s = \frac{e^{x_s}}{1 + e^{x_s}}, \end{aligned} \tag{6}$$

where the linear predictor x_s at lake s includes covariates $\boldsymbol{f}_s = (f_1(s), \dots, f_J(s))'$ in combination with regression parameters $\boldsymbol{\beta} = (\beta_1, \dots, \beta_J)'$. It further has a lake-specific effect w_s that is spatially correlated. For short, we denote effects $\boldsymbol{w} = (w_1, \dots, w_N)'$.

Assuming conditional independence in Equation (6), the log-likelihood of data $\boldsymbol{y}_D = (y_{D,1}, \dots, y_{D,|D|})$ is obtained by

$$\sum_{\boldsymbol{u}_s \in D} \log(p(y_s|\boldsymbol{\beta}, \boldsymbol{w})) = \sum_{\boldsymbol{u}_s \in D} y_s(\boldsymbol{f}_s'\boldsymbol{\beta} + w_s) - \log(1 + \exp(\boldsymbol{f}_s'\boldsymbol{\beta} + w_s)). \tag{7}$$

3.2.2 | Bayesian latent spatial logistic model

The regression parameter $\boldsymbol{\beta}$ is unknown and has a prior probability density function (pdf) $p(\boldsymbol{\beta})$. This pdf is here assumed to be Gaussian with mean vector $\boldsymbol{\mu}_\beta^0$ and covariance matrix $\boldsymbol{\Sigma}_\beta^0$. The spatial effects $\boldsymbol{w}$ are represented by a zero mean

Gaussian process model. We specify a fixed variance $\text{Var}(w_s) = \sigma^2$ and impose a Matern correlation function such that $\text{Corr}(w_s, w_t) = (1 + \phi h_{st}) \exp(-\phi h_{st})$, where h_{st} is the great-circle distance between two lakes centered at locations $\boldsymbol{u}_s$ and $\boldsymbol{u}_t$. Given the currently available lake data, we specify model parameters $\boldsymbol{\mu}_\beta^0$, $\boldsymbol{\Sigma}_\beta^0$, σ and ϕ from an approximate marginal likelihood expression.

Keeping the model parameters fixed in the following, $(\boldsymbol{\beta}', \boldsymbol{w}')'$ are Gaussian distributed. In particular, the linear predictor $x_s = \boldsymbol{f}_s'\boldsymbol{\beta} + w_s$ has mean $\mu_s = \boldsymbol{f}_s'\boldsymbol{\mu}_\beta^0$ and variance $\sigma_s^2 = \boldsymbol{f}_s'\boldsymbol{\Sigma}_\beta^0\boldsymbol{f}_s + \sigma^2$. Let $\boldsymbol{x}_D = \{x_s; \boldsymbol{u}_s \in D\}$ denote the vector of linear predictor variables at the design locations defined via set D. We similarly define the vector $\boldsymbol{w}_D$ of latent effects, $\boldsymbol{\mu}_D$ for the prior mean vector and $\boldsymbol{F}_D$ for the size $|D| \times J$ matrix of covariates at these design locations. Building on properties of Gaussian processes, the joint distribution of $(x_s, \boldsymbol{x}_D')' = (\boldsymbol{f}_s'\boldsymbol{\beta}, \boldsymbol{F}_D\boldsymbol{\beta})' + (w_s, \boldsymbol{w}_D')'$ is Gaussian with mean $(\mu_s, \boldsymbol{\mu}_D')'$ and covariance matrix

$$\text{Var}[(x_s, \boldsymbol{x}_D')'] = \begin{bmatrix} \sigma_s^2 & \boldsymbol{\Sigma}_{s,D} \\ \boldsymbol{\Sigma}_{D,s} & \boldsymbol{\Sigma}_D \end{bmatrix}, \tag{8}$$

with $\boldsymbol{\Sigma}_{D,s}$ being a length $|D|$ vector holding all the cross-covariance terms between variable x_s and the linear predictor variables in the design D, that is, $\boldsymbol{\Sigma}_{s,D} = \boldsymbol{f}_s\boldsymbol{\Sigma}_\beta^0\boldsymbol{f}_D' + \sigma^2\text{Corr}(w_s, \boldsymbol{w}_D)$, while $\boldsymbol{\Sigma}_D$ is a $|D| \times |D|$ matrix with variance-covariance terms within all the design location variables.

By standard Gaussian expressions, the conditional distribution of x_s given $\boldsymbol{x}_D$ is then Gaussian with mean and variance

$$m_s = \mu_s + \boldsymbol{\Sigma}_{s,D}\boldsymbol{\Sigma}_D^{-1}(\boldsymbol{x}_D - \boldsymbol{\mu}_D), \quad \xi_s^2 = \sigma_s^2 - \boldsymbol{\Sigma}_{s,D}\boldsymbol{\Sigma}_D^{-1}\boldsymbol{\Sigma}_{D,s}. \tag{9}$$

In our setting the data are binary, and there is no closed form like Equation (9) for the conditional mean and variance. One can however derive approximate expressions for the expected variance reduction from binomial data. In the VOI approximation below we rely on the following expression from Evangelou and Eidsvik (2017) for the variance reduction associated with binomial measurements

$$\begin{aligned} \chi_s^2 &= \boldsymbol{\Sigma}_{s,D}[\boldsymbol{\Sigma}_D + \boldsymbol{K}_D]^{-1}\boldsymbol{\Sigma}_{D,s}, \\ \boldsymbol{K}_D &= \text{diag}\left\{2 + \exp\left(-\mu_s + \frac{\sigma_s^2}{2}\right) + \exp\left(\mu_s + \frac{\sigma_s^2}{2}\right); \boldsymbol{u}_s \in D\right\}. \end{aligned} \tag{10}$$

Comparing with the variance reduction in Equation (9), we notice an additional $\boldsymbol{K}_D$ for the center matrix that is inverted to get χ_s^2 in Equation (10). This means that the variance reduction is smaller than when observing the linear predictors directly. Moreover, the magnitudes of this matrix $\boldsymbol{K}_D$ depend on the mean μ_s and variance σ_s^2 at the design locations.

3.3 | The value of information for spatial binary data

By using the logistic model formulation, the VOI contribution at lake s in Equation (5) equals

$$\text{VOI}_s(D) = \sum_{\boldsymbol{y}_D \in \{0,1\}^{|D|}} \max\left\{R_s \cdot \mathbb{E}\left(\frac{e^{x_s}}{1+e^{x_s}} | \boldsymbol{y}_D\right) - C_s,\ 0\right\} p(\boldsymbol{y}_D) - \max\left\{R_s \cdot \mathbb{E}\left(\frac{e^{x_s}}{1+e^{x_s}}\right) - C_s,\ 0\right\}. \tag{11}$$

There is no closed-form expression for Equation (11), and we next outline an approximate solution building on the results in Section 3.2.2.

3.3.1 | Approximating the VOI

We rely on an analytical approximation of the VOI developed by Evangelou and Eidsvik (2017). The VOI is computed using the Laplace approximation based on Gaussian approximations in Equations (9) and (10), in combination with normal cumulative distribution function (cdf) fitting of the logistic function. We discuss these in some more detail next.

First, the conditional expectation of $e^{x_s}/(1+e^{x_s})$ in Equation (11) is approximated by linearizing the logistic likelihood and quadratic fitting of the curvature giving Equation (10). In doing so, the integral depends on the unknown conditional mode (approximate Gaussian distributed with variance in Equation (10)) rather than the discrete data. Next, we build upon the idea of approximating the logistic function $g(x_s) = e^{x_s}/(1+e^{x_s})$ by the normal cdf $\Phi(\alpha x_s)$ for an appropriately selected scaling parameter α. Depending on the criterion one uses to minimize the mismatch between the two functions, one gets a different α. We choose $\alpha = 0.59$, which is one of the scaling parameters mentioned in Demidenko (2013). The two functions are then very close in a large span of x_s values. Finally, we compute the complete and incomplete logistic-normal integrals by

$$\begin{aligned}
\Lambda(\mu,\sigma^2) &= \int_{-\infty}^{\infty} \frac{e^x}{1+e^x}\varphi(x;\mu,\sigma^2)dx \approx \int_{-\infty}^{\infty} \Phi(\alpha x)\varphi(x;\mu,\sigma^2)dx = \Phi\left(\frac{\alpha\mu}{\sqrt{1+\alpha^2\sigma^2}}\right)\\
\Lambda_a(\mu,\sigma^2) &= \int_{a}^{\infty} \frac{e^x}{1+e^x}\varphi(x;\mu,\sigma^2)dx \approx \int_{a}^{\infty} \Phi(\alpha x)\varphi(x;\mu,\sigma^2)dx\\
&= \Phi\left(\frac{\mu-a}{\sigma}\right) - \Phi_2\left(\frac{\mu-a}{\sigma}, -\frac{\alpha\mu}{\sqrt{1+\alpha^2\sigma^2}}; \frac{\alpha\sigma}{\sqrt{1+\alpha^2\sigma^2}}\right),
\end{aligned} \tag{12}$$

where $\varphi(x;\mu,\sigma^2)$ denotes the normal probability density function evaluated at x with mean μ and variance σ^2, and $\Phi_2(z_1,z_2;r)$ is the bivariate standard normal cdf with correlation r, evaluated at (z_1,z_2).

The VOI_s in Equation (11) is then approximated by

$$\begin{aligned}
\text{VOI}_s(D) \approx\ & R_s\Lambda_a\left(\frac{\mu_s}{\sqrt{1+\alpha^2\xi_s^2}}, \frac{\chi_s^2}{1+\alpha^2\xi_s^2}\right) - R_s g(a)\Phi\left(\frac{\mu_s - a\sqrt{1+\alpha^2\xi_s^2}}{\chi_s}\right)\\
& - R_s \max\{\Lambda(\mu_s, \xi_s^2+\chi_s^2) - g(a),\ 0\},
\end{aligned} \tag{13}$$

where $a = \log([C_s/R_s]/(1-[C_s/R_s]))$ and $g(a) = 1/(1+e^{-a})$. Evangelou and Eidsvik (2017) use extensive Monte Carlo simulations to study the properties of this approximation for binomial data and Poisson distributed data. Similar expressions have been used to approximate the expected Bernoulli variance in logistic models (Anyosa et al., 2023).

3.3.2 | Search algorithms for optimal designs

Our aim is to find designs with large VOI. Ideally, this entails solving an optimization problem as follows:

$$D^{\dagger} = \arg\max_{D}\{\text{VOI}(D)\}, \quad \text{VOI}(D) > P(D), \tag{14}$$

where $P(D)$ is the cost of gathering the monitoring data from the design D. There may also be interest in maximizing the gap between the information value and the design cost, that is, $\text{VOI}(D) - P(D)$.

In general, the optimal design problem in Equation (14) is NP-hard because of the enormous number of combinations. With no constraints on $|D|$, there are 2^N possible designs. Even if we limit the scope to fixed size designs, there are $N!/[(N-|D|)!|D|!]$ possible designs. With $N = 4748$ and $|D| = 50$ this number of combinations is enormous (about 10^{100}). It is hence infeasible to analyze all available combinations and heuristics are needed.

We next describe a forward selection algorithm aimed to maximize the VOI up to a certain size of designs D. In Algorithm 1, the heuristic approach sequentially adds observation locations $j = 1, 2, \ldots$ to the design. This continues until the maximum size is reached. In the extreme event one continues until size N, but in practice it stops for $|D| << N$, when the VOI increase is negligible from j to $j+1$, or when the VOI is clearly too small to justify purchasing all that data. Instead of choosing just one extra lake in the design at each stage, one can choose more sites at a time. If two lakes are equally good in the forward evaluation, the selection between them is performed randomly.

The forward selection algorithm presented here often gives reasonable designs, but it is only a heuristic search, which has no guarantee of returning the optimal design. More complex search methods for efficient sampling designs include variants of the randomized exchange algorithm (see, e.g., Harman et al., 2020). This defines an iterative search among new (random) combinations of designs. In one of its forms, which we use for our data below, each iteration includes an

Algorithm 1. Forward selection of design

$j = 1$,
$D = \emptyset$ ▷ set of already selected sites
while $j \leq N$ **do**
  for $i = 1, \dots, N$ and $\boldsymbol{u}_i \notin D$ **do**
    $D^{(i)} = D \cup \{\boldsymbol{u}_i\}$ ▷ Candidate design
    $\text{VOI}(D^{(i)}) = \sum_{s=1}^{N} \text{VOI}_s(D^{(i)})$ ▷ VOI of candidate design
  end for
  $i^* = \arg\max_i \{\text{VOI}(D^{(i)}); i = 1, \dots, N \text{ and } \boldsymbol{u}_i \notin D\}$ ▷ optimal new design site
  $D = D \cup \{\boldsymbol{u}_{i^*}\}$
  $j = j + 1$,
end while

exchange where one lake is removed from the design and another is added to the design. The exchange probability is in our case guided by single-location results: $\text{VOI}(\{\boldsymbol{u}_s\})$, that is, assuming $N = |D| = 1$ for all $s = 1, \dots, N$. Lakes with a large single-lake VOI are hence more likely to be added to the design, while the ones with small single-lake VOI are more likely to be removed from the design. Still, the probabilities are positive for including or excluding any lake to the design, and this ensures some randomness helping the optimization approach from getting stuck in a local optimum. For our dataset we also test the approach of Paglia et al. (2022), who used Bayesian optimization and expected improvement to search for promising designs. The Hausdorff distances between designs D are used to form a covariance matrix in a Gaussian process surrogate model for $\text{VOI}(D)$, taking D as input. One learns this Gaussian process from previous VOI evaluations. At each iteration, a batch of promising designs are selected as the ones having high expected improvement in VOI according to the surrogate model. This is a fast calculation. After this selection, all designs in the batch go to the much more costly VOI evaluation.

4 | RESULTS

The modelling, preliminary steps, and greedy algorithm were implemented in R (R Core Team, 2021). The randomized exchange algorithm and the Bayesian optimization algorithm were implemented with Matlab (MATLAB, 2021).

4.1 | Modelling and preliminary steps

We used a standard logistic regression model to determine the important covariates. Candidate covariates were center latitude and longitude coordinates, waterbed area (1000 square kilometers), length of shoreline (kilometers), and by municipality, the agricultural area, population and number of summer residences scaled with municipality area. Additional explanatory variables that are challenging to measure, such as drainage basin, average depth, maximum depth and volume of water mass, were not included in the analysis due to the high number of missing values.

We fitted models that contained each of the seven covariates one at a time. The ones that seemed important on their own based on $-2\hat{l}$, where $\hat{l}$ is the log-likelihood of the logistic regression model, were then analyzed more closely. The covariates that had a significant effect at this point were the latitude and longitude coordinates and the agricultural area scaled with municipality area. We computed the change in the value of $-2\hat{l}$ when each variable on its own was omitted. Only those that lead to a significant increase in the value of $-2\hat{l}$ were retained in the model. As the result of the selection, we chose the latitude coordinate and the agricultural land in the municipality where the lake was located as covariates. We specified $\boldsymbol{\mu}_\beta^0$ and $\boldsymbol{\Sigma}_\beta^0$ as the approximate mean and covariance of $\boldsymbol{\beta}$, given the initial data. Using scaled agricultural land and latitude coordinate as covariates, the model has $\boldsymbol{\mu}_\beta^0 = \hat{\boldsymbol{\beta}} = (1.6, -13.8, 0.02)$, $\text{diag}(\boldsymbol{\Sigma}_\beta^0) = (3.9, 0.5, 0.001)$ and a substantial correlation of -0.6 between intercept and slope with scaled agricultural type. Goodness of fit measures show that the model fits reasonably well to the existing data (deviance statistics).

We then used the Laplace approximation (see, e.g., Shun & McCullagh, 1995) to estimate the covariance parameters of the spatial random effect. For the variability of the spatially structured variables we get an estimate of $\sigma = 1.11$. For the

correlation decay we get $\phi = 0.06$, meaning that the spatial correlation is reduced to 0.05 at a distance of 80 km. We were also interested in seeing whether the selection methods are sensitive to changes in the spatial covariance parameter values. Thus, we varied (σ, ϕ) values. Based on the second derivatives of the marginal log-likelihood function of the parameters, 1 standard deviation up and down (in the log-space) gives $\sigma = 0.99$ as a low value and $\sigma = 1.26$ as a high value for the scale. Similarly, this gives $\phi = 0.047$ as a low value and $\phi = 0.077$ as a high value of the correlation decay. In what follows, we tried five different sets of the spatial covariance parameter values: $(\sigma, \phi) = (1.11, 0.06)$ as benchmark parameter values and in addition, $(\sigma, \phi) = \{(0.99, 0.06), (1.26, 0.06), (1.11, 0.047), (1.11, 0.077)\}$.

Regarding the design, as indicated in the workflow outlined in Section 2.2, we first reduced the set of possible designs. This was done by reducing the set of 54,347 lakes to 4748 lakes. First, from the mean value $\boldsymbol{\mu}_\beta^0$ and the covariates for those lakes, we computed predictive probabilities of a lake being in the target status, as in Equation (6) (see Figure 2). From the revenue R_s and the cost C_s of lake s, we calculated the prior value PV_s, as in Equation (1), for each lake. We selected 1000 lakes which have the first term in the maximum in Equation (1), that is, $R_s \cdot \mathbb{E}(\pi_s) - C_s$, close to zero. We assumed these lakes are among the most interesting in the sense of the VOI evaluations. Second, we calculated the VOI with only single sites in the design, that is, $\text{VOI}_s(\{\boldsymbol{u}_s\})$ (referred to as the self-effect). When calculating self-effect, we assumed that $N = |D| = 1$ for all s. We assumed that if this self-effect is minuscule, then that lake is unlikely to have a significant effect on the total VOI with all N lakes included. There were 3798 lakes that have a self-effect $\text{VOI}_s(\{\boldsymbol{u}_s\}) >$ EUR 138. Partly, these lakes overlap with the first 1000 lakes selected. We combined the first and second selected lakes and limited our optimal design selection into the resulting $N = 4748$ lakes.

4.2 | Selection of data sites

The greedy approach in Algorithm 1 was used to construct a relatively small-size design from the 4748 lakes. In doing so, we conducted the subset selection with greedy forward selection. We ran this approach until $|D| = 300$ lakes were included in the design. The VOI of that forward-selected subsample is shown in Figure 3 (left) with varying spatial covariance parameter values (σ, ϕ). Naturally, the VOI of the design increases as the number of lakes in the design grows.

Our aim was to compare the VOI of a design to the cost of gathering data in that design. We assumed that each selected lake implies one sampling site. Currently, collecting and analyzing one chlorophyll-*a* sample costs EUR 138

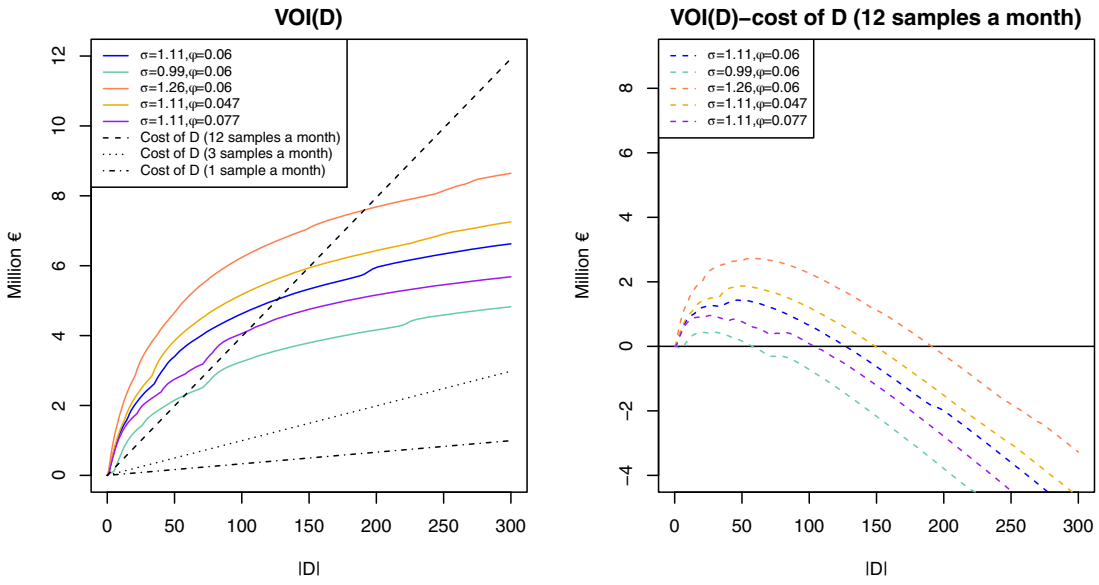


FIGURE 3 Left: The VOI (solid colored curve) and the cost of data gathering (dashed curves) of design of size $|D|$ in million euro plotted with respect to the $|D|$. The color-coded curves show the calculation for different spatial covariance parameter values (σ, ϕ). We assume three alternatives for data gathering: a large amount of data (12 samples a month), an average amount of data (3 samples a month) or a small amount of data (1 sample a month) gathered per site. Right: The difference between the VOI for different spatial covariance parameter values (σ, ϕ) and the cost of gathering a large amount of data in million euro.

(Koski et al., 2020). The total cost of chlorophyll-a data gathering from a design of size $|D|$ was assumed to consist of the samples gathered from four months per year and during six years, as it is the length of the RBMP period. We assumed three data gathering alternatives: either a large amount of data (12 samples a month), an average amount of data (3 samples a month) or a small amount of data (1 sample a month) per site is gathered during that time. Processing of these data gives the ecological classification $y_s = 0$ or $y_s = 1$ for each lake s. Note that we are using the same model for ecological status class y_s, no matter what acquisition and processing is required for the chlorophyll-a samples.

The three dashed curves in Figure 3 (left) illustrate costs of sampling plans. Generally, the studied VOI results of the selected subsamples seem to exceed the cost of gathering that subsample, meaning that the data acquisition is worth doing. When assuming twelve samples in a month, the cost exceed the VOI (calculated with benchmark parameter values (σ, ϕ)) after selecting $|D| = 126$ lakes in the design. Then, VOI(D) = EUR 5.03 million and $P(D)$ = EUR 5.01 million. An even higher VOI value is reached when using $(\sigma, \phi) = (1.26, 0.06)$. In that case, the 12 samples a month cost is reached after selecting $|D| = 192$ lakes, when VOI(D) = EUR 7.60 million and $P(D)$ = EUR 7.63 million.

The gap between the VOI with varying spatial covariance parameter values (σ, ϕ) and the cost of gathering 12 samples a month from the design is shown in Figure 3 (right). This illustrates the excess information value over the cost of the data for different sample sizes. When assuming 12 samples a month and benchmark parameter values (σ, ϕ), the most benefit is achieved when 47 lakes are measured. Then, VOI(D) = EUR 3.26 million and $P(D)$ = EUR 1.83 million, which gives a gap of EUR 1.43 million. When using parameter values $(\sigma, \phi) = (1.26, 0.06)$, the most benefit is achieved when 59 lakes are measured. Then, VOI(D) = EUR 5.03 million and $P(D)$ = EUR 2.31 million, which gives a gap of EUR 2.72 million.

Figure 4 illustrates the selection on the map of Finland when varying the spatial covariate parameter values (σ, ϕ). The circle colors illustrate the order of the 300 selected lakes. The lakes that the algorithm did not include in the design are

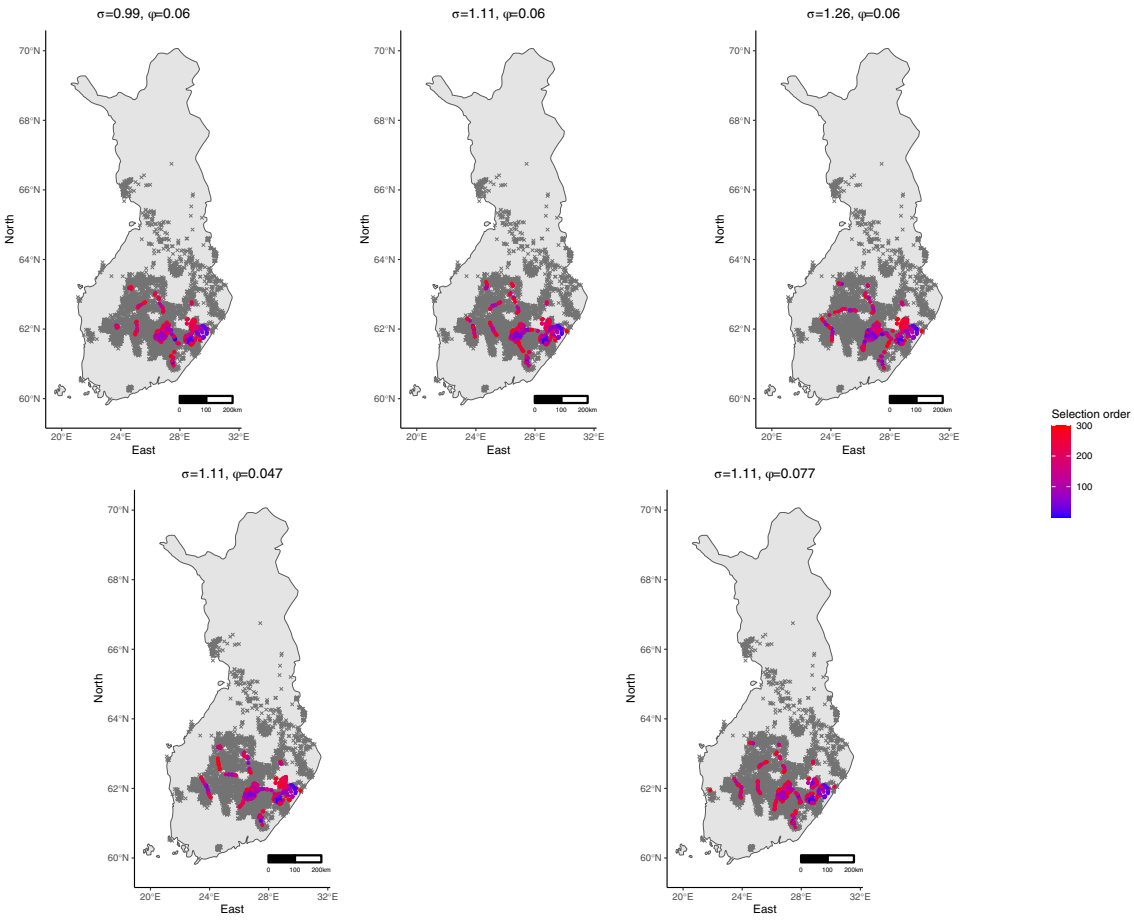


FIGURE 4 Map view indicating the order of 300 selected lakes from the 4748 interesting lakes. The sequential selection is computed using different spatial covariance parameters (σ, ϕ). The red circles are the selected lakes by algorithm in order and the crosses are the lakes excluded in the design.

depicted with crosses. The design selection appears to be similar with very few differences, regardless of the parameter values. All the selected designs are clustered in the southeast region of Finland, which is known to be rich in water areas.

As observed, there are no selected lakes in the northern region. To examine this further, we focused on this region when the 300th lake is selected, as an anecdotal example. When we have selected $|D| = 299$ lakes in the design, there are 4449 potential lakes to be the 300th selected lake. All potential lakes are ranked in descending order based on the VOI when a single lake is added to the current design. From the Northern area, the highest ranked lake is Lake Kattilajärvi (66.16°N, 24.40°E), which is only the 2292nd ranked. Initially, this lake has $\pi_s = \mathrm{P}(y_s = 1) = 0.80$ and a self-effect of $\mathrm{VOI}_s(\{\boldsymbol{u}_s\}) =$ EUR 59. If Lake Kattilajärvi is selected, the total VOI with $|D| = 300$ is $\mathrm{VOI}(D) =$ EUR 6,623,243. According to the sequential selection algorithm, the 300th selected lake is Lake Vääräjärvi (63.30°N, 26.43°E) and the total VOI after that selection is $\mathrm{VOI}(D) =$ EUR 6,626,647.

Figure 5 illustrates what kind of lakes are the most important to measure from the VOI point of view, along with their relationship to the agricultural land (first axis) and the waterbed area (second axis). The color-coded circles show the selection order of the $|D| = 300$ lakes, when using the benchmark parameter values of (σ, ϕ). The crosses are the lakes not included in design. We have denoted the relevant $N = 4748$ lakes with color-coded crosses. The selection order of lakes indicates that the selection covers most lake types but not the ones with very large waterbed and low agricultural land covariates. There is a tendency to choose big lakes early in the design order, but small and average size lakes are also chosen. A closer inspection shows that 48% of the $|D| = 300$ selected lakes are selected from the top 10% largest lakes (19.18–71258.50 ha), while 38% are selected from the next smallest decile (10.25–19.18 ha). The selection further

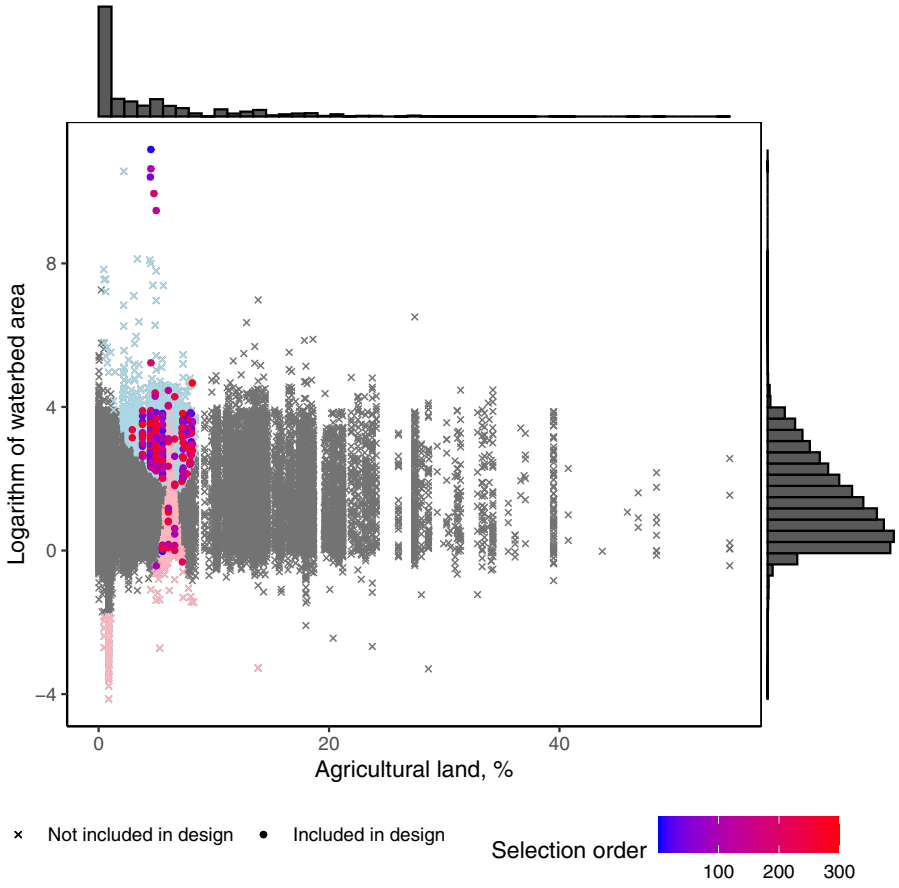


FIGURE 5 The relation between agricultural land (first axis) and waterbed area (second axis) of the full data of 54,347 lakes with the marginal distributions. The circles are the selected lakes by algorithm and the crosses are the lakes excluded in the design. The color-code shows the selection order of the 300 selected lakes. In addition, the color-coded crosses show the 4748 interesting lakes selected based on two criteria: PV_s (light pink) and self-effect VOI (light blue).

seems to prefer high to average agricultural land covariate values. In fact, 18% of the $|D| = 300$ selected lakes are selected from those areas with the top 20% amount of agricultural land (7%–54% agricultural land of the municipal area), 81% are selected from the next 20% (3%–7% agricultural land of the municipal area) and only 1% are selected from municipalities with a lower amount of agricultural land (0%–0.8% agricultural land of the municipal area). This makes sense because there is more ambiguity in the management decision for average agricultural land covariates, leading to high VOI values. From a purely statistical perspective, one would expect very high and low covariates to provide more information about the regression parameter, and in doing so reduce the uncertainty going into the VOI calculations. Here, there is a large amount of data at the initial step, and this element of regression fitting appears to be less relevant in the design.

5 | DISCUSSION

The results in Figures 3–5 show the performance of a forward selection strategy to find useful designs. We now compare different designs of size $|D| = 50$, with the goal of searching for the optimal design. This will tell us if the sequential method performs reasonably or if this way of greedy augmentation of designs overlooks high-value sampling designs. The size of $|D| = 50$ is chosen because the gap between the VOI and the curve for the cost of a scenario with 12 chlorophyll-a samples in Figure 3 appears to be at its largest for this design size.

We search for more optimal designs based on the exchange algorithm and the Bayesian optimization approach of Paglia et al. (2022). For the exchange algorithm, we start the iterative routines with the optimal set of size $|D| = 50$ from the forward evaluation. The exchange of two lakes (one removed from the design and the other added to the design) is based on probabilities vaguely honoring high marginal self-effect of VOI. The Bayesian optimization algorithm starts with 1000 evaluations of the exchange algorithm, and continues with batches of 100 VOI evaluations selected from the expected improvement over 1000 designs.

Figure 6 shows the percentage increase in the VOI as a function of iterations of the two algorithms. This is illustrated for VOI evaluation number on the first axis, and over ten independent runs of the exchange algorithm (solid, red) and

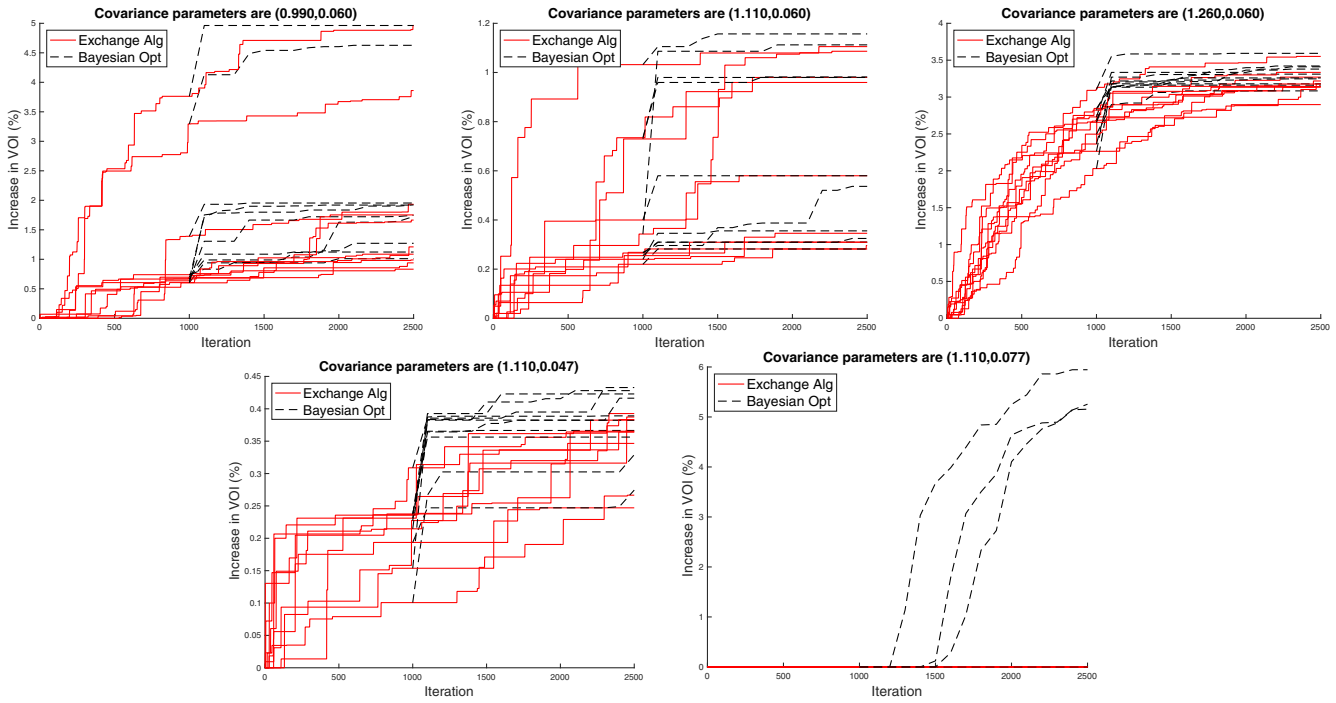


FIGURE 6 VOI results of the exchange algorithm (solid, red) and the Bayesian optimization approach (dashed, black) for 10 independent runs. The five displays reflect different spatial covariance parameters: low variance (top left), benchmark inputs (top center), high variance (top right), low correlation decay (bottom left), high correlation decay (bottom right). Results are shown as percentage VOI increase over evaluations, relative to the sequential forward selection results for a design size of 50. The first 1000 iterations are common. After that the Bayesian optimization algorithm runs batches with a size of 100 using expected improvement of designs.

TABLE 2 VOI(D) computed for different designs of size $|D| = 50$ using different spatial covariance parameters (σ, ϕ). Here, M refers to million euro and K refers to thousand euro.

	VOI(D)				
	$\sigma = 1.11$	$\sigma = 0.99$	$\sigma = 1.26$	$\sigma = 1.11$	$\sigma = 1.11$
Design	$\phi = 0.06$	$\phi = 0.06$	$\phi = 0.06$	$\phi = 0.047$	$\phi = 0.077$
Sequential best	3.40 M	2.15 M	4.65 M	3.86 M	2.75 M
Exchange	3.43 M	2.25 M	4.81 M	3.87 M	2.75 M
Bayes optimization	3.44 M	2.26 M	4.82 M	3.88 M	2.91 M
Highest self-effect	860 K	511 K	1.29 M	1.32 M	546 K
Largest lakes	548 K	-	-	-	-
Geo-spreading	381 K	-	-	-	-
High agriculture	390	-	-	-	-

the Bayesian optimization scheme (dashed, black). The reference case (center, top row) has parameters $\sigma = 1.11$ and $\phi = 0.06$. For the reference case, both the exchange algorithm and Bayesian optimization obtain better designs than the greedy result. The largest VOI improvement for the exchange algorithm is about 1.1% while it is 1.2% using Bayesian optimization. The other displays show VOI increase in cases where the model parameters indicate high/low variance or high/low spatial correlation. Similar to what we see in the reference case, there are designs with higher VOI than that achieved by the sequential forward selection, which indicates that more nuanced algorithms could further improve this. Nevertheless, from what we see here, it is not straightforward to get a much higher value than that obtained by the sequential forward selection (it is only about 0%–5%). For the case with the fast spatial correlation decay parameter, none of the ten exchange algorithm runs, and only three of the ten Bayesian optimization runs, managed to improve the design. This case represents less dependence, and intuitively the sequential method performs better. Still, some of the new designs detected by Bayesian optimization are significantly better, but the search seems more difficult with these parameter settings.

We will now compare designs with a size of 50 based on other criteria. Going beyond the statistical models and decision analytic views, policy makers could have other elements they must consider, and it is insightful to show the VOI results of designs based on a variety of principles. Again, we focus this discussion on designs of size $|D| = 50$. First, we select 50 lakes with the highest self-effect $\text{VOI}_s(\{\boldsymbol{u}_s\})$ from the whole lake data. Second, we calculate the VOI of the 50 lakes with the largest waterbed area from the whole data. Third, we test the set of 50 lakes which are spread out as much as possible on the map of Finland. The spreading was set by selecting the lakes from each county of Finland. There are 18 counties in our data, meaning we randomly selected 2 or 3 lakes from each county. Fourth, we also formed a design with the 50 lakes having highest covariate value (agricultural land in Figure 2, left). Table 2 summarizes the VOI of the greedy selection, the exchange and Bayesian optimization selection (maximum of 10 runs doing 2500 evaluations), as well as the other designs with simpler selection criteria listed above, when varying the statistical covariance parameter values (σ, ϕ). It seems that the VOI of these designs remain very small compared to the results we achieve with the statistical algorithms. Large values of σ seem to produce larger values of VOI.

We therefore recommend that policy makers use statistical methods in the design construction. Making monitoring plans on the highest self-effect alone misses out on the correlations in the statistical model and the interactions of having very similar lakes in a design. For the designs with a size of 50, it gets less than one-fourth of the VOI compared with the more nuanced search approaches. Designs that either focus on geographical coverage, large lake area or high agricultural covariates do not necessarily capture the interesting lakes for ecological purposes. The greedy algorithm succeeds in finding a reasonably good design at moderate computation costs here, and provides a reference for the additional search approaches.

6 | CONCLUSION

We have demonstrated approximate optimal design selection methods that aim to maximize the VOI of the design compared with the cost of the data acquisition of the design. The VOI selection criterion assesses the profitability of designs

when taking into account the costs and benefits of the decisions as well as the associated uncertainty. This approach is exemplified in the context of lake management in Finland. Similar design questions occur in a range of applications, such as the environmental studies brought up in Section 1 related to coral monitoring, animal habitat conservation or the mapping of mine tailings. Decision-makers must plan wisely where to conduct environmental sampling so as to obtain valuable information and to maintain budget limitations.

We calculated the VOI assuming a Bayesian spatial logistic regression model for the ecological status data. Statistical model parameters were obtained from existing data gathered in Finnish lakes. Our VOI calculations relied on approximations of functions and integrals for hierarchical general linear models, which we coupled with the large-size design selection procedures required for the lake monitoring case.

In addition to the heuristic greedy forward selection method, which sequentially adds units into the design, we tested two other heuristics for improved selection result: an exchange algorithm based on randomness and enlightened exchange based on the single-lake VOI, and a selection algorithm based on Bayesian optimization. The VOIs achieved with these statistical approaches were much higher than that of other design criteria based on the initial marginal values, geographical spread, high model covariate values or large lake areas.

We are aware that many considerations must be made in order to calculate the VOI of lake monitoring design in practice, and we are limiting our results. For example, we chose to use chlorophyll-*a* as our ecological indicator of interest, among many, and we focused on the costs of collecting that one indicator. In reality, the lake monitoring process is a more complex exercise. Furthermore, analyzing the monitoring data of one lake produces the ecological status of that lake, and it does not take into account how much monitoring data was used for the classification. In addition, we thought that associating the costs and revenues to the lake areas would have an impact on the results. However, the area seems to have less effect on the selection than we assumed.

Since the problem of optimal design has been widely examined in statistics, there exist many other heuristic methods to solve this problem. We believe it is possible to obtain better designs than we did here. Our purpose was to highlight the possibility of forming a statistically based design for these large-size spatial logistic regression models, and in doing so we see that they clearly outperform designs made from basic principles.

This study does not consider any temporal variation in the ecological status of lakes. Spatio-temporal variation in lake status would thus be interesting to address in future work. In this paper, we relied on earlier studies considering the management decision space and associated costs. In the future it would be relevant to expand the space of management alternatives to a more detailed level, and see how this influences the selection of lakes for the design.

ACKNOWLEDGMENTS

VK acknowledges support from the Emil Aaltonen Foundation and the Kone Foundation. JE acknowledges support from the Norwegian Research Council via relevant grants 305445 and 309960.

CONFLICT OF INTEREST STATEMENT

The authors have no conflict of interest to declare.

DATA AVAILABILITY STATEMENT

The data supporting the analyses are downloadable from https://nextcloud.jyu.fi/index.php/s/2BjQjkzWsaBYRr8. The data that support the findings of this study are openly available in Koski&Eidsvik2022 at https://nextcloud.jyu.fi/index.php/s/PmqmRDTp4F4sC3Z.

ORCID

Vilja Koski https://orcid.org/0000-0002-5970-3582

REFERENCES

Abbas, A. E., & Howard, R. A. (2015). *Foundations of Decision Analysis*. Pearson Higher Ed.

Anyosa, S., Eidsvik, J., & Pizarro, O. (2023). Adaptive spatial designs minimizing the integrated bernoulli variance in spatial logistic regression models-with an application to benthic habitat mapping. *Computational Statistics & Data Analysis*, *179*, 107643.

Aroviita, J., Mitikka, S., & Vienonen, S. (2019). *Pintavesien tilan Luokittelu ja Arviointiperusteet Vesienhoidon Kolmannella Kaudella*. Finnish Environment Institute (SYKE) (In Finnish). https://helda.helsinki.fi/handle/10138/306745Placeholder Text.

Arthur, J., Hachey, M., Sahr, K., Huso, M., & Kiester, A. (1997). Finding all optimal solutions to the reserve site selection problem: formulation and computational analysis. *Environmental and Ecological Statistics*, *4*, 153–165.

Canessa, S., Guillera-Arroita, G., Lahoz-Monfort, J. J., Southwell, D. M., Armstrong, D. P., Chadès, I., Lacy, R. C., & Converse, S. J. (2015). When do we need more data? A primer on calculating the value of information for applied ecologists. *Methods in Ecology and Evolution*, *6*(10), 1219–1228.

Demidenko, E. (2013). *Mixed Models: Theory and Applications with R*. John Wiley & Sons.

Dudgeon, D., Arthington, A. H., Gessner, M. O., Kawabata, Z., Knowler, D. J., Lévêque, C., Naiman, R. J., Prieur-Richard, A. H., Soto, D., Stiassny, M. L., & Sullivan, C. A. (2006). Freshwater biodiversity: Importance, threats, status and conservation challenges. *Biological Reviews of the Cambridge Philosophical Society*, *81*(2), 163–182.

Dykstra, O. (1971). The augmentation of experimental data to maximize [X'X]. *Technometrics*, *13*(3), 682–688.

Eidsvik, J., Mukerji, T., & Bhattacharjya, D. (2015). *Value of Information in the Earth Sciences: Integrating Spatial Modeling and Decision Analysis*. Cambridge University Press.

European Parliament (2000). Directive 2000/60/EC, of the European Parliament and Council of 23 October 2000 establishing a framework for Community action in the field of water policy.

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

Foss, K. H., Berget, G. E., & Eidsvik, J. (2022). Using an autonomous underwater vehicle with onboard stochastic advection-diffusion models to map excursion sets of environmental variables. *Environmetrics*, *33*(1), e2702.

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*(529), 348–361.

Hays, S., Kumari, B., Stewart-Koster, B., Boone, E., & Sheldon, F. (2021). Site reduction in redundant ecosystem sampling schemes. *Environmental and Ecological Statistics*, *28*, 1–20.

Higgins, J., Zablocki, J., Newsock, A., Krolopp, A., Tabas, P., & Salama, M. (2021). Durable freshwater protection: A framework for establishing and maintaining long-term protection for freshwater ecosystems and the values they sustain. *Sustainability*, *13*(4), 1950.

Jauslin, R., Panahbehagh, B., & Tillé, Y. (2022). Sequential spatially balanced sampling. *Environmetrics*, *33*(8), e2776.

Jeppesen, E., Søndergaard, M., & Liu, Z. (2017). Lake restoration and management in a climate change perspective: An introduction. *Water*, *9*(2), 122.

Koski, V., Kotamäki, N., Hämäläinen, H., Meissner, K., Karvanen, J., & Kärkkäinen, S. (2020). The value of perfect and imperfect information in lake monitoring and management. *Science of the Total Environment*, *726*, 138396.

MATLAB. (2021). *Version 9.11.0 (R2021b)*. The MathWorks Inc.

Nguyen, L., Ulapane, N., & Miro, J. V. (2018). *Adaptive sampling for spatial prediction in environmental monitoring using wireless sensor networks: A review*. In *2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA)* (pp. 346–351). IEEE.

Nõges, P., van de Bund, W., Cardoso, A. C., Solimini, A. G., & Heiskanen, A.-S. (2009). Assessment of the ecological status of European surface waters: A work in progress. *Hydrobiologia*, *633*, 197–211.

Official Statistics of Finland. (2020a). *Buildings and free-time residences [e-publication]*. Statistics Finland http://www.stat.fi/til/rakke/index·en.html

Official Statistics of Finland. (2020b). *Utilised Agricultural Area [e-publication]*. Natural Resources Institute Finland http://www.stat.fi/til/kaoma/index·en.html

Paglia, J., Eidsvik, J., & Karvanen, J. (2022). Efficient spatial designs using Hausdorff distances and Bayesian optimization. *Scandinavian Journal of Statistics*, *49*(3), 1060–1084.

Prentius, W., & Grafström, A. (2022). Two-phase adaptive cluster sampling with circular field plots. *Environmetrics*, *33*(5), e2729.

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Reich, B. J., Pacifici, K., & Stallings, J. W. (2018). Integrating auxiliary data in optimal spatial design for species distribution modelling. *Methods in Ecology and Evolution*, *9*(6), 1626–1637.

Shun, Z., & McCullagh, P. (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(4), 749–760.

Søndergaard, M., Jeppesen, E., Lauridsen, T. L., Skov, C., Van Nes, E. H., Roijackers, R., Lammens, E., & Portielje, R. (2007). Lake restoration: Successes, failures and long-term effects. *Journal of Applied Ecology*, *44*(6), 1095–1105.

Stankey, G., Clark, R., & Bormann, B. (2005). *Adaptive Management of Natural Resources: Theory, Concepts, and Management Institutions*. USDA Forest Service General Technical Report PNW.

Thilan, A., Menéndez, P., & McGree, J. (2023). Assessing the ability of adaptive designs to capture trends in hard coral cover. *Environmetrics*, *34*, e2802.

Williams, B. K., & Brown, E. D. (2020). Scenarios for valuing sample information in natural resources. *Methods in Ecology and Evolution*, *11*(12), 1534–1549.

Woods, D. C., Overstall, A. M., Adamou, M., & Waite, T. W. (2017). Bayesian design of experiments for generalized linear models and dimensional analysis with industrial and scientific application. *Quality Engineering*, *29*(1), 91–103.

How to cite this article: Koski, V., & Eidsvik, J. (2024). Sampling design methods for making improved lake management decisions. *Environmetrics*, e2842. https://doi.org/10.1002/env.2842

IV

RISK AVERSION IN THE VALUE OF INFORMATION ANALYSIS: APPLICATION TO LAKE MANAGEMENT

by

Koski, V. & Karvanen, J. 2024

Submitted to Stochastic Environmental Research and Risk Assessment

Risk aversion in the value of information analysis: application to lake management

Vilja Koski and Juha Karvanen
Department of Mathematics and Statistics,
University of Jyvaskyla, P.O. Box 35, 40014 Jyväskylä Finland
✉vilja.a.koski@jyu.fi

November 13, 2024

Abstract

We analyze the relationship between the value of information (VOI) and the decision-maker's risk aversion in the lake monitoring where one needs to decide about whether to implement costly management actions or not. We calculate the value of perfect as well as imperfect information for risk neutral and risk averse decision-makers. The risk aversion is measured using the certain equivalent and the Arrow-Pratt risk aversion measures, which are defined using the derivatives of the utility function. We consider two utility functions for a risk averse decision-maker, an exponential utility and a power utility function, and demonstrate their use with lake management data from Finland. The results show that, in this context, a risk averse decision-maker's VOI may be lower or higher than a risk neutral decision-maker's VOI, depending on the prior probability of the lake being in the need of management actions and the cost of the actions. The risk aversion seems to have a clear impact on decisions. This may encourage the decision-makers to contemplate their risk preferences instead of hastily assuming the risk neutrality.

Keywords: imperfect information, perfect information, risk aversion, risk neutrality, utility function, value of information.

1 Introduction

The value of information (VOI) analysis is a common tool for real-life decision-makers in many fields of science, for example in finance and medicine [Eidsvik et al., 2015]. In brief, VOI is the expected monetary amount one should be willing to pay for additional data in decision-making situation. In the VOI analysis, the decision-maker is often assumed to be a risk neutral. However, in real life, a commonly held belief is that humans are risk averse [Davies and Satchell, 2007]. The relation of VOI and risk aversion has been actively studied in the economics and operations research but is rarely considered in other applications. This paper focuses on risk averse decision-making in environmental monitoring.

We apply the VOI analysis to the lake monitoring data of Finnish lakes. Due to the EU Water Framework Directive (WFD) [European Parliament, 2000], Finland is implementing a monitoring program of its inland waters. The classification of the lakes into five ecological status classes (high, good, moderate, poor and bad) is based on monitoring data on several ecological indicators. We are only interested in one of the ecological indicators, the chlorophyll-a concentration, which usually correlates well with the lake status. If the status of a lake is moderate or worse, the directive obliges its member countries to implement management actions to improve it. The question is whether the costly actions are needed or not. In the earlier literature, this kind of setting with a binary decision is often called a two-action problem. In our setting, the variable of interest is the binary ecological status of the lake which indicates whether the lake is in need of management actions or not. We have the continuous chlorophyll-a data, which is reflecting the status imperfectly. Then, the posterior distribution of the ecological status given the chlorophyll-a data is obtained by Bayes' rule.

We calculate the value of perfect and imperfect information for both a risk neutral and a risk averse decision-maker. The degree of risk aversion is specified via the parameters of an exponential utility or a power utility function. Since the parameters of the utility function can be difficult to determine directly, we suggest the use of a certain equivalent (see Section 2.3) as a way for the decision-maker to express the risk preferences. The value of imperfect information from the standpoint of a risk averse decision-maker is a topic that is not fully addressed in earlier research, which is summarized in Table 1.

The earliest studies on risk averse decision-maker's VOI seem to have focused on investigating who has a higher VOI, the risk neutral or the risk averse decision-maker. Hilton [1981] proves that

there is no general monotonic relationship between the degree of risk aversion and VOI. The prove is a numerical counterexample for a two-action decision problem with a binary state of nature. Mehrez [1985] shows for a class of unfavorable projects with nonpositive expected monetary value that a risk averse decision-maker will never pay more for perfect information than will a risk neutral decision-maker. Also, Freixas and Kihlström [1984] conclude that under certain restrictions, the demand for information decreases as risk aversion increases. Their idea is reiterated by [Willinger, 1989], in a different context. Willinger suggests that the relation of VOI and risk aversion is highly model specific. Eeckhoudt and Godfroid [2000] refer to works by Freixas and Kihlström and Willinger and aim to explain in simple terms why increased risk aversion does not always induce a greater value of information in a numerical illustration of a two-action decision problem with a binary response variable. Nadiminti et al. [1996] show that the relationship between risk aversion and the demand for information depends on the method of payment for the information, where the information can either be costless or costly, and the payment for costly information can be either ex-ante or contingent upon its positive incremental value.

More recent studies have focused on investigating the monotonicity of the VOI, and the general consensus is that the relationship of the degree of risk aversion and VOI is not monotonic in general but depends on the decision situation. Delquié [2008] shows that VOI is maximal when the decision-maker is indifferent between the two prior alternatives, and it is lower as the preference for one alternative over the others gets stronger. Delquié calculates VOI for imperfect information instead of perfect in the case of continuous variables. Bickel [2008] deepens the understanding of the value of imperfect information relative to perfect information in two-action linear-loss setting, where linear-loss refers to a situation with a continuous variable of interest, unlike in our problem setting. Abbas et al. [2013] analyze the relationship between risk aversion and the value of perfect information in a two-action and a continuous response setting when the initial wealth of a decision-maker is deterministic. They also consider the value of partition information, where a decision-maker receives information specifying that the outcome falls into an interval or a set of intervals. In their context, the VOI is called monotonic according to the degree of risk aversion if it is monotonic in the region where the original decision without the additional information remains the same. Sun and Abbas [2014] study the sensitivity of VOI with various measures of risk aversion in two-action decision problems when the initial wealth is unknown. As an application example, a decision-maker is

choosing whether or not to make an investment with uncertain return.

As far as we know, the effect of risk aversion on VOI is rarely considered outside economics or artificial examples. de Palma et al. [2012] analyze the decisions of drivers on whether to acquire information and which routes to take on simple congested road networks, varying the degree of risk aversion from risk neutral to very risk averse. Their study suggests that very risk averse drivers generally have lower VOI. Considering a risk averse decision-maker alongside a risk neutral one is an important perspective to study also in the environmental monitoring to obtain more accurate VOI estimates and in time, to better allocate the limited resources.

In addition, although the study by de Palma et al. [2012] addresses a real-life problem, it still does not utilize real data, as neither do the other articles reviewed in Table 1. In this study, we aim to use the real data to calculate the VOI.

The structure of this article is as follows. Section 2 introduces the used data and the statistical framework of the study. Section 3 interprets the results of VOI analysis of lake management application when varying the degree of risk aversion. Section 4 concludes with discussion of the results and further study directions.

2 Materials and methods

In this section, we start by introducing the data and formulating the decision problem in lake management. Then, we discuss the definition of risk aversion and formulate the value of information with differing risk preferences. We show how the value of imperfect information is assessed in this context.

2.1 Lake monitoring data

In order to put the WFD into practice in Finland, the River Basin Management Planning (RBMP) is implemented every six years [Aroviita et al., 2019]. During a six-year period, the monitoring of lakes is implemented, including data acquisition from several parameters representing biotic structure of the lake and supported by the physical and chemical properties of water as well as hydrological and morphological features. In this study, we are limiting our analysis to the chlorophyll-*a* concentration which is indicative of water body productivity and therefore generally correlates

Table 1: A summary of the literature reviewed in Introduction.

	Response x	Actions a	Type of information	Utility function	Application
Hilton [1981]	Binary	Binary	Perfect	Power	Information system
Mehrez [1985]	Continuous	Binary	Perfect	Exponential	Numerical example
Willinger [1989]	Continuous, normally distributed	Discrete	Perfect	Exponential	Investments
Nadiminti et al. [1996]	Binary	Binary	Perfect and imperfect	Exponential	Credit approval problem
Eeckhoudt and Godfroid [2000]	Binary	Binary	Perfect	Logarithmic	"The newsboy problem"
Delquié [2008]	Continuous	Binary	Imperfect	Exponential (and others)	Theoretical results only
Bickel [2008]	Continuous	Binary	Ratio= Imperfect/perfect	Exponential	Oil drilling
de Palma et al. [2012]	Continuous	Binary	Perfect	Exponential	Traffic equilibrium
Abbas et al. [2013]	Continuous	Binary	Perfect and partial	Power	Theoretical results only
Sun and Abbas [2014]	Binary	Binary	Perfect	Exponential, Gaussian and linear plus exponential	Investments

well with the ecological status of lentic water bodies suffering from human-induced eutrophication. Samples are collected from the observation sites in summertime (approximately from May to August). The monitoring data of chlorophyll-a is produced by the official Finnish lake monitoring program and stored in the open source database of the Finnish Environment Institute (`http://www.syke./en-US/Open_information`).

We use a dataset already introduced and used by Koski et al. [2020] who calculated the value of perfect and imperfect information in lake management assuming a risk neutral decision-maker. The data we are using is gathered from the years 2006–2012 and it consists of 6742 observations from 166 water bodies. We have selected the most frequently sampled water bodies with at least 3 summertime observations per year. We aggregate annual and local observations into means of annual medians per water body, which is the standard current approach in ecological status assessment of water bodies [Aroviita et al., 2019]. Of those 166 water bodies, 79 did not need management actions (high or good ecological status) while 87 needed them (moderate, poor or bad status).

2.2 Problem formulation

Assume a lake management situation where the decision-maker needs to decide about the management actions under uncertainty about the ecological condition of the lake. The ecological condition is defined by a binary random variable, where the decision-maker needs to decide whether the lake is in the need of management actions ($x = x_0$) or not ($x = x_1$). The binary variable has the probability $p(x) \geq 0$ of the state $x \in \Omega$ such that $\sum_{x \in \Omega} p(x) = 1$. We are unable to observe x directly, but we can measure the value y of a continuous random variable with the density $p(y)$ reflecting the state of x. Here, the variable y represents chlorophyll-a concentration of the lake, which indicates the ecological condition.

Then, the decision-maker can choose between alternatives $a \in A$: either to leave a lake untreated ($a = a_0$) or implement the management actions to bring the lake to a satisfying condition ($a = a_1$). The values $v(x, a)$ in monetary units associated with the alternatives are adopted from Koski et al. [2020] and Koski and Eidsvik [2024]. According to the valuation study by Ahtiainen [2008], a single lake in a condition where it does not need management actions with no performed actions is valued $v(x_1, a_0)$ = EUR 3 million/3000 hectare = EUR 1000 per hectare while the value of a lake in a condition where it needs actions is set to EUR 0 per hectare. Under the management actions,

it costs EUR 200 per hectare to bring the lake to a sufficiently good condition (source: Finnish Environmental Institute). Regardless of whether the management actions have been successful, the resulting value is EUR 1000 − EUR 200 = EUR 800 per hectare. A more detailed description of the lake valuation is provided by Koski et al. [2020]. See Table 2 for a summary of this two-action situation with binary x.

Table 2: A summary of costs and the monetary values for an example lake where the value of the target ecological status equals EUR 1000 per ha [Ahtiainen, 2008].

		Monetary value $v(x, a)$ (EUR/ha)	
		Ecological status x	
Alternative a	Cost of alternative (EUR/ha)	x_0: needs management actions	x_1: does not need management actions
a_0: no actions	0	0	1000
a_1: actions	200	1000-200=800	1000-200=800

2.3 Utility and certain equivalent

Assuming a risk neutral decision-maker, the values $v(x, a)$, $x \in \{x_0, x_1\}$, $a \in \{a_0, a_1\}$ are sufficient measure to describe the decision-maker's appreciation of different scenarios. However, since we are particularly interested in a risk averse decision-maker, these values need to be extended to utilities. We consider a strictly increasing utility function $u(v)$ that takes units of value as input and returns units of utility [von Neumann and Morgenstern, 1944]. The monetary value of a certain outcome is the same for all decision-makers, but each decision-maker accounts for the same value differently through the utility function.

The utility function varies based on the personal risk tolerance of each decision-maker. A risk neutral decision-maker should make decisions by maximizing the expected value and disregarding the variance. Thus, the utility function is linear. A risk-averse decision-maker prefers alternatives with low uncertainty compared to those with high uncertainty, even if the latter alternative has an equal or higher expected (monetary) value. For a risk averse decision-maker, the utility function

is concave. Some examples of a concave utility function used in the literature are an exponential function having a form of $a - \exp(-\gamma v)$, $\gamma > 0$ (Fig. 1, left), a power function having a form of v^α, $0 < \alpha < 1$ (Fig. 1, right) and a logarithmic function. The opposite of a risk-averse decision maker is a risk seeking decision-maker, who prefers alternatives with high uncertainty over those with low uncertainty, if the expected (monetary) value of the outcome of the high uncertainty alternative is higher. In this case, the utility function is convex. The optimal decision is invariant under a linear transformation of a utility function [Keeney and Raiffa, 1979].

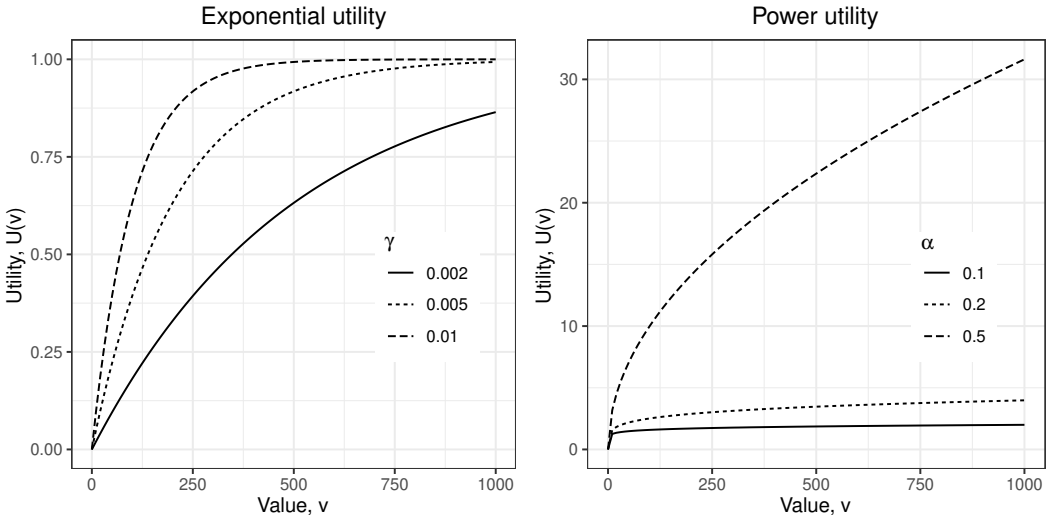


Figure 1: Examples of exponential and power utility functions for a risk averse decision-maker.

A decision-maker may find it difficult to directly specify the values of parameters γ and α that control the degree of risk aversion. Sometimes it is more approachable to specify the degree of risk aversion via a certain equivalent (CE, also known as a certainty equivalent), which is defined in the same units as the value function. CE means the lowest amount of guaranteed cash that one would accept instead of taking the risk of receiving a larger amount from the uncertain decision situation. To compare CE and other risk aversion measures (see Section 2.4), we define CE of a lottery and denote it as CE_0 to distinguish it from the CE of the decision-making situation between alternatives a. Suppose the decision-maker has an initial wealth w. Formally, the CE of a lottery is defined as

$$\mathrm{CE}_0 = u^{-1}\left(\mathbb{E}(u(v(x)+w))\right) - w = u^{-1}\left(\sum_x u(v(x)+w)p(x)\right) - w, \tag{1}$$

where u^{-1} is the inverse the utility function, and the prospects of a binary random variable x have values $v(x)$.

2.4 Risk aversion measures

There are several measures for the risk aversion expressed by the utility function. One of the most commonly used is the Arrow-Pratt measure of absolute risk aversion (ARA) [Arrow, 1965, Pratt, 1964]. It is defined as

$$r(v) = -\frac{u''(v)}{u'(v)}, \tag{2}$$

where u' and u'' are the first- and the second-order derivatives of the utility function, respectively, and v is the (monetary) value. The greater the value of $r(v)$, the larger the risk aversion. For a risk neutral decision-maker, the measure is zero because $u(v)$ is a linear function. The idea is to measure risk aversion using the second derivative of the utility function and to normalize it by the first derivative, which accounts for the magnitude of the utility function. If $r(v)$ is a constant over all v, it is referred to as a constant absolute risk aversion measure (CARA). A linear utility function (for a risk neutral decision-maker) and an exponential utility function (for a risk averse decision-maker) are the only utility functions to meet the CARA condition [Eidsvik et al., 2015]. As the first example of a utility of a risk averse decision-maker, we are using an exponential utility $u(v) = 1 - \exp(-\gamma v)$, where parameter γ controls the risk aversion. Then, the measure becomes constant with respect to v: $r(v) = \gamma$. The parameter γ is alternatively referred to as the risk aversion coefficient. If, for instance, the value is expressed in the units of EUR, then the unit of γ is EUR^{-1}. The risk aversion coefficient can also be parameterised as a risk tolerance $1/\gamma$.

If the decision-maker's risk attitude varies over v, i.e. the decision-maker changes from risk averse to risk seeking or vice versa, one should use a relative risk measure. The Arrow–Pratt measure of relative risk aversion (RRA) is defined as

$$R(v) = vr(v) = -\frac{vu''(v)}{u'(v)}. \tag{3}$$

Like for absolute risk aversion, the corresponding term constant relative risk aversion (CRRA) is used. As a specific example of that and the second example of a utility of a risk averse decision-maker, we use a power utility function $u(v) = v^\alpha$, where $0 < \alpha < 1$ is a risk aversion parameter. It follows $R(v) = 1 - \alpha$.

2.5 Value of information in lake management

Additional data could assist the decision-makers in choosing among the alternatives in lake management situation. VOI is a concept of the decision theory that is useful in this context [Howard and Abbas, 2015, Eidsvik et al., 2015]. It assesses the value of additional information to solve the problem, before it is gathered. The goal is then to compare VOI with the actual cost of the decision of decide whether the additional information is worth gathering or not. If the VOI exceeds the cost, the decision-maker should commit to gather the information.

According to the principles of the decision theory, the decision-maker always chooses the alternative that maximizes the expected utility. If the decision-maker with initial wealth w chooses alternative a, then they gain the value $v(x, a)$ and their wealth becomes $v(x, a) + w$. The maximum expected utility between the alternatives is the prior value

$$\max_{a \in A} \left\{ \sum_x u(v(x, a) + w)p(x) \right\}.$$

This can be also understood as the certain equivalent of the decision situation without information. Next, we calculate the expected utility when the information is available. Assume that the decision-maker pays the price $v^\star$ to get the information but will not know how the uncertainty will resolve before the decision is made. The expected utility is then the posterior value

$$\sum_x \max_{a \in A} \{u(v(x, a) + w - v^\star)\}p(x).$$

It can be also understood as the certain equivalent of the situation where information is available for free. The value of (perfect) information VOI(x) is the price $v^\star$ at which the above expected utilities are equal:

$$\sum_x \max_{a \in A} \{u(v(x, a) + w - v^\star)\}p(x) = \max_{a \in A} \left\{ \sum_x u(v(x, a) + w)p(x) \right\}. \tag{4}$$

VOI can be computed by iteratively varying $v^\star$ until the equation is satisfied. A unique solution always exists because u is a strictly increasing function.

Equation (4) is a general definition, but it can also be presented in an easier form by setting assumptions. When assuming a linear or an exponential utility function, the expected utility becames independent of the decision-maker's initial wealth w. This is because these utility functions fulfill the CARA assumption [Howard and Abbas, 2015]. Under CARA, the value of (perfect) information

can be expressed as

$$\mathrm{VOI}(x) = u^{-1}\left(\sum_x \max_{a\in A}\{u(v(x,a))\}\,p(x)\right) - u^{-1}\left(\max_{a\in A}\left\{\sum_x u(v(x,a))p(x)\right\}\right), \tag{5}$$

which is independent of the initial wealth w. Denoting $\mathrm{VOI}(x)$ implies that the VOI is calculated for a perfect information x.

In Equations (4) and (5), it is assumed that the perfect knowledge about the state of x is obtained by gathering the data. However, in many cases that is not possible, but the decision-makers need to settle for imperfect data that indicates the state of x. The value of imperfect information is the price $v^\star$, so that

$$\int_y \max_{a\in A}\left\{\sum_x u(v(x,a)+w-v^\star)p(x|y)\right\}p(y)dy = \max_{a\in A}\left\{\sum_x u(v(x,a)+w)p(x)\right\}, \tag{6}$$

where $p(x|y)$ is the posterior distribution of x given the uncertainty y. Again, it can be solved by iteratively varying the price until equation is satisfied. If again assuming a linear utility function (for a risk neutral decision-maker) or an exponential utility function (for a risk averse decision-maker), VOI can be expressed as

$$\mathrm{VOI}(y) = u^{-1}\left(\int_y \max_{a\in A}\left\{\sum_x u(v(x,a))p(x|y)\right\}p(y)dy\right) - u^{-1}\left(\max_{a\in A}\left\{\sum_x u(v(x,a))p(x)\right\}\right). \tag{7}$$

Denoting $\mathrm{VOI}(y)$ implies that it is calculated for an imperfect information. Note that the Equation (7) is again independent of the initial wealth w.

Because of the integration of the continuous probability distribution, we are unable to obtain $\mathrm{VOI}(y)$ in Equation (6) and Equation (7) directly. To obtain an approximation of the value of imperfect information, we utilize a Monte Carlo type of approach using empirical data, as done by Koski et al. [2020]. We approximate Equation (6) by finding $\mathrm{VOI}(y) = v^\star$ that fulfills the condition

$$\frac{1}{n}\sum_{i=1}^{n}\max_{a\in A}\left\{\sum_x u(v(x,a)+w-v^\star)\hat{p}(x|y)\right\} = \max_{a\in A}\left\{\sum_x u(v(x,a)+w)p(x)\right\}, \tag{8}$$

where y_i are the n values sampled randomly from the distribution $\hat{p}(y_i|x)$ fitted to the empirical data of chlorophyll-a concentration. The posterior distribution of the ecological status x given by the data y_i is obtained by Bayes' rule. When using a linear or an exponential utility function, the approximation of $\mathrm{VOI}(y)$ (Eq. (7)) can be again expressed simpler by

$$\widehat{\mathrm{VOI}}(y) = u^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\max_{a\in A}\left\{\sum_x u(v(x,a))\hat{p}(x|y_i)\right\}\right) - u^{-1}\left(\max_{a\in A}\left\{\sum_x u(v(x,a))p(x)\right\}\right). \tag{9}$$

For details on the approximation, see Koski et al. [2020].

3 Application to lake management

This section demonstrates the effect of the degree of risk aversion on the value of information in a lake management example. We calculated the value of perfect information as well as imperfect information and used an exponential utility function $u(v) = 1 - \exp(-\gamma v)$ and a power utility function $u(v) = v^{\alpha}$ for a risk averse decision-maker with different values of the parameters γ and α. The risk aversion measures used are then $r(v) = \gamma$ for the exponential utility (the risk aversion coefficient) and $R(v) = 1 - \alpha$ for the power utility. To evaluate the sensitivity of VOI, we varied the prior probability $p(x = x_1)$ to be 0.1, 0.5 and 0.9, and the cost of the lake management actions to be not only EUR 200 but also EUR 900. This means that after implementing the management actions, the resulting value of the lake equals EUR 800 or EUR 100. All coding was implemented in `R` [R Core Team, 2023].

3.1 Relationship of certain equivalent and risk aversion function

To begin with, we studied the relationship of CE and the risk aversion functions in the decision situation of lake management. For this purpose, we assumed that a lake is either in the need of management actions or not, but there is no possibility of management actions ($a = a_0$). For instance, when assuming $p(x = x_1) = 0.5$ and the values for lake in poor and good condition to be EUR 0 and EUR 1000, respectively, the CE of this situation for a risk neutral decision-maker can be calculated as $\text{CE}_0 = \sum_{x \in \{0,1\}} (u(v(x))p(x)) = \text{EUR } 0 \cdot 0.5 + \text{EUR } 1000 \cdot 0.5 = \text{EUR } 500$. Thus, the risk neutral decision-maker would be willing to sell the situation at EUR 500 or a higher price. Similar calculations can be presented when assuming $p(x = x_1) = 0.1$ and $p(x = x_1) = 0.9$.

In the case of exponential utility, the risk aversion coefficient γ equals zero for a risk neutral decision-maker and increases as the degree of risk aversion increases. The top row of Figure 2 shows that the more a decision-maker avoids risk, the lower price they are willing to sell the risky situation and receive price CE for certain. When CE has shrunk to half of the risk neutral decision-maker's CE, then γ equals approximately 0.0025. In the case of power utility function, the relative risk aversion function $R(v) = 1 - \alpha$ equals zero for a risk neutral decision-maker and increases to one as the degree of risk aversion increases. To compare with the exponential utility, the decrease of CE seems to be more linear than exponential as the risk aversion increases (Fig. 2, bottom row).

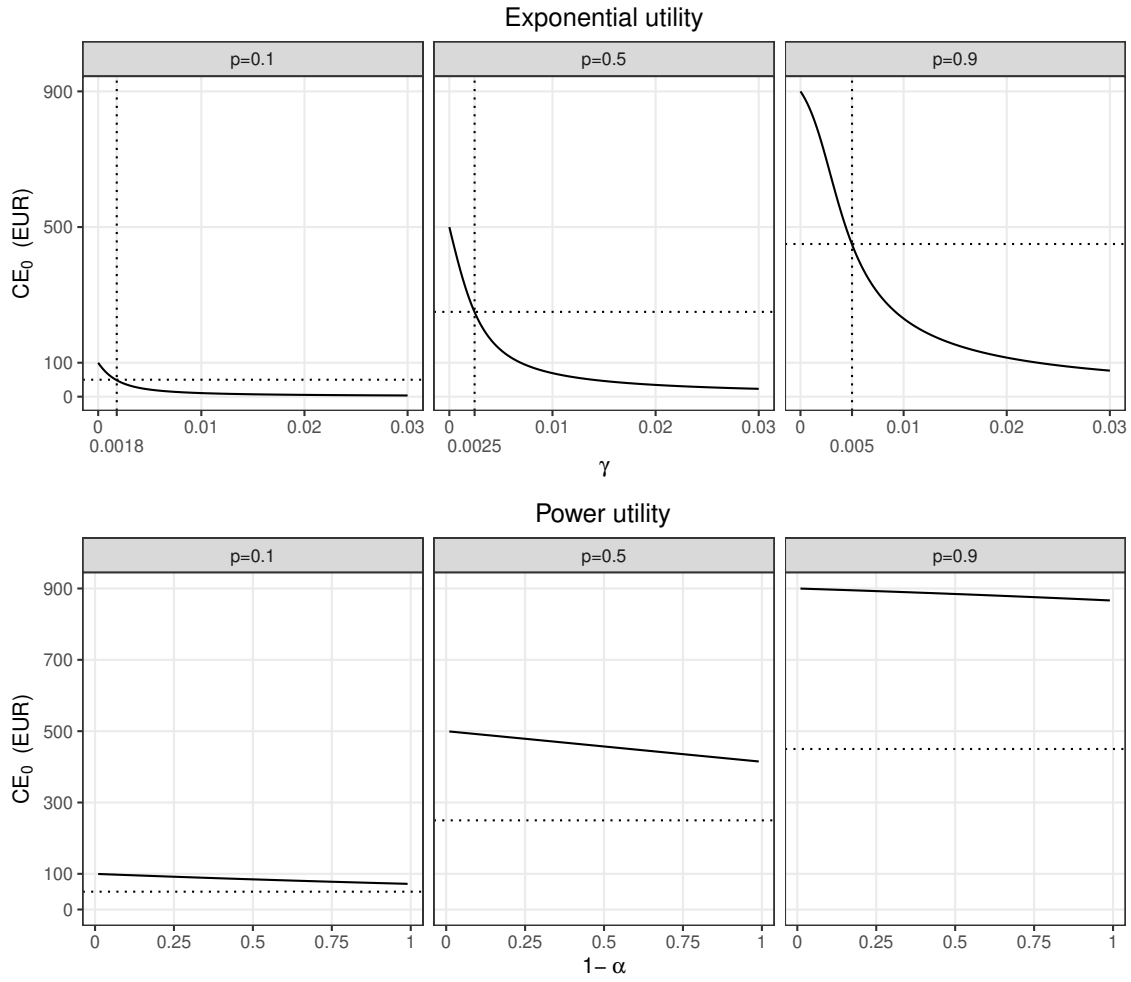


Figure 2: The relationship of CE and the risk aversion coefficient γ in lake management decision situation, when using the exponential utility function (top row). The relationship of CE and the relative risk aversion function $R(v) = 1 - \alpha$, when using power utility function (bottom row). Here, it is assumed the initial wealth EUR 1000 per hectare for power utility, $p(x = x_1) = \{0.1, 0.5, 0.9\}$ and the values for the lake in poor and good condition to be EUR 0 and EUR 1000 per hectare, respectively.

3.2 Relationship of VOI and risk aversion

Figure 3 shows the value of perfect and imperfect information as a function of risk aversion function and the CE when the decision-maker has an exponential utility function. The risk aversion coefficient

γ (primary horizontal axis) is varied from 0 to 0.02, where 0 responds a risk neutral decision-maker. The CE (secondary horizontal axis) is varied according to the value of γ. It seems that the relationship between VOI and the degree of risk aversion is monotonic in every decision region, where the decision based on prior knowledge remains similar. However, VOI may either decrease or increase as the risk aversion increases, depending on the prior and the cost. The value of perfect information, VOI(x), is the absolute maximum value worth paying from additional information, and value of imperfect information, VOI(y), is always less than that. However, VOI(y) seems to depend on VOI(x) so that the maximum value for both is reached at the same degree of risk aversion.

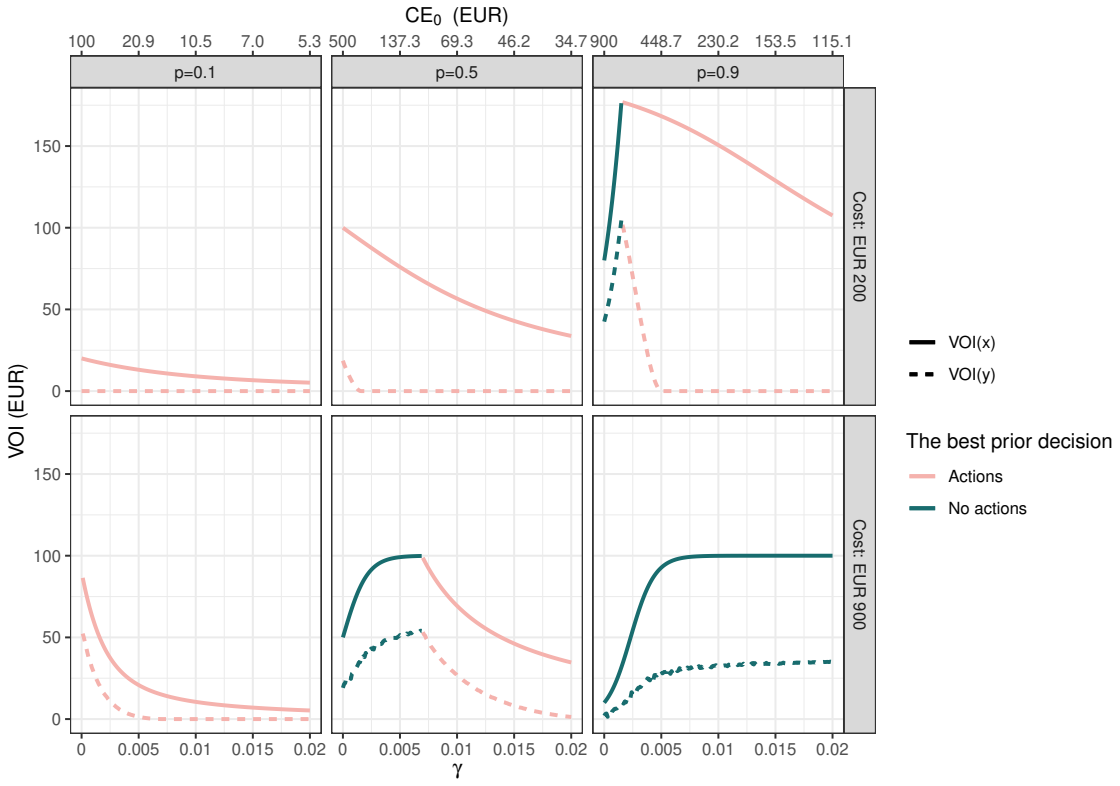


Figure 3: Value of perfect and imperfect information as a function of the risk aversion coefficient γ (primary horizontal axis) and CE_0 (secondary horizontal axis), when using an exponential utility function. VOI is calculated using different values of the prior probability $p(x = x_1)$ and the cost of management action. The decision about lake managements (color coding) between two alternatives (actions/no actions) is based on the prior information only.

First, we fixed the cost of the management actions to be EUR 200 per hectare and varied the prior (Fig. 3, top row). When a priori we are quite certain that the lake needs management actions ($p(x = x_1) = 0.1$), VOI is low and, based on the prior information only, the best decision that maximizes the expected utility is to implement the management actions. Here, VOI(x) is decreasing as the degree of risk aversion is decreasing and VOI(y) is zero regardless of the degree of risk aversion. Thus, regardless of the risk preference, it is profitable to implement the actions without additional monitoring. When a priori we are uncertain about the status of the lake ($p(x = x_1) = 0.5$), the same pattern can be seen, but VOI is generally somewhat higher. If in turn we are a priori more certain that the lake is in target status ($p(x = x_1) = 0.9$), the best decision based on the prior information is not to implement the management actions for a risk neutral and for a little risk averse decision-maker. If the risk aversion increases enough ($\gamma = 0.0016$ and CE = EUR 792) the best decision changes to implement the actions. This seems reasonable: a risk averse decision-maker wants to assure a good lake status and avoid value losses if the prior assumption is wrong. Here, VOI(x) is first increasing as the degree of risk aversion is increasing but then decreases after $\gamma = 0.0016$. VOI(y) decreases faster to zero even VOI(x) remains positive. It seems that if the decision-maker is risk averse enough, the imperfect information is not enough to change the prior information sufficiently.

Second, we also tried the cost to be EUR 900 per hectare (Fig. 3, bottom row). If a priori we think that the lake needs management actions ($p(x = x_1) = 0.1$), the best decision based on the prior information is to implement the actions. VOI is higher for a risk neutral decision-maker and decreases fast when the degree of risk aversion increases. If a priori the status of the lake is uncertain ($p(x = x_1) = 0.5$), the best decision based on the prior information is not to implement the management actions for a risk neutral and for a little risk averse decision-maker. If the risk aversion increases enough ($\gamma = 0.007$ and CE = EUR 99) the best decision changes to implement the actions. Here, VOI(x) is first increasing as the degree of risk aversion is increasing but then decreases fast after $\gamma = 0.007$. VOI(y) is approximately EUR 50 lower than VOI(x) at any level of risk aversion. If a priori it is quite certain that the lake is in target status ($p(x = x_1) = 0.9$), based on the prior information, it is more profitable to not to implement the management actions for both a risk neutral and for a risk averse decision-maker. VOI(x) is low for a risk neutral decision-maker but increases fast to be EUR 100 when the degree of risk aversion increases. VOI(y) increases as well but not as fast.

In addition, we studied the relationship of VOI and risk aversion when the decision-maker has a power utility function. Figure 4 shows the value of perfect and imperfect information as a function of risk aversion function and CE. The parameter α is varied from 0.01 to 1, that is, the risk aversion function $R(v) = 1 - \alpha$ is varied from 0 to 0.99 (primary horizontal axis), $R(v) = 0$ responding a risk neutral decision-maker. The CE is varied according to the value of the risk aversion function (secondary horizontal axis). Now, the VOI is depending on the decision-maker's initial wealth and it is set to EUR 1000 per hectare. This is selected because it is the value of a lake being in high or good ecological status. Compared with the results of a decision-maker with an exponential utility function, the overall patterns seem quite similar. However, there are no extreme value points for VOI(x) and VOI(y) when the risk aversion increases.

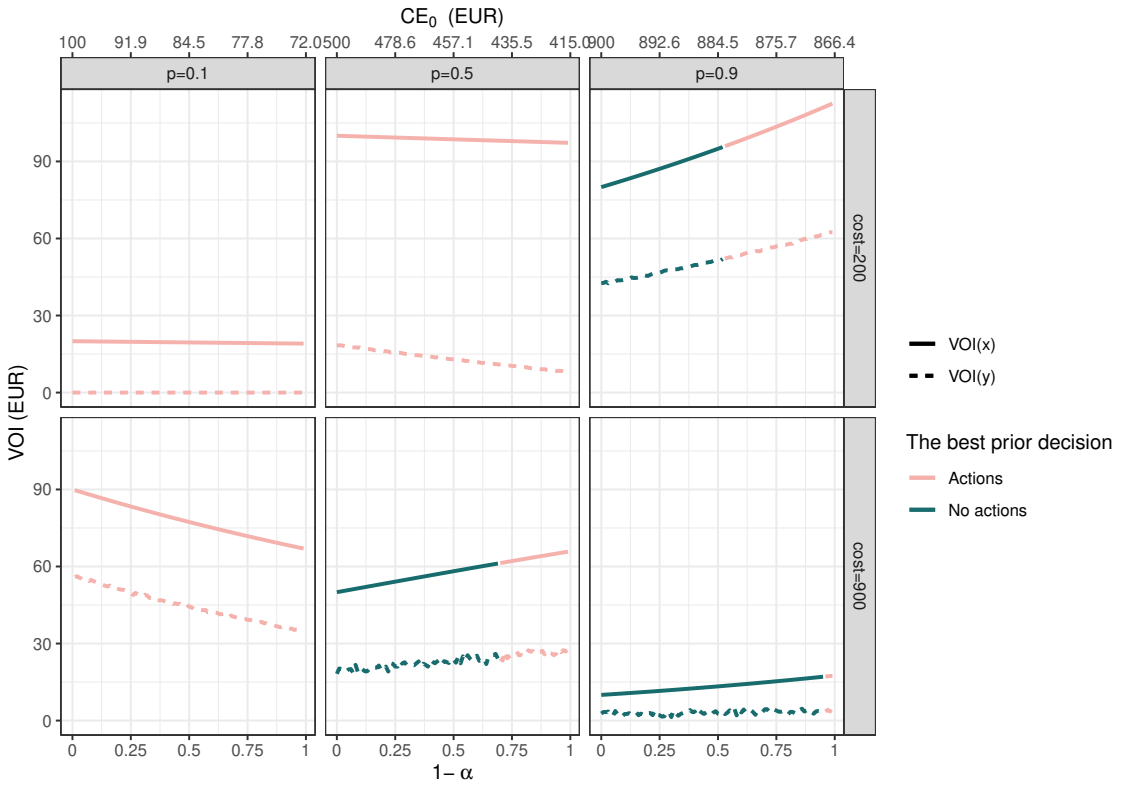


Figure 4: Value of perfect and imperfect information as a function of the relative risk aversion function $R(v) = 1 - \alpha$ (primary horizontal axis) and CE_0 (secondary horizontal axis), when using a power utility function. VOI is calculated using different values of the prior probability $p(x = x_1)$ and the cost of management action. The decision about lake managements (color coding) between two alternatives (actions/no actions) is based on the prior information only.

3.3 Effect of risk aversion on the prior decision

Yet another interesting question is how the risk aversion affects the prior decision. Figures 3 and 4 show that the best decision that maximizes the expected utility, based on only prior knowledge between implementing management actions or not to implement them, depends on the degree of risk aversion. Three cases can be recognized: (1) the best prior decision is to implement the management actions regardless of the degree of risk, (2) the best prior decision is not to implement them regardless of the degree of risk or (3) the best prior decision depends on the degree of risk aversion. In addition, an interesting detail is that generally, while the best prior decision is to implement the actions, the VOI is decreasing as the risk aversion is increasing, and while the prior based decision is to not to implement the actions, the VOI is increasing as the risk aversion is increasing. This is true at least in the case of an exponential utility function but not always in the case of a power utility function. Intuitively, this makes sense since when the management actions are not implemented, it is worth paying more for the additional information, the more the risk aversion increases. Respectively, when the management actions are implemented, it is worth paying less for the information, the more the risk aversion increases in order to save money.

We studied the impact of the prior probability $p(x = x_1)$, the cost of the lake management action, and the risk aversion (an exponential utility varying the risk aversion coefficient γ between 0 and 0.02) on the best prior decision. Figure 5 shows that for the large values of the prior probability and the cost, the decision based on only prior knowledge is always not to implement the management actions, regardless of the risk aversion. If, on the other hand, the prior is large and the cost is small, or the prior probability is small and the cost is large, or if both are small, then the best prior decision is always to implement the management actions, regardless of the risk aversion. Between these two areas there are values for the prior probability and the cost where the degree of risk aversion is important for decision-making. The larger the values of the prior probability and the cost are, the larger the degree of risk aversion seems to be with which the best prior decision changes (from no actions to actions). In general, the more risk averse the decision-maker is, more likely they are to choose to implement the management actions, depending on the prior probability and the cost of the management actions. Note that the final decision between the two alternatives (actions/no actions) depends on whether the decision-maker is acquiring the additional data or making the decision based on the prior knowledge only.

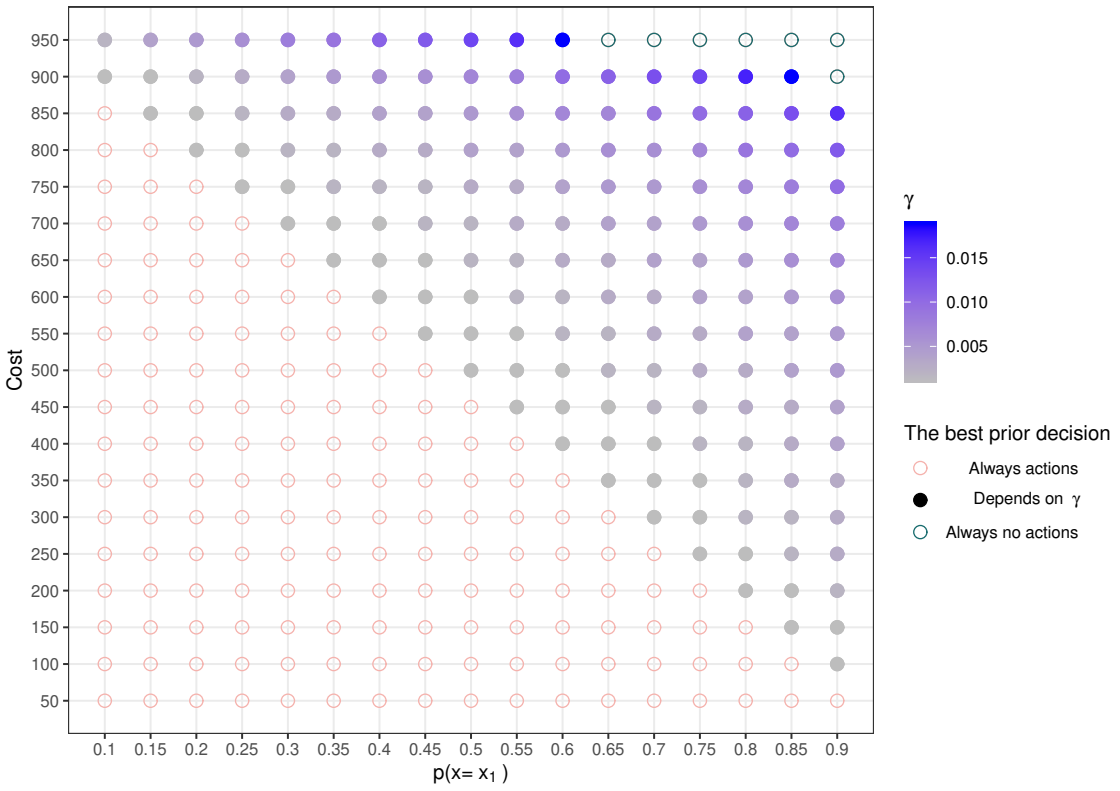


Figure 5: The best prior decision as a function of the prior probability $p(x = x_1)$, the cost of management action and the degree of risk aversion. The best decision is to implement the management actions regardless of the degree of risk (hollow red dots), or not to implement them regardless of the degree of risk (hollow blue dots) or then the best decision is depending on the degree of risk aversion (solid coloured dots). The color code for γ indicates the value where the best prior decision changes from no actions to actions.

3.4 Comparison of VOI and the monitoring costs

To complete our VOI analysis, we also compared the obtained VOI with actual monitoring costs in the cases of a risk neutral and a risk averse decision-makers. Based on the information from the Finnish Environmental Institute, one sample of chlorophyll-a costs EUR 138 [Koski et al., 2020]. In order to be able to map the status of the lake, multiple samples must be taken annually and possibly from different locations.

Assume a lake with an area of 1000 hectares and assume that we are a priori uncertain whether

the lake needs management actions or not ($p(x = x_1) = 0.5$). The cost of the management actions is assumed to be EUR 900 (per hectare). A risk neutral decision-maker would value the additional information that gives a certain knowledge of the status of the lake EUR $50 \cdot 1000$ hectare $=$ EUR 50000. On the other hand, a risk averse decision-maker with an exponential utility and a risk aversion coefficient $\gamma = 0.005$ would value the same knowledge EUR $100 \cdot 1000$ hectare $=$ EUR 100000. However, the value of perfect information calculated here is only a theoretical upper limit for the value of information because there are no real data that would give certain (perfect) information about the status of the lake.

A more realistic scenario is to assume that instead of certain knowledge, we obtain an imperfect data about the status. A risk neutral decision-maker would value the imperfect additional information about the status of the lake EUR $18 \cdot 1000$ hectare $=$ EUR 18000. A risk averse decision-maker with an exponential utility and a risk aversion coefficient $\gamma = 0.005$ would value the same knowledge EUR $51 \cdot 1000$ hectare $=$ EUR 51000. With the data cost mentioned above, for a risk neutral decision-maker it would be profitable to gather 130 chlorophyll samples spread over observation years and locations whereas a risk averse decision-maker could gather up to 369 samples. Typically, monitoring is performed at observation sites every year or every few years. Annually during a growing season, samples are taken several times, usually 2–12 times, depending on the need [Aroviita et al., 2019]. For a six-year period, this means 12–72 samples.

4 Conclusion

We have studied the relationship between the risk aversion and the value of perfect and imperfect information in the context of a lake management application. We calculated the value of perfect information theoretically and estimated the value of imperfect information based on the real monitoring data. The examples in earlier studies on risk aversion have been mostly numerical without a real life data.

A risk averse decision-maker's VOI may be lower or higher than a risk neutral decision-maker's VOI, depending on the prior probability for a lake to need management actions and the cost of the actions. Generally, it seems that if a priori it is quite sure that a lake needs management actions, a risk neutral decision-maker values the additional data higher than a risk averse decision-maker.

In contrast, if a priori it is quite sure that a lake is not in the need of management actions, a risk averse decision-maker values the additional data higher than a risk neutral decision-maker. These conclusions hold for both an exponential utility function and a power utility function.

In VOI analysis, it is often assumed risk neutrality for the sake of simplicity. Our examples, as well as earlier literature, provide evidence that VOI may strongly depend on the degree of risk aversion. These results imply that decision-makers could make better decisions if they quantify their risk aversion. This is the main message of the study for practical applications.

The decision-maker's utility function can be determined using elicitation methods, one of which is finding out CE. Usually the elicitation consists of a set of questions that investigate decision-maker's risk aversion, and the answers are used to estimate the utility function. A review of different elicitation schemes is provided by Farquhar [1984]. Kirkwood [2004] also showed, that using a simple form of utility function, which requires very little utility elicitation of the decision-maker, will be sufficient for an accurate decision analysis. Specifically, an exponential utility function is usually sufficient to describe the decision-maker's risk preference.

The exponential utility function fits the case of environmental monitoring particularly well also because it makes the VOI calculations independent of the decision-maker's initial wealth. In general, the determination of monetary values makes the application of VOI analysis in environmental monitoring difficult. Thus, determining the wealth of the decision-maker is also difficult in this context.

From the environmental management point of view, an interesting question of the VOI analysis is the amount of VOI and comparison to the actual costs of the data gathering. An assumption of the decision-maker's risk aversion is often a more realistic scenario and based on Figures 3 and 4, VOI for a risk averse decision-maker can be even higher than for a risk neutral decision-maker. Compared with the results of the study of Koski et al. [2020], this study gives even more evidence that the lake monitoring is cost-effective.

The study leaves room for future research. In our VOI analysis, we are only considering monetary value as a criterion to be optimized. However, in the lake management situation, we could have been interested in optimizing for both monetary and biodiversity criteria. In such a multiple criteria context, VOI analysis involves a trade-off between competing interests [Eyvindson et al., 2019].

Funding

V.K. acknowledges support from the Kone Foundation.

Competing Interests

The authors have no conflict of interest to declare.

Author Contributions

Both authors contributed to the study conception. Data preparation and analysis was performed by V.K. The first draft of the manuscript was written by V.K., and V.K. and J.K. both reviewed, read and approved the final manuscript.

References

A. E. Abbas, N. O. Bakır, G.-A. Klutke, and Z. Sun. Effects of risk aversion on the value of information in two-action decision problems. *Decision Analysis*, 10(3):257–275, 2013. doi: 10.1287/deca.2013.0275. URL `https://doi.org/10.1287/deca.2013.0275`.

H. Ahtiainen. *Järven tilan parantamisen hyödyt. Esimerkkinä Hiidenvesi.* Finnish Environment Institute (SYKE), Helsinki, 2008. ISBN 978-952-11-3284-1. (In Finnish.) Available online at: `https://helda.helsinki.fi/handle/10138/38353` (Accessed May 22th, 2019).

J. Aroviita, S. Mitikka, and S. Vienonen. *Pintavesien tilan luokittelu ja arviointiperusteet vesienhoidon kolmannella kaudella.* Finnish Environment Institute (SYKE), Helsinki, 2019. ISBN 978-952-11-5074-6. (In Finnish.) Available online at: `https://helda.helsinki.fi/handle/10138/306745` (Accessed February 17th, 2020).

K. Arrow. *Aspects of the theory of risk-bearing.* Yrjö Jahnssonin Säätiö, 1965. URL `https://books.google.fi/books?id=hnNEAAAAIAAJ`. Reprinted in: Essays in the Theory of Risk Bearing, Markham Publ. Co., Chicago, 1971, 90–109.

J. E. Bickel. The relationship between perfect and imperfect information in a two-action risk-sensitive problem. *Decision Analysis*, 5(3):116–128, 2008. doi: 10.1287/deca.1080.0118. URL https://doi.org/10.1287/deca.1080.0118.

G. B. Davies and S. E. Satchell. The behavioural components of risk aversion. *Journal of Mathematical Psychology*, 51(1):1–13, 2007. ISSN 0022-2496. doi: https://doi.org/10.1016/j.jmp.2006.10.003. URL https://www.sciencedirect.com/science/article/pii/S0022249606001222.

A. de Palma, R. Lindsey, and N. Picard. Risk aversion, the value of information, and traffic equilibrium. *Transportation Science*, 46(1):1–26, 2012. doi: 10.1287/trsc.1110.0357. URL https://doi.org/10.1287/trsc.1110.0357.

P. Delquié. The value of information and intensity of preference. *Decision Analysis*, 5(3):129–139, 2008. doi: 10.1287/deca.1080.0116. URL https://doi.org/10.1287/deca.1080.0116.

L. Eeckhoudt and P. Godfroid. Risk aversion and the value of information. *The Journal of Economic Education*, 31(4):382–388, 2000. doi: 10.1080/00220480009596456. URL https://www.tandfonline.com/doi/abs/10.1080/00220480009596456.

J. Eidsvik, T. Mukerji, and D. Bhattacharjya. *Value of Information in the Earth Sciences: Integrating Spatial Modeling and Decision Analysis*. Cambridge University Press, 2015.

European Parliament. Directive 2000/60/EC, of the European Parliament and Council of 23 October 2000 establishing a framework for Community action in the field of water policy, 2000. URL http://eur-lex.europa.eu/resource.html?uri=cellar:5c835afb-2ec6-4577-bdf8-756d3d694eeb.0004.02/DOC_1&format=PDF.

K. Eyvindson, J. Hakanen, M. Mönkkönen, A. Juutinen, and J. Karvanen. Value of information in multiple criteria decision making: an application to forest conservation. *Stochastic Environmental Research and Risk Assessment*, (33):2007–2018, 2019. ISSN 1436-3240. doi: https://doi.org/10.1007/s00477-019-01745-4.

P. H. Farquhar. State of the art – utility assessment methods. *Management Science*, 30(11):1283–1300, 1984. doi: 10.1287/mnsc.30.11.1283. URL https://doi.org/10.1287/mnsc.30.11.1283.

X. Freixas and R. Kihlström. Risk aversion and information demand, in: M. Boyer and R.E. Kihlström (ed). *Bayesian Models in Economic Theory*, pages 93–104, 1984.

R. W. Hilton. The determinants of information value: Synthesizing some general results. *Management Science*, 27(1):57–64, 1981. doi: 10.1287/mnsc.27.1.57. URL https://doi.org/10.1287/mnsc.27.1.57.

R. A. Howard and A. E. Abbas. *Foundations of Decision Analysis*. Pearson Higher Ed, 2015.

R. Keeney and H. Raiffa. *Decisions with Multiple Objectives: Preferences and Value Trade-Offs*. Wiley, 1979. Reprinted Cambridge University Press, Cambridge (1993).

C. W. Kirkwood. Approximating risk aversion in decision analysis applications. *Decision Analysis*, 1(1):51–67, 2004. doi: 10.1287/deca.1030.0007. URL https://doi.org/10.1287/deca.1030.0007.

V. Koski and J. Eidsvik. Sampling design methods for making improved lake management decisions. *Environmetrics*, n/a(n/a):e2842, 2024. doi: https://doi.org/10.1002/env.2842. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/env.2842.

V. Koski, N. Kotamäki, H. Hämäläinen, K. Meissner, J. Karvanen, and S. Kärkkäinen. The value of perfect and imperfect information in lake monitoring and management. *Science of The Total Environment*, 726:138396, 2020. ISSN 0048-9697. doi: https://doi.org/10.1016/j.scitotenv.2020.138396. URL https://www.sciencedirect.com/science/article/pii/S0048969720319094.

A. Mehrez. The effect of risk aversion on the expected value of perfect information. *Operations Research*, 33(2):455–458, 1985. ISSN 0030364X, 15265463. URL http://www.jstor.org/stable/170756.

R. Nadiminti, T. Mukhopadhyay, and C. H. Kriebel. Risk aversion and the value of information. *Decision Support Systems*, 16(3):241–254, 1996. ISSN 0167-9236. doi: https://doi.org/10.1016/0167-9236(95)00023-2. URL https://www.sciencedirect.com/science/article/pii/0167923695000232.

J. W. Pratt. Risk aversion in the small and in the large. *Econometrica*, 32(1/2):122–136, 1964. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1913738.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL https://www.R-project.org/.

Z. Sun and A. Abbas. On the sensitivity of the value of information to risk aversion in two-action decision problems. *Environment Systems and Decisions*, 34:24–37, 03 2014. doi: 10.1007/s10669-013-9477-y.

J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, New Jersey, 1944. ISBN 9780691130613. URL http://www.jstor.org/stable/j.ctt1r2gkx.

M. Willinger. Risk aversion and the value of information. *The Journal of Risk and Insurance*, 56 (1):104–112, 1989. ISSN 00224367, 15396975. URL http://www.jstor.org/stable/253017.