FACULTY OF SCIENCE AND TECHNOLOGY

MASTER THESIS

Study programme / specialisation:
Industrial Asset Management

The spring semester, 2022
Open

Author: **Reihaneh Shahali**

Supervisor(s):
Reidar B. Bratvold, Aojie Hong

Thesis title:
Machine learning methods for assessing value-of-information

Credits (ECTS): 30

Keywords:
Machine learning
Value of Information
Value of Flexibility
Simulation-Regression
Monte Carlo Simulation
Decision analysis

Pages: ……71……
+ supplemental material/other: 0

Stavanger, 15th June 2022

Acknowledgements

This thesis concludes my master's study in Industrial Asset Management at the University of Stavanger.

I would like to express my gratitude to those who have helped and supported me during this thesis project. First and foremost, I would like to express my deepest appreciation to my supervisors, Professor Reidar B. Bratvold and Professor Aojie Hong, for their impeccable advice, vast knowledge, kindness, patience, and encouragement during this project and my master's study. It has been a pleasure and wonderful experience to work together with the experts in the field of decision analysis and programming.

I could not undertake this journey without my parents and family for providing me the best educations and giving me endless emotional support throughout my studies and life. I would also give my special thanks to Mohammadmahdi Ataei and my friends, I am truly grateful for their help and encouragement throughout the entire process.

Abstract

One of the most useful features of decision analysis is its ability to distinguish between constructive and wasteful information gathering. Value-of-information (VOI) and sequential information gathering (Value-of-Flexibility, VOF) analyses evaluate the benefits of collecting additional information before making a decision.

Traditionally, VOI has been assessed by constructing a decision tree or influence diagram model where a Bayesian framework has been used to update probabilities given new information. In this research, we evaluate the use of machine learning (ML) methods such as Ordinary-Least-Square (OLS), Random Forest (RF), K-Nearest Neighbor (KNN), Support Vector Regression (SVR), and Extreme Gradient Boost (XGB) for VOI calculations.

In this study, VOI will be estimated using a simulation-regression approach. In the simulation-regression approach, VOI is computed by simulating the model parameters, the data, and prospect values, then regressing the prospect values on the data (Eidsvik, et al., 2015, Eidsvik, Dutta, et al., 2017, Dutta, et al., 2019). Simulation-regression approach is considered to be one solution to overcome the computational issue by constructing efficient approximations for the VOI.

In addition, VOI and Value-of-Flexibility (VOF) analyses are implemented in a case study of estimating the CO_2 storage capacity of Utsira formation located in the North Sea using the simulation-regression approach.

Table of Contents

Acknowledgements ii
Abstract iii
List of Figures v
List of Tables vii
Nomenclature viii
Chapter 1 - Introduction 1
Chapter 2 –VOI Analysis and Simulation-Regression Approach 3
2.1 Monte Carlo Simulation 3
2.2 Value of Information 4
2.2.1 Prior Value 6
2.2.2 Posterior Value 7
2.3 Value Discretization 12
2.4 Simulation-Regression Approach 13
2.5 Regression methods 15
2.5.1 Ordinary Least Squares Linear Regression (OLS) 16
2.5.2 *K*-Nearest Neighbors (KNN) 17
2.5.3 Random Forest (RF) 17
2.5.4 Extreme Gradient Boosting (XGB) 18
2.5.5 Support Vector Regression (SVR) 18
2.5.6 Piecewise regression 19
2.6 Cross-Validation 19
2.6.1 Error Metrics 20
2.6.2 Comparing Different Regression Methods 20
2.7 Value of Flexibility (VOF) 33
2.8 Conclusions 38
Chapter 3 – Case Study at Utsira Formation 39
3.1 Utsira Formation Reservoir Model 39
3.2 Decision Frame 41
3.2.1 $CO2$ Storage Capacity Estimation and Uncertainties 41
3.2.2 Decision Frame and Influence Diagram 42
3.3 Workflow 43
3.3.1 Generate samples 44
3.3.2 Feature Selection 45
3.3.4 VOII Analysis 46
3.3.5 VOF Analysis 48
3.4 Summary 55
Chapter 4 - Conclusions and Recommendations 56
References 57
Appendix 1 – Cross-Validation on Hyperparameters 60
Appendix 2 – Sensitivity analysis on SVR and OLS performance using Piecewise Regression 61
Appendix 3 – $R2$ scores of Utsira case study 63

List of Figures

Figure 2.1 - Schematic of Monte Carlo simulation procedure (Bratvold and Begg, 2010) 3
Figure 2.2 - The sensitivity analysis on effect of test accuracy on VOI for $CO2$ storage investment case 12
Figure 2.3 - Histogram of VOII with 100 iterations 15
Figure 2.4 - Least Square method 16
Figure 2.5 - Piecewise regression with linear regression (Kalvelagen, E, 2018) 19
Figure 2.6 - Different regression methods results on a linear function with thickness as the only uncertainty 21
Figure 2.7 - The sensitivity analysis on number of Monte Carlo samples vs linear regression accuracy on a linear function. 23
Figure 2.8 - Different regression methods results on a non-linear function with pressure as the only uncertainty 24
Figure 2.9 - Piecewise OLS regression with 2 and 10 splits 25
Figure 2.10 - Cross-validation on hyperparameter maximum leaf nodes for RF regression with pressure as an uncertainty and feature 26
Figure 2.11 - The sensitivity analysis on effect of hyperparameter maximum leaf nodes in RF regression results 27
Figure 2.12 - The sensitivity analysis on hyperparameter regularization parameter in SVR with values of (a) 1, (b) 10, and (c) 40. 28
Figure 2.13 - Multivariate regression plots. (a) OLS, (b) Piecewise OLS with 4 splits 29
Figure 2.14 - Multivariate regression plots. (a) XGB, (b) KNN. 29
Figure 2.15 - Multivariate regression plots. (a) Piecewise SVR, (b) RF 30
Figure 2.16 - The sensitivity analysis on hyperparameter C for Piecewise SVR 32
Figure 2.17 - The sensitivity analysis on number of splits in Piecewise OLS Regression 32
Figure 2.18 - Sequential information gathering schematic 34
Figure 2.19 - Sequential information gathering decision tree 36
Figure 2.20 - Illustration of sequential information gathering scheme 37
Figure 3.1 - Utsira Formation figure with thickness range. (Source: MRST, SINTEF, 2016b, Lie, 2019) 39
Figure 3.2 - Influence Diagram of $CO2$ capacity estimation case study 42
Figure 3.3 - Workflow of the $CO2$ capacity estimation case study 44
Figure 3.4 - Tornado Diagram, features vs NPV 45
Figure 3.5 - The sensitivity analysis of noise vs $R2$ score 47
Figure 3.6 - The sensitivity analysis of VOII vs noise 47

Figure 3.7 - VOF analysis with sensitivity analysis on tests costs..... 49
Figure 3.8 - The sensitivity analysis on tests costs vs decisions..... 50
Figure 3.9 – Sequential information gathering with different regression method selection scenarios - cost $[2,2,2] million..... 52

List of Tables

Table 2.1 - VOI analysis of regression methods with uncertain thickness 22
Table 2.2 - VOI analysis of multivariate regressions with uncertain thickness and pressure..... 31
Table 3.1 - VOI analysis of multivariate regressions with uncertain thickness, pressure, and porosity..... 46
Table 3.2 - Senitivity analysis on tests costs vs best information gathering sequence 51
Table 3.3 - $R2$ scores of sequential information gathering - cost $[2,2,2] million 52
Table 3.4 - VOF analysis with different regression method selection scenarios- cost $[2,2,2] million 53
Table 3.5 - VOF analysis with different regression scenarios with cost set of $[14,2,2] million..... 54

Nomenclature

CAPEX	-	Capital Expenditure
CCS	-	Carbon Capturing and Storage
CDF	-	Cumulative Distribution Function
CSLF	-	Carbon Sequestration Leadership Forum
EVwI	-	Expected Value with Information
EVwoI	-	Expected Value without Information
KNN	-	K Nearest Neighbors Regression
MCS	-	Monte Carlo Simulation
NPD	-	Norwegian Petroleum Directorate
NPV	-	Net Present Value
OLS	-	Ordinary Least-Squares Linear Regression
PDF	-	Probability Distribution Function
PoV	-	Posterior Value
PV	-	Prior Value
RF	-	Random Forest Regression
SVR	-	Support Vector Regression
VOF	-	Value of Flexibility
VOI	-	Value of Information
VOII	-	Value of Imperfect Information
VOPI	-	Value of Perfect Information
XGB	-	Extreme gradient Boosting

Chapter 1 - Introduction

An essential element in every decision making situation which makes decision making difficult is uncertainty (Bratvold and Begg, 2010). As illustrated and discussed by Bratvold and Begg, (2010), and Eidsvik, et al. (2015), information can be gathered to further inform the decision. However, information sources usually have a cost that must be balanced with the value the information might provide in the given decision context. Value of information analysis (VOI) is being used for this issue. Most influence diagram applications use decision trees, or convert the model to a decision tree, and calculate the VOI. However, where the distributions of probabilities are continuous, discretization methods such as value discretization, CDF discretization, three-point shortcuts, etc. are needed to calculate VOI (Modeling for Decision Insights, lecture notes, 2021). In these cases, the decision tree has a lot of branches which makes VOI analysis suffer from the curse of dimensionality and does not scale well with the number of uncertainties (Modeling for Decision Insights, lecture notes, 2021). The simulation-regression method using Monte Carlo simulation is one method that has the potential to reduce the curse of dimensionality with an approximation of VOI relative to a decision tree calculated VOI. Therefore, simulation-regression approach is considered to be one solution for the computational issue in VOI calculation. In this study, the simulation-regression approach, and the impact of the regression method on VOI calculation are evaluated.

Carbon Capturing and Storage (CCS) is a much-discussed approach to reduce carbon emissions. The higher the CO_2 concentration, the higher the greenhouse effect, and the more rapid is the resulting climate change (Nordbotten, Celia, 2012). Many projects have been suggested and some have been initiated for CCS. Site selection for storage is one of the challenges in these projects. Utsira formation is a known saline formation suitable for CCS, which is located in the North Sea. However, CCS is very expensive, and any investment in CO_2 injecting and storage comes with significant uncertainties related to the reservoir, formation properties, and costs.

Gathering information can bring value to the decision making and help to make good decisions. However, this information can be costly and the decision maker should take it into account before choosing to have the information or not (Bratvold and Begg, 2010). In some cases, we can bring value to the decision making by investing in flexibility like gathering information sequentially and having multiple decision points instead of having only one decision to make for gathering

information (Begg, et al., 2002). For instance, in sequential information gathering cases, if the first decision is “have the first test”, we can observe the result of the test and then decide whether we should have a second or third test. In some cases, after choosing to have the first test, based on the observed result of it, the additional value that the further tests bring is not worth its cost, and the decision maker can say no to further tests.

This thesis consists of 4 chapters. Following this introduction, Chapter 2 explains the concepts of VOI and the simulation-regression approach, studies and evaluates the simulation-regression approach, and studies the impact of the regression methods on VOI calculation in the simulation-regression method. Chapter 3 includes a workflow provided for VOI calculation using the simulation-regression approach to have satisfactory accuracy, and a case study of CO_2 storage using this workflow. Lastly, Chapter 4 contains the conclusions and future recommendations of this thesis.

Chapter 2 –VOI Analysis and Simulation-Regression Approach

This chapter starts with the concept of VOI and examples. Then, the simulation-regression approach and different regression methods are introduced. Lastly, we focus on the simulation-regression accuracy, study the impact of regression methods on VOI calculation, and discuss the importance of selecting the best regression method for VOI calculation in the simulation-regression approach in order to have the maximum accuracy with this approach.

2.1 Monte Carlo Simulation

Uncertainty means not being sure about the trueness of a statement or outcome of a system (Bratvold and Begg, 2010). When we face uncertainty in a variable that influences the outcome we are interested in, for example, market demands when we want to produce a product, since it is mostly impossible to assess the uncertainties directly, a good approach is modeling the uncertainty. Monte Carlo Simulation (MCS) can be used to model and quantify uncertainties (Bratvold and Begg, 2010). According to Bratvold and Begg, (2010), MCS is a very popular mathematical technique that helps construct a model of uncertainties (Kenton, 2021). MCS works by randomly sampling probability distribution functions (PDF) representing uncertain input variables, using them to calculate the value of interest, which can be the net present value (NPV), based on the

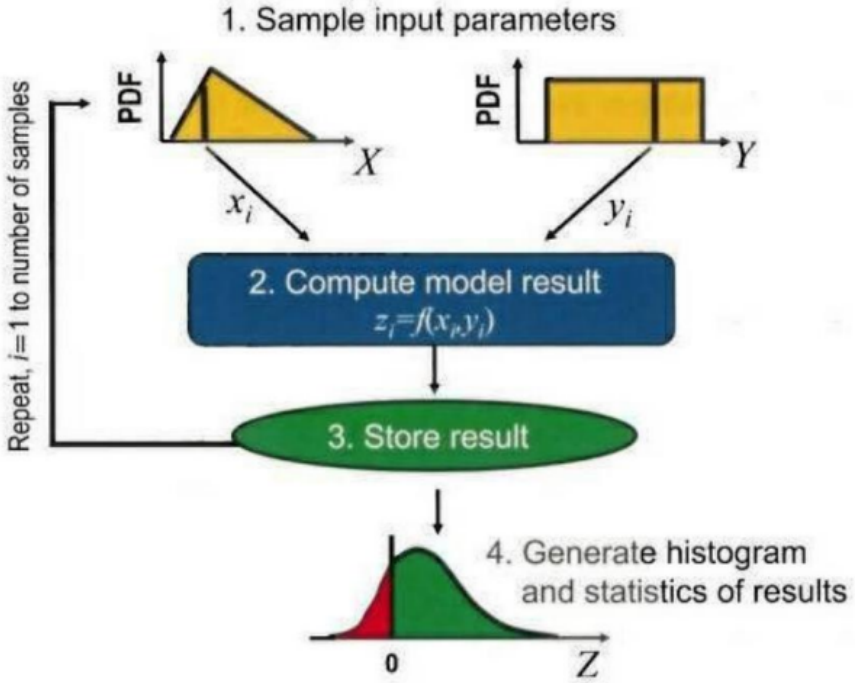


Figure 2.1 - Schematic of Monte Carlo simulation procedure (Bratvold and Begg, 2010)

model, and then repeating these steps multiple times with different inputs to the model (Bratvold and Begg, 2010, Eidsvik, et al., 2015). With MCS we can produce a large number of samples of possible outcomes based on the probability models (Kenton, 2021). Bratvold and Begg, (2010) showed these steps in Figure 2.1.

The MCS method has the following advantages:

- It is easy to use, and it can include complicated, non-linear mathematics without additional complexity.
- There is no need to estimate input variables distributions while using MCS because it is not limited to working with theoretical probability distributions.
- It can also cover extreme outcomes.
- It is able to solve complex problems where an analytical solution is not available.
- It is possible and easy to define dependencies among uncertainties in the model by this method.
- There is commercial software available to implement it.
- The probability of making errors using it in solving problems might be lower compared to the analytical approach.
- It is easy to make changes to the model quickly and investigate its performance.
- It is a well-known approach that both decision makers and analysts are familiar with it, and they are more likely to accept its results (Bratvold and Begg, 2010).

And, disadvantages of MCS method are:

- The probability distributions of samples are likely to have errors in comparison with the distributions they come from.
- It is computationally demanding since it is needed to be implemented on a large dataset in order to increase the accuracy (Bratvold and Begg, 2010).

2.2 Value of Information

According to Bratvold and Begg (2010), minor uncertainties are not the uncertainties making decisions difficult. With uncertainties holding the potential to affect the outcome severely, making decisions would be a hard task (Eidsvik, et al., 2015, Bratvold and Begg, 2010). We might think

of uncertainty as something that must be avoided or reduced as much as possible. Therefore, getting more information on uncertain variables might help in making decisions. However, information gathering might not be worth it since it might not change the initial decision (Bratvold and Begg, 2010). For example, let's assume the data available now about a reservoir, results in an initial decision to invest in a CO_2 injection project, and the decision maker is risk neutral, and hence she uses expected value as the decision metric. After conducting a sensitivity analysis on expected value by varying an uncertain input from its minimum to maximum, the decision might remain "Invest", where she shouldn't consider gathering more information, or change to "Not Invest", where there might be value in gathering information considering its cost. Information gathering comes with costs and efforts, and can inform about the uncertain variable (Bratvold and Begg, 2010). If the decision remains as the initial decision after gathering more information, information gathering was nothing but an extra and unnecessary cost. It didn't change the initial decision, it just made the decision maker more confident about her decision. VOI is not for improving confidence in decision making or reducing uncertainties (Bratvold, et al. 2009). These two do not bring value to the outcome (Bratvold, et al., 2009).

Therefore, Information must have 4 attributes to be considered valuable (Bratvold, et al. 2009, Eidsvik, et al., 2015): 1- Relevant. It must be dependent on and related to the distinction of interest. It must be able to change our beliefs. 2- Material. The Information must have the ability to change the initial decision. If the decision is still the same as before, the information is not reasonable to be gathered. 3- Economic. The value that the information brings, must be greater than its cost. 4- Observable. The results of the test must be observable to be used in the decision making process (Bratvold, et al., 2009, Eidsvik, et al., 2015).

As discussed by Eidsvik, et al. (2015); there are four steps for VOI analysis. First, introducing the decision situation with clarity, uncertainty x, and possible alternatives A. Without being clear and understanding what the situation is, making a good decision would be impossible. Second, identifying possible information gathering methods and the type of information that can be helpful. Third, making a spatial model of the situation. And forth, VOI analysis to see whether the information to be received is worth its costs (Eidsvik, et al., 2015).

Assuming the decision maker is risk neutral, VOI is the difference between expected value without (prior value) and with information (posterior value) (Bratvold, et al., 2009, Bratvold and Begg, 2010, Eidsvik, et al., 2015).

$$VOI = \begin{bmatrix} expected\ value \\ with\ information \\ (Posterior\ value) \end{bmatrix} - \begin{bmatrix} expected\ value \\ without\ information \\ (prior\ value) \end{bmatrix}.$$

$$VOI = PoV(x) - PV.$$

VOI cannot be negative and its lower bound is always zero because the decision maker can always say "No" to new information and not pay for it (Bratvold and Begg, 2010).

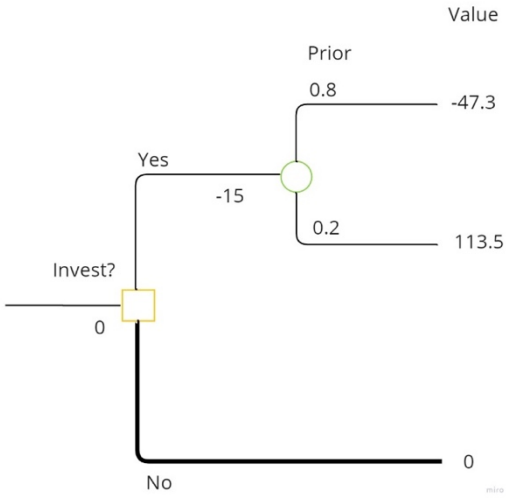
2.2.1 Prior Value

Assuming the decision maker is facing a decision situation with an uncertain variable x with a prior probability distribution $p(x)$. She must choose an alternative a, which maximizes the expected value, from available alternatives A (Eidsvik, et al. 2015). Thus, the Prior value can be calculated as:

$$PV = \max_{a \epsilon A}\{E(v(x,a))\} = \max_{a \epsilon A}\left\{\int_x v(x,a)p(x)dx\right\}.$$

Consider a case where a decision maker is facing an uncertain situation whether to invest in a CO_2 storage project. The target value is net present value and is a function of the storage capacity and costs.

She is uncertain about whether this project is profitable. She assumes a discrete probability distribution $[0.2, 0.8]$ for having positive or negative NPV respectively. Now, she can structure this decision tree as:



The bold line shows the optimal decision in this case where not investing is the best choice. If we define $a = 1$ the "Invest" and $a = 0$ the "Not Invest" alternative, the prior value is:

$$E[v(x, a = 1)] = 0.8 \times -47.3 + 0.2 \times 113 = -15,$$

$$E[v(x, a = 0)] = 0.$$

$$PV = \max\{E|v(x, a = 1), E|v(x, a = 0)\} = \max\{-15, 0\} = 0$$

2.2.2 Posterior Value

Now, the decision maker is considering gathering information to see whether it can increase the possibility of having good outcomes and change the initial decision (Bratvold and Begg, 2010). This information can be tests, studies, data, or experiments and as mentioned before, this information must be higher than its cost, be able to change the decision, observable, and relevant to the uncertainties (Bratvold and Begg, 2010). Based on the accuracy of the test, this information is defined by a likelihood probability $p(y|x)$ which is the probability of test y given uncertainty x (Eidsvik, et al., 2017).

Perfect Information

Perfect information is when the data is always correct, i.e., 100% accurate as the information y removes all uncertainty about x (Bratvold and Begg, 2010). Posterior value (PoV) is:

$$PoV(x) = \int_x \max_{a \epsilon A} \{E(v(x,a)|y)\} p(x)\, dx,$$

According to Bratvold and Begg (2010) the concept of perfect information can be really helpful as the value of perfect information is the most the decision maker is willing to pay for any kind of information. Perfect information is a hypothetical calculation to assess the maximum any other information should be worth it, and it gives a limit on value of any information gathered by a survey, test, human, etc. (Bratvold and Begg, 2010). In addition, as mentioned earlier, VOI cannot be negative. Therefore, VOI must be in the range of zero and the expected value of perfect information (Bratvold and Begg, 2010).

Imperfect Information

All available tests and studies provide imperfect information and are inaccurate due to instrument accuracy limit, human error, etc. (Bratvold and Begg, 2010). Thus, the posterior value and VOI with imperfect information y is calculated by:

$$PoV(y) = \int_y \max_{a \epsilon A} \{E(v(x,a)|y)\} p(y)\, dy.$$

To calculate the posterior value, pre-posterior, or the total probability $p(y)$, is needed. Pre-posterior probabilities are the probabilities that the test will indicate the high or low values, and is given by:

$$p(y) = \int_x p(y|x)p(x)dx.$$

And the expectation is given by:

$$E[v(x,a)|y] = \int_x v(x,a)p(x|y)dx,$$

$p(x|y)$ is the posterior and is calculated by Bayes' rule:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}.$$

In the CO_2 storage investment case, the PoV with perfect information is:

$$PoV(x) = \sum_x \max_{a\epsilon A}\{E(v(x,a)|y)\}p(x)$$

$$= \max\{v(0,0), v(0,1)\}p(x=0) + \max\{v(1,0), v(1,1)\}p(x=1)$$

$$= \max\{0, -43.7\}(0.8) + \max\{0, 113.5\}(0.2) = 22.7.$$

The VOI is then:

$$VOI = PoV(y) - PV,$$

$$VOI(x) = PoV(x) - PV = 22.7 - 0 = 22.7.$$

This shows that the maximum the decision maker should pay for any information is 22.7. This is a relatively high VOI and thus the decision maker may want to assess the value of imperfect information. If the value of perfect information had been very small (less than what any relevant information gathering activity would cost), the decision maker should now conclude that it is not worthwhile gathering more information before making the decision.

In order to consider this case with imperfect information, the decision maker needs a likelihood probability for the test. Assume there is a test available that has an accuracy of 90%, the likelihood probabilities are given by $p(y=1|x=1) = p(y=0|x=0) = 0.9$.

First, pre-posteriors, should be calculated for each outcome of the test and since this example is discrete, pre-posteriors are calculated by:

$$P(y) = \sum_j P\left(y\middle|x_j\right)P(x_j),$$

Therefore, pre-posteriors are:

$$p(y=0) = \sum_x p(y|x)p(x) = p(y=0|x=0)p(x=0) + p(y=0|x=1)p(x=1)$$

$$= 0.9 \times 0.8 + 0.1 \times 0.2 = 0.74,$$

$$p(y=1) = \sum_x p(y|x)p(x) = p(y=1|x=0)p(x=0) + p(y=1|x=1)p(x=1)$$

$$= 0.1 \times 0.8 + 0.9 \times 0.2 = 0.26.$$

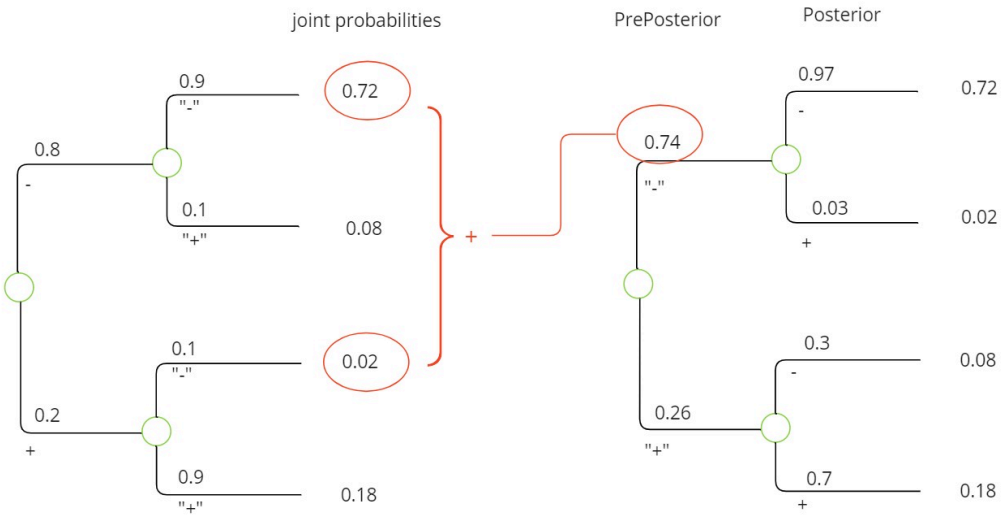
Second, posteriors for each outcome of x given the different outcomes of y are calculated as:

$$p(x=0|y=0) = \frac{p(y=0|x=0)p(x=0)}{p(y=0)} = \frac{0.9 \times 0.8}{0.74} = 0.97,$$

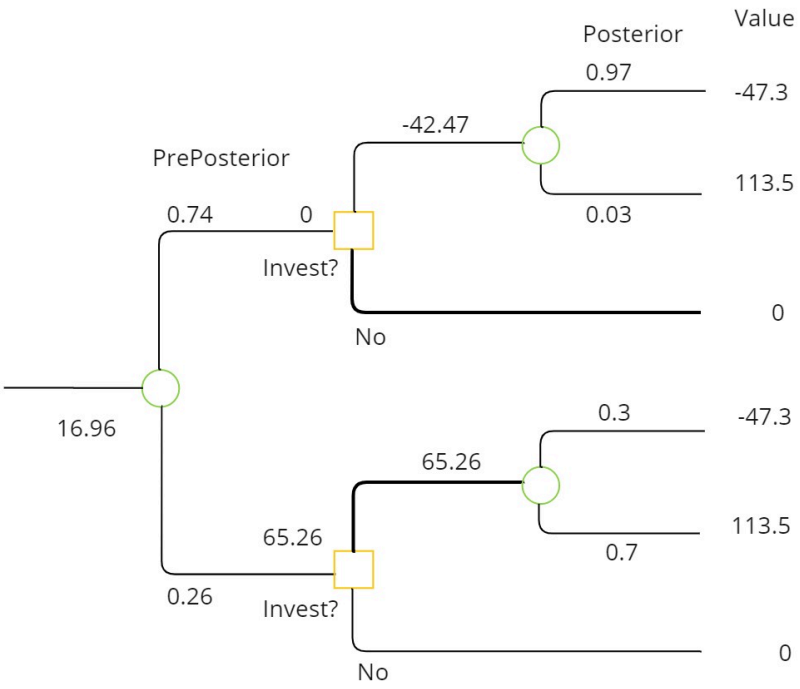
And correspondingly for the rest:

$$p(x=1|y=0) = 0.03, p(x=0|y=1) = 0.3, p(x=1|y=1) = 0.7.$$

These results can also be illustrated by flipping the decision tree using Bayes' rule:



The decision situation after having information is then shown as:



And then, PoV(y) can be calculated:

$$PoV(y) = \sum_{y} \max_{a \epsilon A}\{E(v(x,a)|y)\}p(y)$$

$$= \max\{E[v(x,a)|y=0]\}p(y=0) + \max\{E[v(x,a)|y=0]\}p(y=1)$$

$$= \max\{0,-42.47\}(0.74) + \max\{0,65.26\}(0.26) = 16.96.$$

Thus, if the test with 90% accuracy costs more than 16.96, the decision maker should reject it. And as we can see from the decision tree above, the information is 1-Relevant since it is informing about the uncertainty, 2- Material since the decision can change based on the result of the test, 3-Economic if it costs less than 16.96, and 4-Observable since the result of the test is observable as "−" and "+".

The VOI will be reduced if the accuracy of the test is reduced. The more accurate the test, the more valuable it becomes (assuming that the specificity and sensitivity of the test are the same) (Modeling for Decision Insights, lecture notes, 2021). The minimum value of the test is when the accuracy is 50% since it doesn't give any information about the uncertainty and cannot change prior probabilities, i.e., the test is not relevant for the underlying uncertainty. If it doesn't matter what the test indicates, it cannot change the prior probability. Having a test with accuracy less than 50% is more informative, since we would know whatever test says, it is likely to be wrong. In the Figure 2.2, a sensitivity analysis is done for VOI as a function of accuracy.

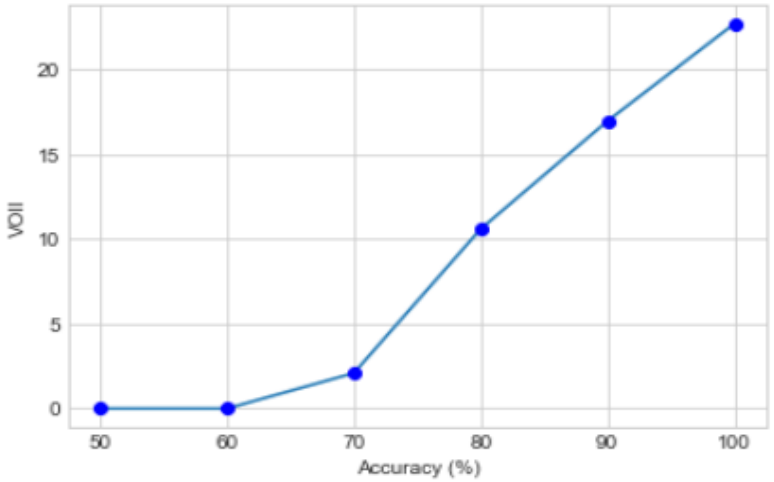


Figure 2.2 - The sensitivity analysis on effect of test accuracy on VOI for CO_2 storage investment case

Figure 2.2 shows that with increase in accuracy of the test result, the VOI is elevated and with accuracy below 60%, the information doesn't have value.

2.3 Value Discretization

The probability distribution of parameters in decision making situations can be discrete or continuous depending on the situation at hand. If continuous distributions are used, they need to be discretized for use in decision trees and perform mathematical operations. There are multiple methods to discretize a continuous distribution such as 3-Point Shortcuts and N-Point Discretization methods including Value Discretization, CDF Discretization, and Moment Matching. Each of these methods has its advantages and disadvantages. For instance, 3-point shortcuts are simple to use but have limitations in approximating the tails of the distribution and should not be used where the extreme values and their probabilities need to be identified from the

distribution (Modeling for Decision Insights, lecture notes, 2021). In the Moment Matching, optimization is needed for the probabilities which might need complex calculations. The CDF discretization also cannot capture values around the tail, and the value discretization method is computationally demanding when the number of uncertain variables and the number of value points increase. However, it can capture values from the tails of the distributions (Modeling for Decision Insights, lecture notes, 2021).

In value discretization method, the value range is discretized into N sections. Then, the mid-value of each section i is used to calculate the probability density using this mid-value (x_i) and PDF, $f_i = f(x_i)$. Then, f_i is normalized to find the probability for x_i, $P_i = \frac{f_i}{\sum_{i=1}^{N} f_i}$ and $\sum_{i=1}^{N} P_i = 1$. Value discretization is a great approach to cover the PDF shape including the tails and extreme values (Modeling for Decision Insights, lecture notes, 2021). In this study, value discretization method is selected among other methods due to its ability to represent the distribution and accuracy which are both important elements in VOI analysis.

2.4 Simulation-Regression Approach

Simulation-regression approach is based on using Monte Carlo simulation with regression methods to calculate PoV and VOI (Eidsvik, et al., 2017). In this study, we use simulation-regression approach to calculate VOI for a sequential information gathering scheme. As discussed by Eidsvik, et al. (2017) and Eidsvik, et al. (2015), the simulation-regression method includes multiple steps:

1. Generate B samples of parameter x based on prior probability $p(x)$ as $x^1, \dots, x^B$. the parameter x represents uncertainties, which are porosity, thickness, pressure, depth, temperature, and cost in the case study of this thesis.
2. For each sample of x (x^b), and for each alternative a, generate values $v_a^b = v(x^b, a)$ and data samples y^b by using forward modeling. In our case study, values are calculated with NPV functions mentioned in Chapter 3.
3. Fit a regression model for each alternative a values:

$$\hat{v}_a^b = F_a(y^b),$$

Which approximates the conditional expectation $E[v(x,a)|y^b]$, and the regression method is chosen by a 10-fold cross-validation.

4. Calculate the prior value. Prior value (PV) ensures that the value of information never gets negative. PV is calculated as (Eidsvik, et al. 2015):

$$PV = \max_{a \epsilon A} \left[\frac{1}{B} \sum_{b=1}^{B} v_a^b \right],$$

5. Approximate posterior value by:

$$\begin{aligned} PoV(y) &= \int_y \max_{a \epsilon A} \{E(v(x,a)|y)\} p(y)\, dy \\ &\cong \frac{1}{B} \sum_{b=1}^{B} \max_{a \epsilon A} E[v(x,a)|y^b] \\ &\cong \frac{1}{B} \sum_{b=1}^{B} \max_{a \epsilon A} \hat{v}_a^b. \end{aligned}$$

6. Calculate VOI:

$$VOI = PoV(y) - PV.$$

The regression function depends on the number of samples B, and the regression method chosen, which we discuss in the following sections. In addition, VOI should be compared with the cost of the information to see if the information is worthwhile.

Imperfect information in the simulation regression approach can be represented by adding a measurement error as a noise value to change the test results, y^b, before building the regression model. In this study, the noise value is represented by a normal distribution MCS with a mean of x^b and standard deviation of samples variable's average multiplied by the percentage error as:

$$y^b = x^b + \left(\frac{1}{B} \sum_{b=1}^{B} x^b \right) . \frac{\%noise}{100}.$$

Thus, when the noise is 0%, the information would be perfect as it represents the actual value of uncertainty x^b, $y^b = x^b$. By using MCS, correlation effects and biased results can be produced since each repetition of VOII calculation has a different result. Therefore, we need to repeat the calculation of VOII and then find the average. In this study, we used 100 repetitions for calculating VOII. A sensitivity analysis should be done on the number of iterations to minimize the computing time needed balanced with getting reasonable result. For example, for the CO_2 storage investment example, Figure 2.3 shows the histogram of VOII with 10% noise to show the range of calculated VOII.

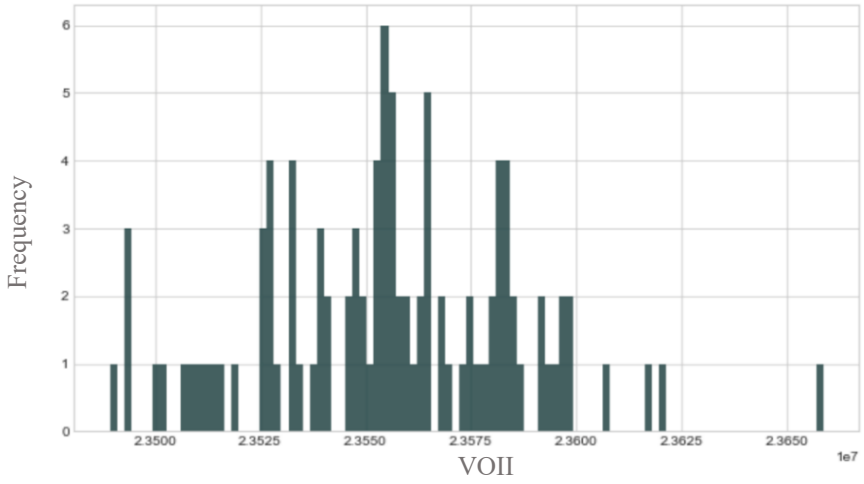


Figure 2.3 - Histogram of VOII with 100 iterations

2.5 Regression methods

Regression methods are important elements in the simulation-regression approach. In this section, we introduce and evaluate 6 different regression methods, Ordinary Least Square Linear Regression (OLS), *K*-Nearest Neighbor (KNN), Random Forest (RF), Extreme Gradient Boosting (XGB), Support Vector Regression (SVR), and Piecewise Regression, on different types of datasets and focus on their impact in VOI calculation in the simulation-regression approach.

Regression methods are models built to relate independent variables to a target variable (dependent or response variable) and based on these models, we would be able to describe the relationship between these variables and predict the target variable with some new inputs (Agami Reddy, 2011). Given datasets including both independent and dependent variables, we can fit a regression model and use this model for another dataset to predict target values based on regressor

variables of a new dataset (Agami Reddy, 2011). In doing this, we need to find a suitable regression model. Factors like the type of data, shape of the data, number of variables, and error metrics should play into this choice (Chugh, 2020). As an aid in selecting a suitable model, scatter plot data underlying each uncertainty and calculated values can provide an initial understanding of the behavior of the system. For example, by plotting the 2D or 3D scatter plots we might observe linearity or non-linearity making the dependency structure clearer.

2.5.1 Ordinary Least Squares Linear Regression (OLS)

Ordinary least squares is one of the most common methods (Agami Reddy, 2011). This method is based on minimizing the sum of squared errors or differences between data and model which is given by $(\sum_{i=1}^{n} D_i^2)^{\frac{1}{2}}$. This method is also called Method of the Moments Estimation since it is related to squared errors (Agami Reddy, 2011).

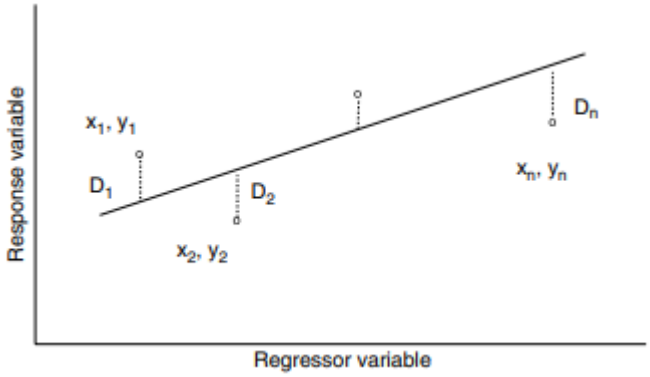


Figure 2.4 - Least Square method

A multivariate linear model (Polynominal) is given by:

$$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_i X_i + \varepsilon\,, \quad i = 1, \ldots, n$$

Where ε is the error in the model. The goal of this method is to determine β_0 , β_1,…, β_n parameters to be able to relate y to X_1, X_2,…, X_n variables in the best way. This model is simple and works well for linear systems. It finds single or multivariate linear regression models from data (Agami Reddy, 2011). If one finds the linear regression the best regression model for a specific dataset, she can also use forward, backward, and stepwise regression methods as a features selection technique to remove features with less influence on the target variable and reduce dimensionality of the system, making the regression faster.

2.5.2 *K*-Nearest Neighbors (KNN)

K-Nearest Neighbor is a pattern recognition method that is commonly used in regression and classification (Agami Reddy, 2011). The KNN method is a great tool for modeling non-linear systems. It is based on a distance measure that with a given k, finds the distance of the specific input from its k nearest neighbors (Agami Reddy, 2011). This method needs 1- a definition of distance, and 2- a number of neighbors (k) to take into account. The KNN algorithm is easy to understand, fast, and is described as follows (Gupta, Sehgal, 2021):

1. Select the k, as the number of neighbors.
2. Calculate the distance between the test point and each data point.
3. Sort the distances.
4. Select the *k*-nearest neighbors from sorted distances.
5. Calculate the average of *k*-nearest neighbors values to assign the value for the test point in regression.

KNN method is extremely sensitive to *k* values which makes it easy to overfit the model with a very small *k*. In addition, by selecting a very large *k*, the regression model cannot capture the behavior of the dataset. Thus, 10-fold cross-validation must be done on the regression to find the *k* producing the best fitted model while preventing overfitting. In this study, a different number of *k* in the range [1,500] is assigned to 10-fold cross-validation to find the best fit.

2.5.3 Random Forest (RF)

Random Forest Regression is one of the supervised learning algorithms which uses the ensembling learning method (Bakshi 2020). The ensembling method merges machine learning algorithms' predictions to provide an accurate prediction. In the random forest workflow, several independent trees run in parallel during training time and use the mean of the classes as the prediction of all trees. RF regression models are suitable for capturing non-linear relationships. One disadvantage of RF is that it can easily be overfitted and the number of trees should therefore be chosen carefully (Bakshi, 2020).

The RF regression is made by growing trees based on independent and identically distributed random vectors (Θ), averaging over k of the trees $\{h(x, \Theta_k)\}$, controlling overfitting, and

improving the prediction accuracy (Breiman, 2001). In summary, RF regression method tries to fit trees on several sub-samples of the whole dataset, improves accuracy, and prevents overfitting by averaging (Pedregosa, et al. 2011). A different number of maximum leaf nodes in the range [1,100] is assigned to 10-fold cross-validation.

2.5.4 Extreme Gradient Boosting (XGB)

Gradient Boosting method is based on making a strong model in predicting using a loss function from multiple weak models (Natekin and Knoll 2013). According to Natekin and Knoll (2013), the main concept of the gradient boosting family is building a model sequentially and by iterations. At each iteration, a new weak model is produced using the errors of the model built so far and making it more accurate. The loss function for this model can be customized by the user. Friedman, (2001) introduced the gradient boosting algorithm as below:

1. Initialize the model with a constant as a first guess.
2. Compute the negative gradient.
3. Fit a new base-learner function.
4. Find the best gradient decent step.
5. Update the model.
6. Repeat from step 2.

XGB is based on the gradient boosting algorithm with some improvements in prediction and accuracy. It is a combination of classification and regression tree (CART) (Babajide, Saeed 2016), and is a powerful method built to be useful for large and complicated datasets. A different number of the hyperparameter maximum depth in the range [1,10] is assigned to 10-fold cross-validation.

2.5.5 Support Vector Regression (SVR)

The support vector algorithm looks for non-linearity in the data to provide a model to predict (Raj, 2020). SVR uses Support Vector Machines (SVM) which is a supervised machine learning method in classification and regression (Raj, 2020). In support vector regression, the hyperplane is the straight line needed to fit the data and the SVM tries to generate it where the data point on both sides of it are called support vectors. Thus, support vectors can change the hyperplane position.

The SVR method considers a threshold instead of minimizing the error that other regressions do (Raj, 2020). This method is slow such that for large datasets, linear SVR is being used even though it only considers a linear kernel. Another disadvantage of this method is that it cannot perform well on data with noise. We used scikit-learn SVR regressor in this study (Pedregosa et al., 2011). SVR cannot handle more than 10,000 data. Since in this study we are generating 100,000 samples and this is a large dataset for SVR, we should split the data set into at least 10 splits to have a maximum of 10,000 data in each split.

2.5.6 Piecewise regression

Piecewise regression splits the dataset and does a regression, mostly linear regression, on each split. In this study, we implemented this technique with linear and support vector regression. We used KBinsDiscretizer of scikit-learn library for splitting datasets and then fitted a regression model for each split. The whole process is done by mlinsights extensions of scikit-learn library.

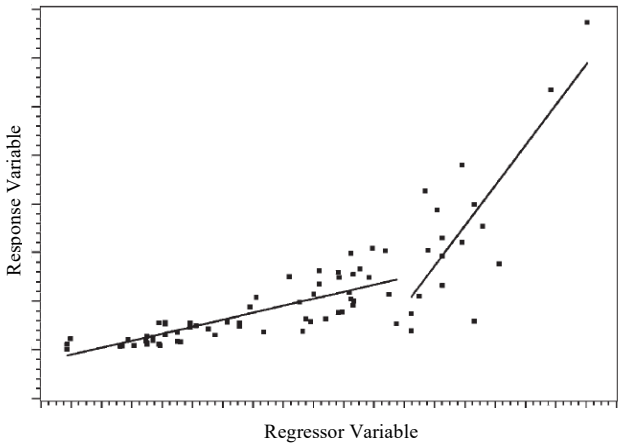


Figure 2.5 - Piecewise regression with linear regression (Kalvelagen, E, 2018)

2.6 Cross-Validation

Overfitting and underfitting are very common concepts in machine learning methods. In underfitting cases, the model cannot capture the relation between the independent variables and the target variable, resulting in low scores on both training and test scores. This can happen due to either the model being too simple, the need for more features, or the need for less regularization (Nikolaiev, 2021). In overfitting cases, the model also captures part of the noise in the data and fits to it, which can cause a poor prediction for new data (Agami Reddy, 2011). In these cases, the

model works well on the training data but performs poorly on the test. Hyperparameters in regression models are to control models from overfitting and underfitting. A method to control hyperparameters and find the best regression fitted to the data is k-fold cross-validation. k-fold cross-validation can tell if the model is overfitted or underfitted and shows the prediction ability of the model (Agami Reddy, 2011). In this study, we use $k = 10$, which becomes 10-fold cross-validation and is a common approach in machine learning field (Grootendorst, 2019).

2.6.1 Error Metrics

In every machine learning model, one of the most important steps is to check how well the model is fitted and what the error of the model is (Chugh, 2020). There are different methods to evaluate the model regressed such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R-squared score, etc. each of these methods brings insight to the regressed model and evaluates its performance. For example, MSE represents the difference between the actual value and the value predicted, and R-squared shows how well the model is able to predict the values. In this study, the R-squared measure is selected as cross-validation score. R^2 score, is calculated by:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}.$$

Where, $\bar{y}$ is the average of data and $\hat{y}$ is the predicted value, and y_i actual value of data points. The R^2 score closer to 1, the better the model represents the dataset (Chugh, 2020).

2.6.2 Comparing Different Regression Methods

In this section, we discuss the impact of the regression method on different data types and show how selecting the best regression method is affecting VOI calculation in the simulation-regression approach.

In VOI calculation in the simulation-regression method, the regression method chosen should be a function of the dataset and its behavior. For example, in a CO_2 capacity estimation case with only one uncertainty (thickness) which has a linear relationship with the target value (NPV), regression results using different methods are shown in Figure 2.6. Before comparing different

regression results to find the best fit, cross-validation is used in each regression to identify the best fit each method can build.

As expected, the OLS result is a perfect fit to a linear system while the RF does not result in nearly as good fits to the dataset. KNN and XGB also provide good results for a linear function.

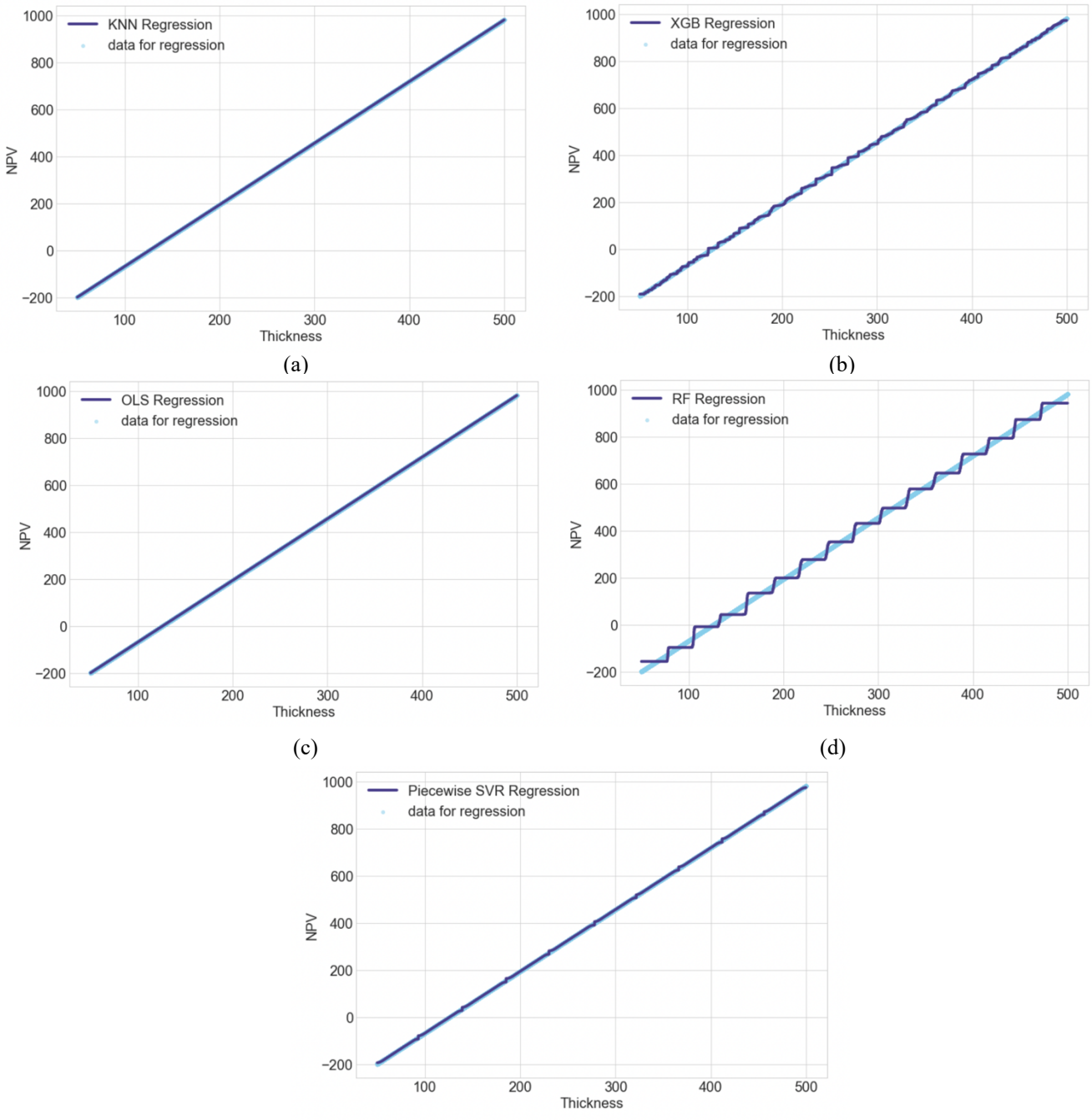


Figure 2.6 - Different regression methods results on a linear function with thickness as the only uncertainty

In Figure 2.6, light blue markers represent the data for the regression and dark blue lines the regressed model. (a) KNN, (b) XGB, (c) OLS, (d) RF, and (e) Piecewise SVR with 10 splits.

We use the value discretization method for comparison purposes as the correct answer. Value discretization's accuracy increases with an increase in the number of values. Thus, we increase the number of value points in the range of 3 to 100 to find where the VOI calculated stops changing by 2 decimal digits. After doing sensitivity analysis on the number of values in the value discretization, 30 as the number of values is selected as the correct answer for VOI calculation. In Table 2.1, there is still a difference between linear regression results and the value discretization which is due to using MCS and it decreases by increasing the number of samples in the simulation-regression method. Moreover, the expected value without information (EVwoI) of regressions using the simulation-regression method is different from the one calculated using value discretization. This difference is also a result of using MCS and gets smaller with increasing the number of samples. All of the variations of the simulation-regression method have the same EVwoI since we used the same set of MC samples for all the methods. It is obvious that we can achieve different EVwoI each time running MCS and this value gets closer to the value discretization by increasing the number of samples. In addition, each regression method can provide a reasonable fit in a specific running time which can be time-consuming for some methods. For example, XGB can provide a perfect fit to a linear function with a 2.9% error in VOI calculation, which takes 182.8 seconds to run, but a simple linear regression takes only 0.8 seconds with the same accuracy in VOI calculation. It is not reasonable to use XGB to regress a linear function.

Table 2.1 - VOI analysis of regression methods with uncertain thickness

	Value Discretization *30 values*	*OLS*	*XGB*	*KNN*	*RF*	*SVR Piecewise*
EVwoI	390.68	390.58	390.58	390.58	390.58	390.58
EVwI	407.47	407.85	407.85	407.85	406.75	407.85
VOI	16.79	17.26	17.26	17.26	16.16	17.26
VOI Error	-	2.9%	2.9%	2.9%	3.7%	2.9%
Running time (s)	0.1	0.8	182.8	0.67	1.2	154

Lastly, the accuracy of the regression increases with an increase in the number of samples. A sensitivity analysis is done on the number of samples on linear regression result, and Figure 2.7 shows how the accuracy of a linear regression increases with the number of samples. Thus, the error resulting from too few samples in the MCS contributes to the overall error in the VOI. For example, with 10,000 and 10,000,000 samples, the percentage error in VOI calculation is 25% and 1% respectively.

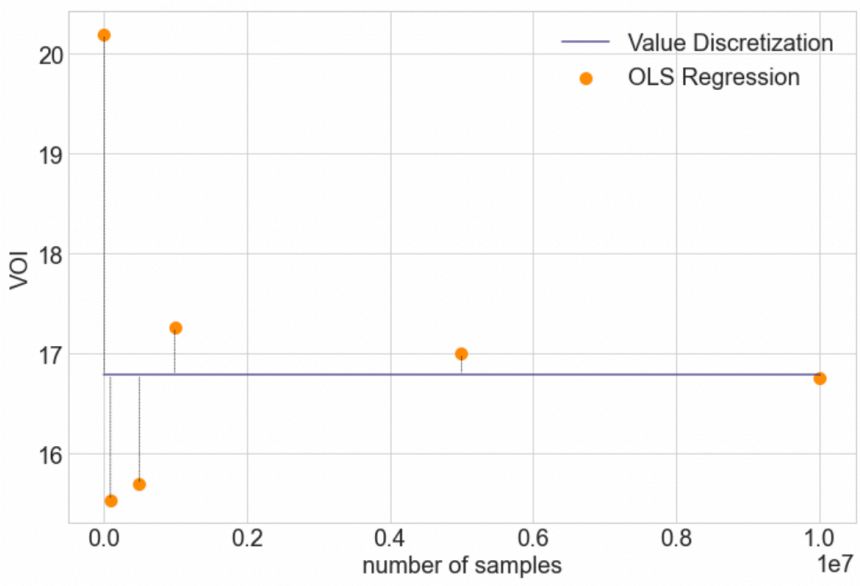


Figure 2.7 - The sensitivity analysis on number of Monte Carlo samples vs linear regression accuracy on a linear function.

Figure 2.8 shows an example where the NPV is not a linear function of the underlying uncertainty, which in this case is pressure.

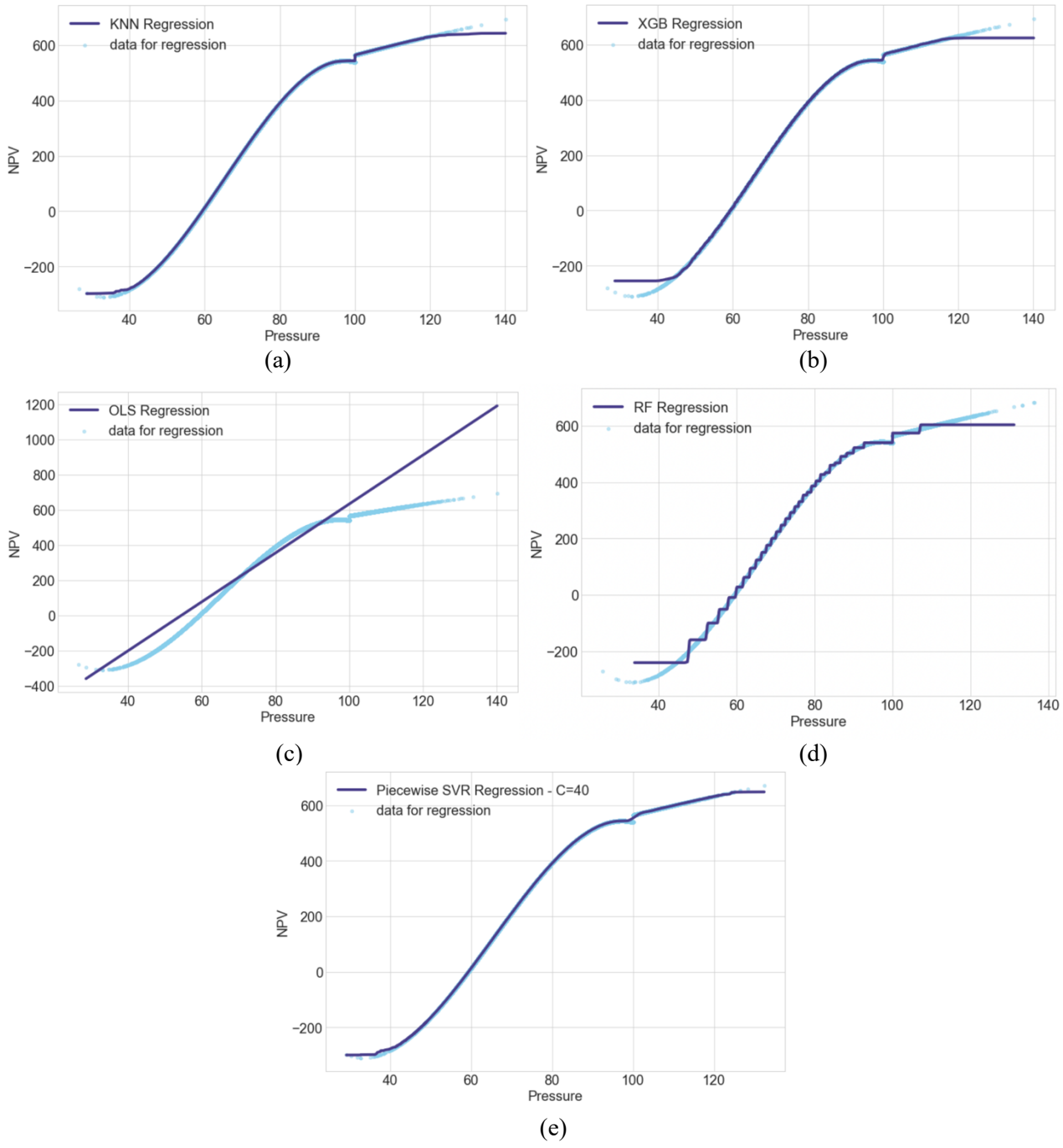


Figure 2.8 - Different regression methods results on a non-linear function with pressure as the only uncertainty

As we can see from Figure 2.8, linear regression cannot fit a non-linear relationship of dependent and independent variables. In this case, KNN, Piecewise SVR, and XGB are the best

fitting methods, and RF provides a slightly better fit than OLS. These results are also achieved in R^2 scores achieved by cross-validation which are 0.92, 0.99, 0.99, 0.96, and 0.99 for Linear, XGB, KNN, RF, and Piecewise SVR respectively.

As discussed earlier, in piecewise regression we split the dataset and fit a regression for each split. The piecewise linear correlation approach identifies linearity in smaller parts of the overall system. As shown in Figure 2.9, we split the dataset into 2, and 10 splits and then used linear regression for each set, compared the predicted result with value discretization, and the case with a single linear regression for the entire range. In this case, the correct VOI about pressure from the value discretization calculation is 3.61, while with simple linear regression, which is shown in Figure 2.8, VOI is 2.85 with a 21% error. With piecewise regression, VOI is 3.30 and 3.37 with 2 and 10 splits, and 8% and 6% errors respectively. Thus, piecewise regression can be a very powerful approach to deal with non-linearity. The accuracy of the models in piecewise regression improves with an increase in the number of splits. In Appendix 2, a sensitivity analysis of the number of splits on the prediction results is presented including cases with a large number of splits. The splits have an equal number of data points and the way they have been split shows the higher density of the data in the middle of the plots than in the tails which explains the inaccurate results in tails.

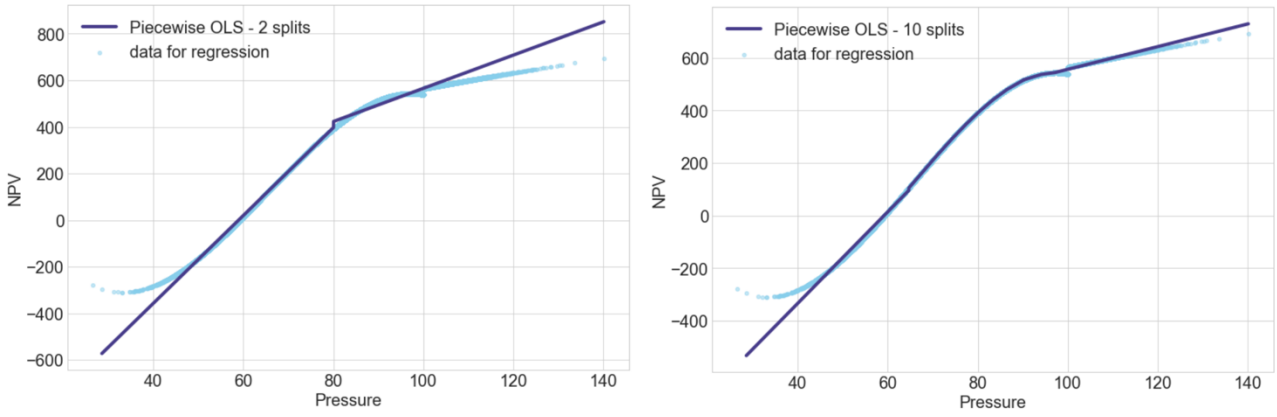


Figure 2.9 - Piecewise OLS regression with 2 and 10 splits

Results of the regression methods are sensitive to their hyperparameters. We should do 10-fold cross-validation to find optimal values for these hyperparameters, which are k in KNN, maximum leaf nodes in RF, maximum depth in XGB, and regularization parameter C in SVR, to avoid overfitting and underfitting. For instance, 10-fold cross-validation is done on RF regression with only pressure as the uncertainty to study its behavior and find the optimal leaf nodes. A range of [1,100] for maximum leaf nodes is given for cross-validation, and as shown in Figure 2.10, maximum leaf nodes in a range of [30,100] are having the highest test score where no overfitting or underfitting is happening. We selected 30 to avoid long running time while having satisfactory accuracy.

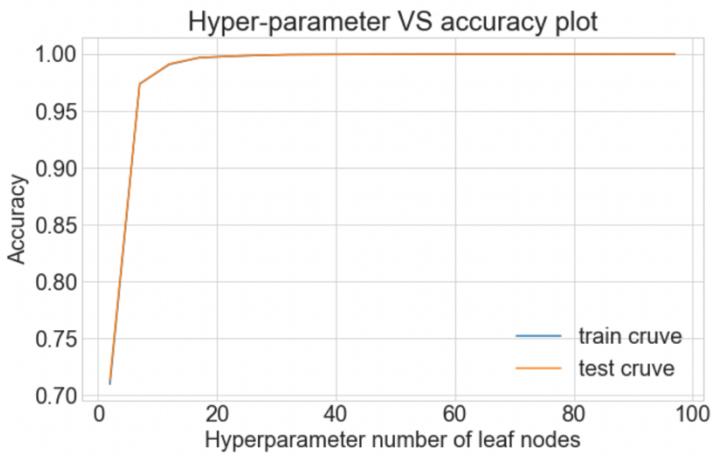


Figure 2.10 - Cross-validation on hyperparameter maximum leaf nodes for RF regression with pressure as an uncertainty and feature

As shown in Figure 2.10, RF result depends on maximum leaf nodes and its accuracy enhances with an increase in the maximum leaf nodes hyperparameter. We can choose between 30 and 100 based on the running time of regression and accuracy needed since having more leaf nodes results in a longer running time. In Appendix 1, R^2 scores of all regression methods with 3 uncertainties and features, thickness, pressure, and porosity are shown.

Figure 2.11 shows RF regression results for a non-linear relationship where pressure is the only uncertainty and we use the maximum leaf nodes of (a) 5, (b) 15, and (c) 30.

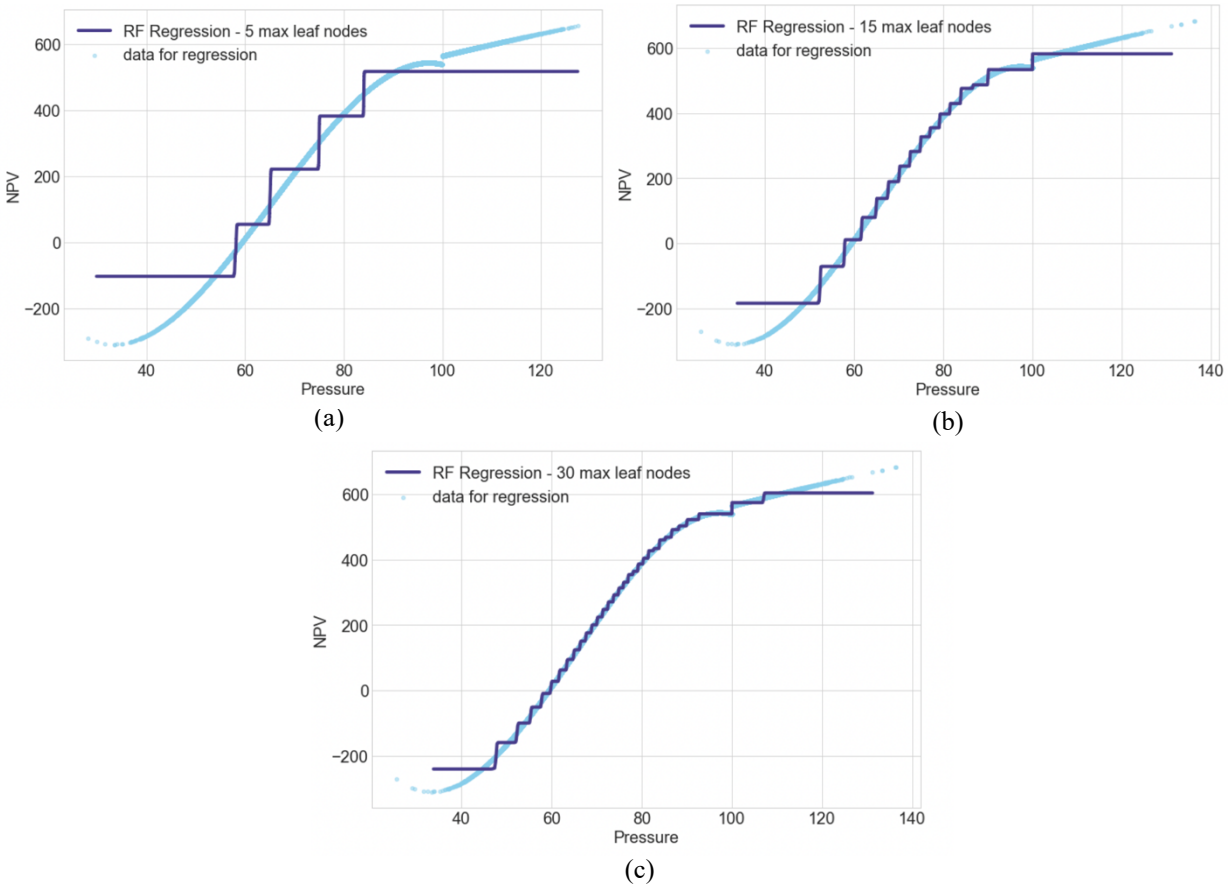


Figure 2.11 - The sensitivity analysis on effect of hyperparameter maximum leaf nodes in RF regression results

In addition, cross-validation is done on hyperparameter C, which is the regularization parameter in SVR. C must be positive, and a sensitivity analysis is done on it with values in a range of [1,50]. The plots of results are shown in Figure 2.12, for (a) $C = 1$, (b) $C = 10$, and (c) $C = 40$. Lastly, $C = 40$ is selected.

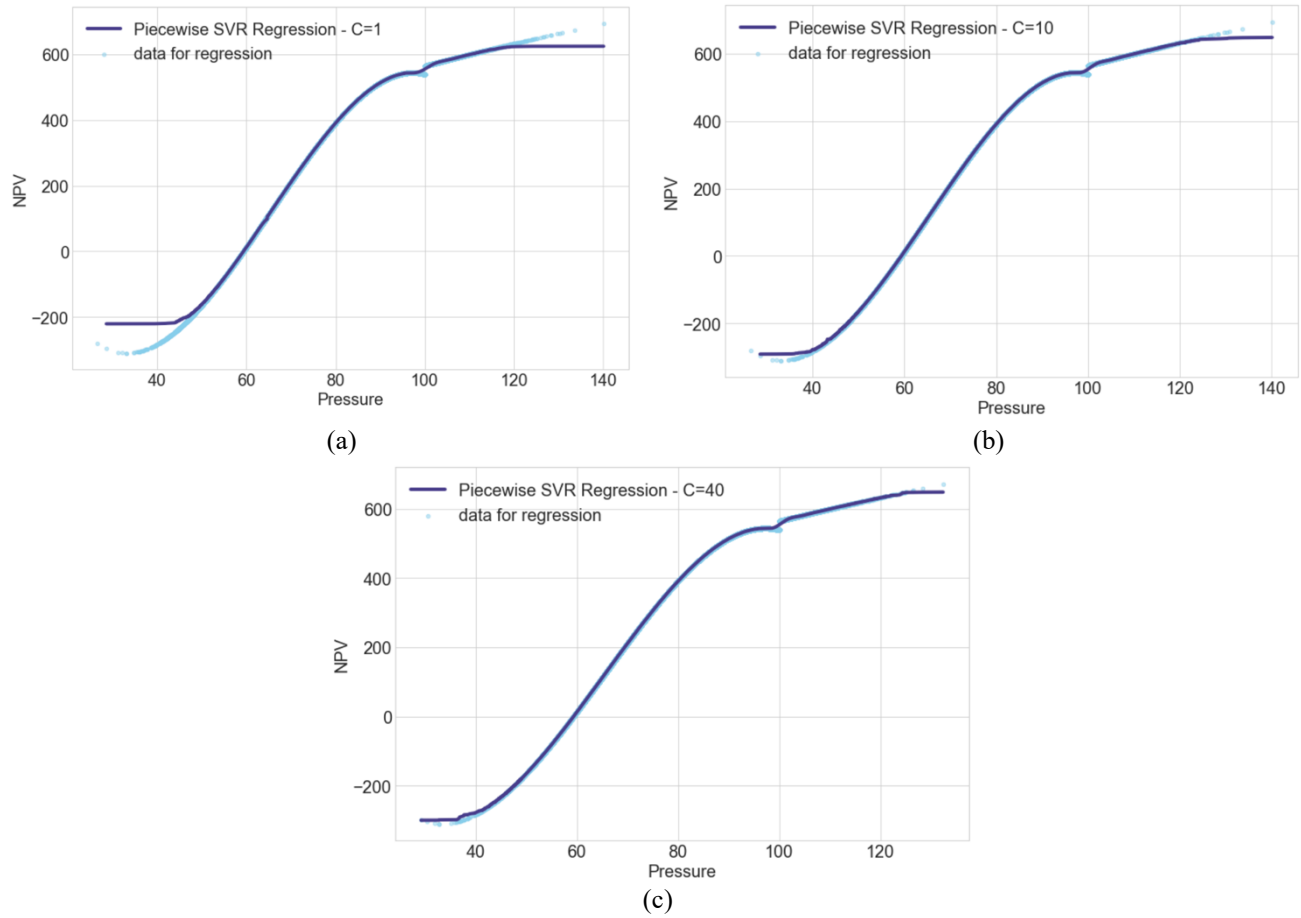
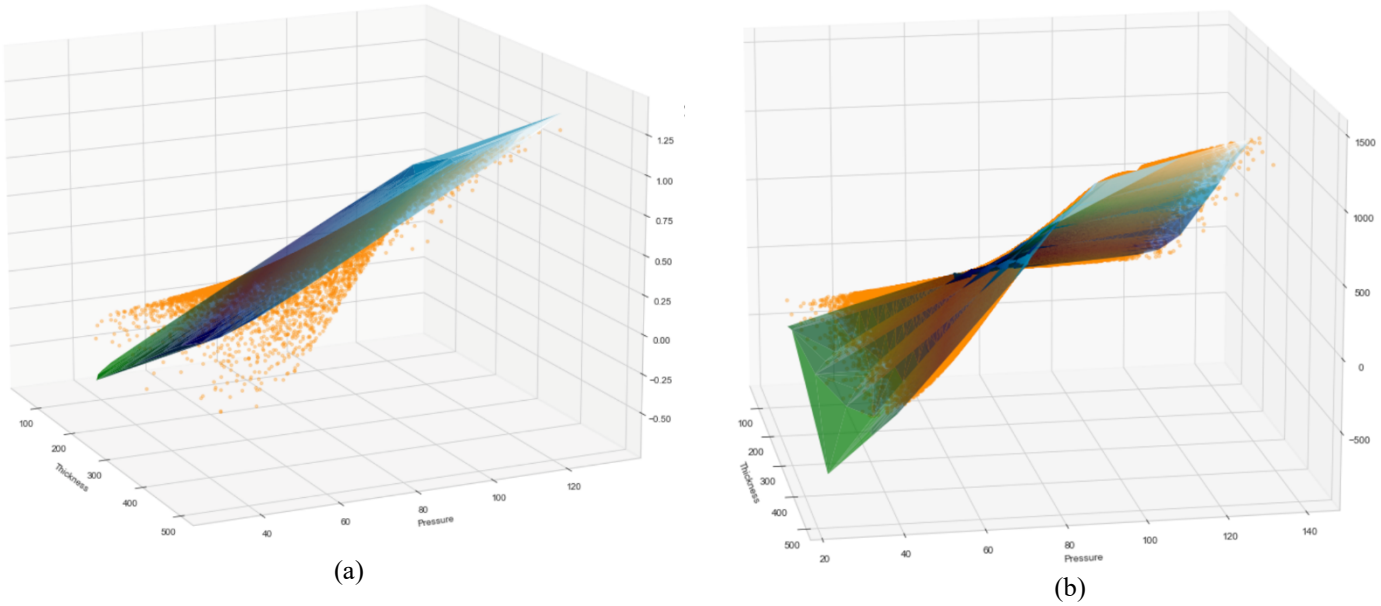


Figure 2.12 - The sensitivity analysis on hyperparameter regularization parameter in SVR with values of (a) 1, (b) 10, and (c) 40.

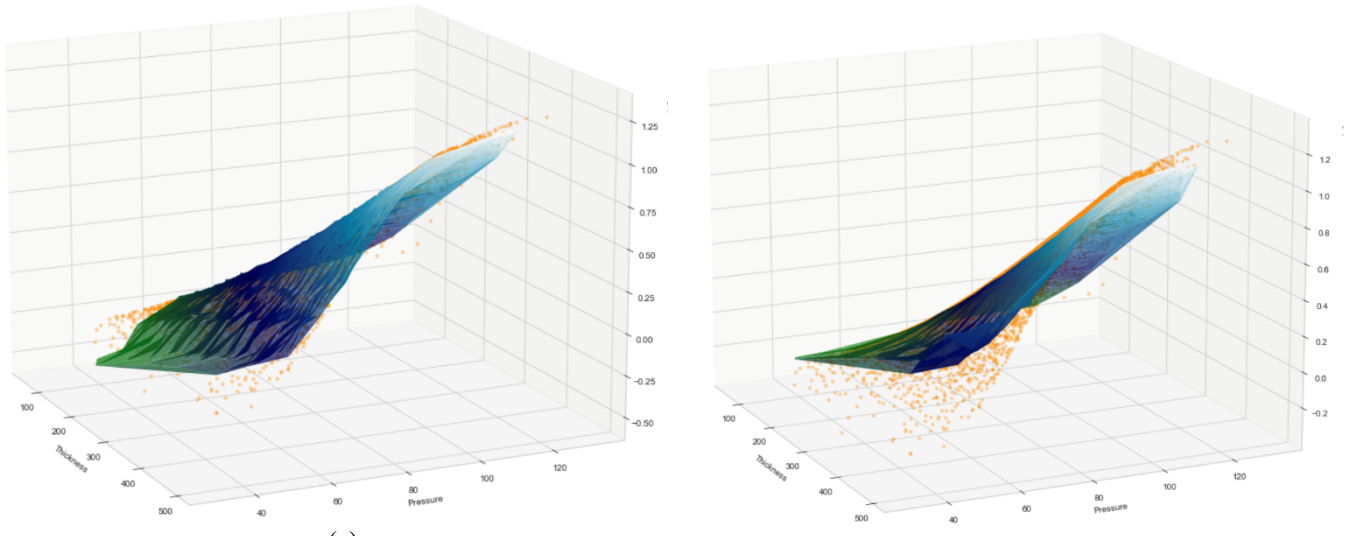
As shown in Figure 2.12, decreasing *C* leads to decreasing accuracy of the prediction and regression cannot cover tail and extreme values very well which play important roles in VOI analysis.

Now, to further investigate the different regression methods' abilities in fitting and predicting different data types and impacting VOI, the case with both uncertainties (thickness and pressure) is considered, and multivariate regression models are constructed. The relationship between target and independent variables is now non-linear. For RF, KNN, SVR, and XGB, 10-fold cross-validation is conducted to control overfitting and underfitting by the hyperparameters the maximum leaf nodes, *k*, *C* regularization parameter, and the maximum depth, respectively. Plots of multivariate regression results are shown in Figures 2.13, 2.14, and 2.15.



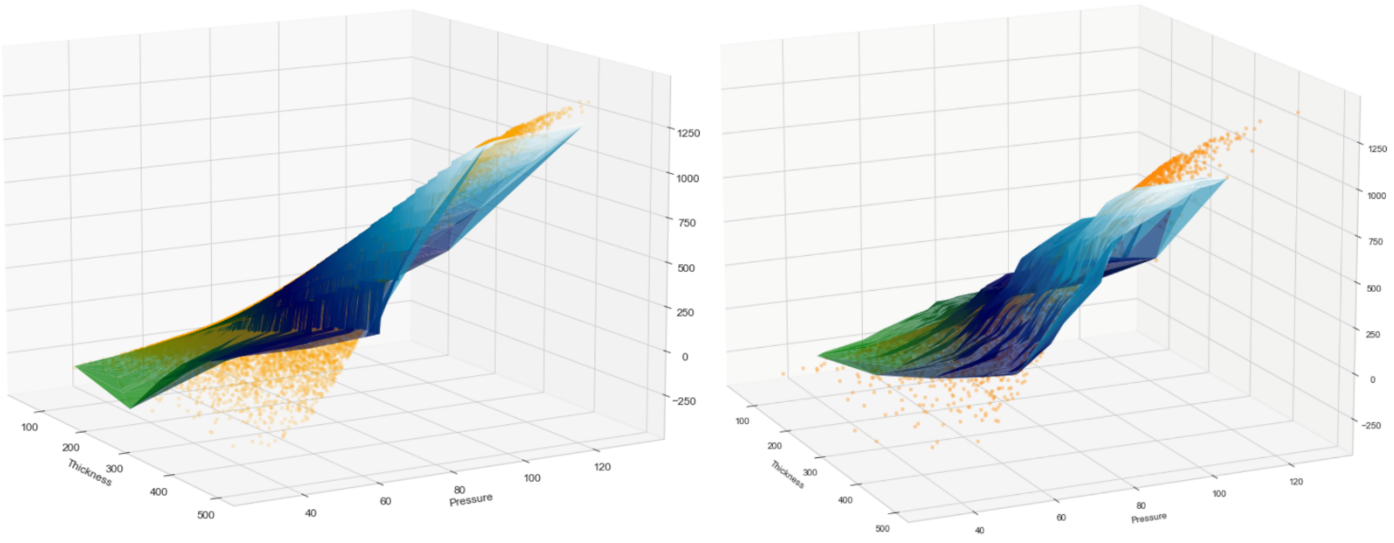
(a) (b)

Figure 2.13 - Multivariate regression plots. (a) OLS, (b) Piecewise OLS with 4 splits



(a) (b)

Figure 2.14 - Multivariate regression plots. (a) XGB, (b) KNN.



(a) (b)

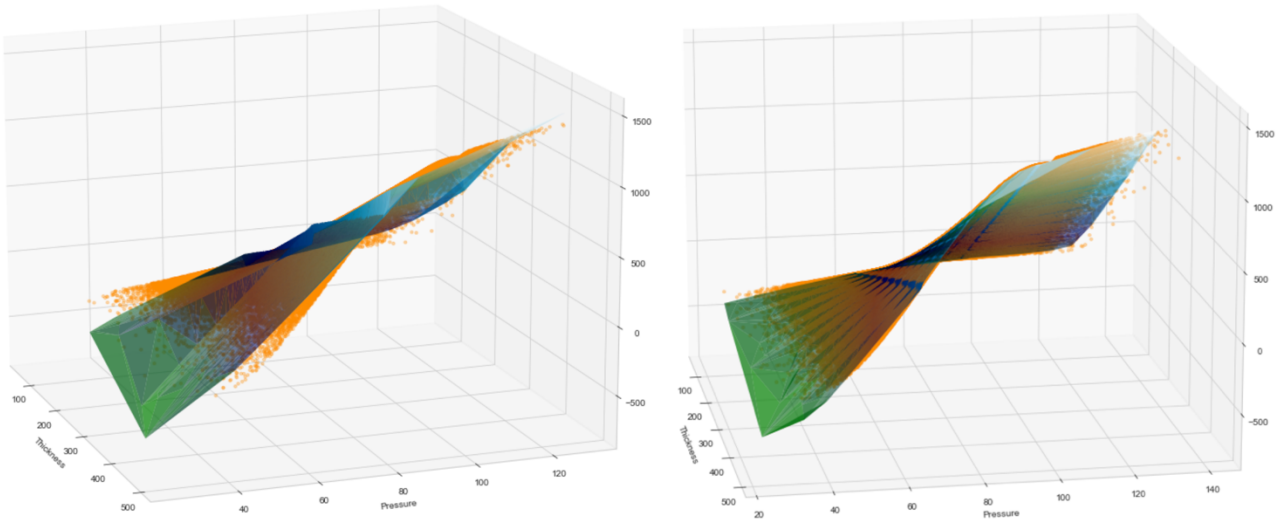
Figure 2.15 - Multivariate regression plots. (a) Piecewise SVR, (b) RF

As shown in Figures 2.13, 2.14, and 2.15 linear regression cannot capture the behavior of the system as well as other regression methods like KNN and XGB. The relationship between the uncertain inputs and the NPV is a curved surface (orange dots) while the regressed surface using linear regression is a linear and flat surface. However, Piecewise OLS with 4 splits improves the result of simple OLS. In Table 2.2, the predicted results are compared with the value discretization method and R^2 scores confirm the conclusion based on the figures. SVR is weak in capturing the tail values as shown in Figures 2.12 and 2.15, which are important in VOI analysis. We should also mention the difference between the expected values without information in two cases where only thickness is uncertain in Table 2.1 and where thickness and pressure are both uncertain in Table 2.2. This is the result of including pressure uncertainty in the system which is material and influences the NPV. In addition, VOI is not additive (Bratvold and Begg, 2010), which means that the VOIs achieved by value discretization for thickness and pressure are 16.79 and 3.61, respectively but the VOI for having both at the same time is 24.37 which is higher than the sum of 3.61 and 16.79.

Table 2.2 - VOI analysis of multivariate regressions with uncertain thickness and pressure

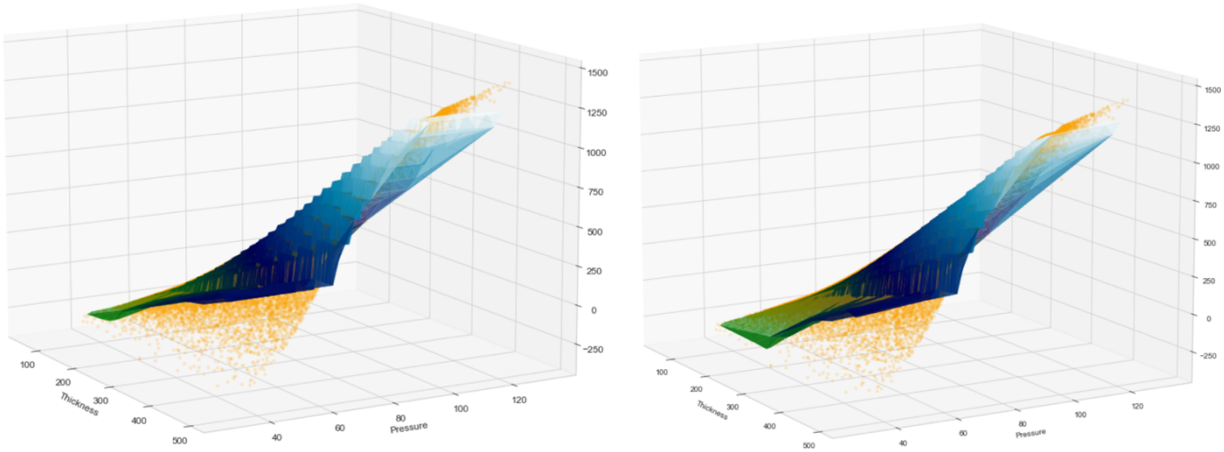
	Value Discretization 70 values	*OLS*	*OLS Piecewise 10 splits*	*XGB 5 max depth*	*KNN K=50*	*RF 70 max leaf nodes*	*SVR Piecewise C=40 10 splits*
EVwoI	353.73	353.92	353.92	353.92	353.92	353.92	353.92
EVwI	378.10	373.67	376.42	376.46	376.39	375.26	375.75
VOI	24.37	19.75	22.50	22.54	22.47	21.34	21.83
R^2 *scores*	-	0.93	0.99	0.99	0.99	0.96	0.99
Run Time (s)	319.5	2.1	7.3	418.7	1.2	2.8	396.8
VOI Error	-	19.6%	7.6%	7.5%	7.7%	12.4%	10%

In addition, the running time for these two approaches (value discretization and simulation-regression) can also be informative for evaluating the simulation-regression approach. Value discretization becomes computationally demanding with an increase in the number of uncertainties and discretized value points. With the same sensitivity analysis on the number of value points in value discretization we had earlier, 70 as the number of values is selected as the correct answer. With 2 uncertainties and 70 value points for each, there would be 70×70 data points. For cases with more than 2 uncertainties, like the CO_2 storage capacity estimation problem with 6 uncertainties, or cases with more value points, VOI calculation becomes highly computationally demanding. Thus, when the number of uncertainties increases, we can find the best regression model fitted using 10-fold cross-validation and assess the running time before conducting VOI analysis with the simulation-regression approach. All these methods are comparable with each other when cross-validation has been done to find the best possible fit for each method. For example, SVR with 2 different values for its hyperparameter (a) $C = 4$, and (b) $C = 40$ yields different results as shown in Figure 2.16. As we can see, the model fit is better with a higher *C* and the regression improves. Thus, we cannot compare the best case of one regression method with a medium accurate case for another.



(a) (b)

Figure 2.17 - The sensitivity analysis on number of splits in Piecewise OLS Regression



(a) (b)

Figure 2.16 - The sensitivity analysis on hyperparameter C for Piecewise SVR

Figure 2.17, shows the sensitivity analysis of the number of splits in Piecewise OLS with (a) 2 and (b) 10 splits, with an increase in the number of splits, the model gets the shape of the data better. The number of splits is chosen by 10-fold cross-validation. First, we split the data set with the given number of splits. Second, we conduct a regression on each split and calculate the R^2 scores. Then, we increase the number of splits and repeat the previous steps to find the best fit which in

this case is 10 splits. Increasing the number of splits by more than 10 is not improving the R^2 scores in this case and causes overfitting.

2.7 Value of Flexibility (VOF)

Flexibility can bring value in 4 general situations (Begg, et al., 2002): 1- when it has less cost than information gathering. 2- gathering the information is not possible. 3- when there is a need to manage the residual uncertainty after having information. And 4- when it actually creates value. Flexibility is combined with creativity and brings new solutions to the problem. Flexibility provides the opportunity to learn from intermediate outcomes and then apply this learning in making the decision. This is dynamic decision making which differs from static decision making where there is no such learning. It can create value since the main decision can be split into multiple steps where the decision at each step uses the learning from the outcomes of the previous decisions (Begg, et al., 2002).

As mentioned, decision making can be static or sequential (dynamic). Similarly, the information-gathering scheme can be categorized as being static or sequential (Eidsvik, et al., 2015, Eidsvik, et al., 2017). There are some cases where the information gathering is not done at a single point in time and is not limited to only one decision (Eidsvik, et al., 2015). For example, gathering information relevant for multiple uncertainties at the same time, or making a decision sequentially for each test after receiving information from the previous one. Gathering information sequentially brings flexibility and value to the decision situation since tests for information are having costs (Eidsvik, et al., 2015).

There are three models representing sequential decision making and information-gathering schemes (Modeling for Decision Insights, lecture notes, 2021). 1- myopic model, where we ignore future decisions and information. 2- naïve model, where we only ignore the future information and learning concept, and 3- dynamic decision making, introduce the full model of the decision situation, and there is nothing ignored (Alyaev, et al., 2019, Modeling for Decision Insights, lecture notes, 2021). Each model has its pros and cons. Moving from myopic to dynamic, the calculation becomes more complex, and more value adding. In a dynamic decision making model, with an increase in the number of decision points, decision alternatives at each point, number of uncertainties, and number of the possible outcome of uncertainties, the complexity increases

exponentially, resulting in the curse of dimensionality (Modeling for Decision Insights, lecture notes, fall 2021. Eidsvik, et al. 2017). In this study, we apply a dynamic decision making model to reveal the inherent value of sequential decision making.

For instance, assume there is a case with two uncertainties, $x1$ and $x2$, related to a profit of an investment situation and tests are available for these uncertainties, $y1$ and $y2$, with specific costs for each test, $P1$ and $P2$, respectively. The decision maker can decide if they want information about both uncertainties before investing in the project. This information-gathering scheme is called static where all the information is gained at the same time (Eidsvik, et al., 2015). Another option she has is to include flexibility (Begg, et al. 2002). She might think of obtaining information sequentially. For instance, perform the test for one of the uncertainties now, observe the results of the test, and then decide whether they want to continue with the second test or not (Eidsvik, et al., 2015). This way of thinking can reduce costs as it also considers the case with only one test.

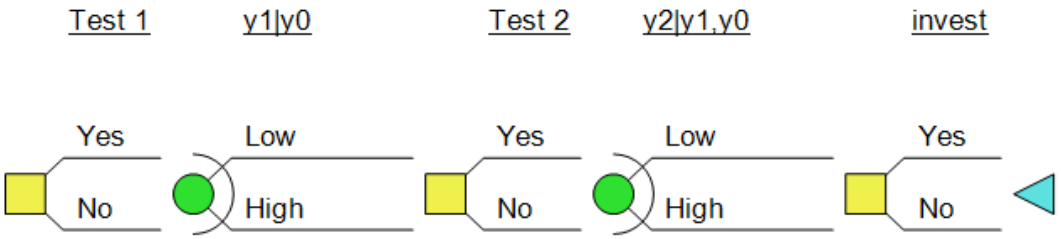


Figure 2.18 - Sequential information gathering schematic

Figure 2.18 illustrates the situation. In the first step, the decision maker observes the results of the first test and decides whether to continue with the second test, or stop gathering information and a make decision to invest or not. Posterior value at this point is:

$$PoV(y_1) = \int_{y_1} \max_{a \epsilon A} \{E(v(x,a)|y_1)\} p(y_1)\, dy_1,$$

Observation of $x1$ affects the decision about whether to observe $x2$. The decision maker should only go on with the second test if the additional value considering its cost, $P2$, is more than the value with having only the first test (Eidsvik, et al., 2015).

$$\int_{y_2} \max_{a \epsilon A}\{E(v(x,a)|y_1,y_2)\}p(y_2|y_1)\,dy_2 \; - \; P_2 \; > \; \max_{a \epsilon A}\{E(v(x,a)|y_1)\},$$

$$Continue\ testing \; - P_2 \; > \; Stop\ testing,$$

Thus, the PoV of the sequential information gathering with having only two uncertainties is calculated by:

$$PoV_{seq}(y_2|y_1) = \int \max \left\{ \begin{matrix} \int_{y_2} \max_{a \epsilon A}\{E(v(x,a)|y_1,y_2)\}\, p(y_2|y_1)dy_2 - \; P_2, \\ \max_{a \epsilon A}\{E(v(x,a)|y_1)\} \end{matrix} \right\} p(y_1)dy_1,$$

Then, the VOI must be calculated using PoV and PV. This VOI should be compared with the cost of test 1, P_1. If VOI is larger than the cost P_1, the decision maker should choose this sequential information gathering scheme. This approach can also be extended to different available test sequences. For example, start with test 1, continue with test 2, or start with test 2 and continue with test 1.

In this study, we used flexibility in information gathering using the simulation-regression method to assess the VOI. We considered all possible sequences for information gathering to find the best information gathering sequence. For example, for a situation where we have 3 uncertainties x_1, x_2, x_3, and 3 tests available, y_1, y_2, y_3, there are $3! = 6$ possible sequences for information gathering, $y_1y_2y_3$, $y_1y_3y_2$, $y_2y_1y_3$, $y_2y_3y_1$, $y_3y_1y_2$, and $y_3y_2y_1$. Consider the first sequence, $y_1y_2y_3$. The sequential decision making of this case is illustrated in Figure 2.19. In this study, for CO_2 storage capacity estimation, we calculate value of flexibility with tests for thickness, pressure, and porosity since they are material uncertainties.

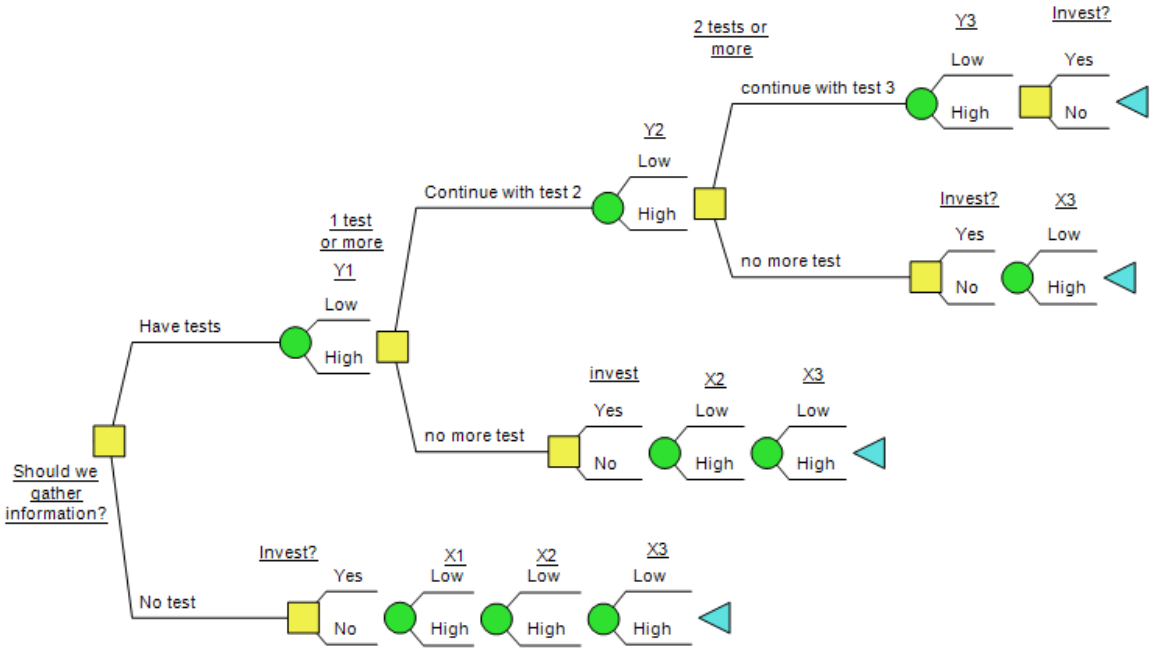


Figure 2.19 - Sequential information gathering decision tree

Therefore, we repeat this decision making process in Figure 2.19 for all other sequences as well and find the VOI for each of them. Lastly, the last step would be comparing these VOIs to find the best sequence as shown in the diagram in Figure 2.20.

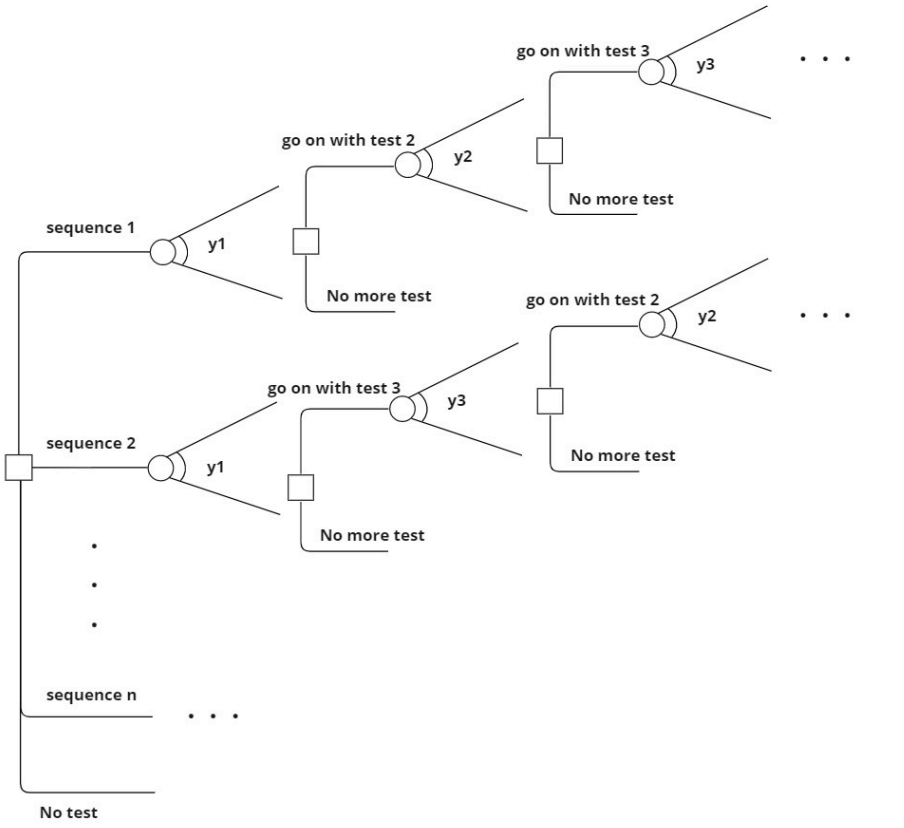


Figure 2.20 - Illustration of sequential information gathering scheme

2.8 Conclusions

In this chapter, we introduced the value of information analysis, the value discretization method, and the simulation-regression approach. We introduced different regression methods and evaluated their ability to predict different data types in VOI analysis by simulation-regression method.

We evaluated the simulation-regression's accuracy in VOI calculations influenced by the regression methods and the number of Monte Carlo samples. The simulation-regression method's accuracy in VOI calculations depends on the regression method selected. The regression method selected for the simulation-regression method must be evaluated based on the data before VOI calculations, and we introduced R^2 scores as a great means to find the best fit. Thus, in order to have the best accuracy possible for the simulation-regression method, one should find the best fitted model for specific data.

In addition, since the simulation-regression method uses MCS, and MCS is sensitive to the number of samples, a large number of samples is needed to increase the accuracy of the VOI calculation.

Chapter 3 – Case Study at Utsira Formation

In this chapter, we introduce a workflow for the simulation-regression approach based on the conclusions of the previous chapter, frame a sequential decision making situation for the Utsira formation CO_2 capacity storage estimation with related uncertainties, implement our workflow on the case, and calculate the VOI and VOF.

3.1 Utsira Formation Reservoir Model

In CCS projects, CO_2 is being stored in deep underground geological storage, deep ocean storage, and mineral carbonation (Aminu, et al., 2017). According to Allen, et al. (2018), Utsira is a saline aquifer that has the largest storage capacity compared with other geologic formations (Temitope, et al., 2016) with an average top-surface depth of almost 900 meters (ranging from 300 to 1400 meters) and a critical point of 31°C and 73.8 bars.

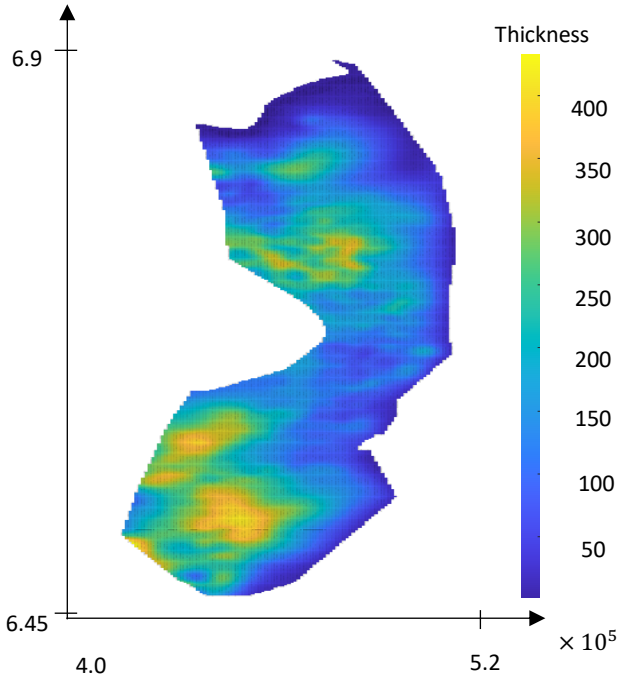


Figure 3.1 - Utsira Formation figure with thickness range. (Source: MRST, SINTEF, 2016b, Lie, 2019)

There are some parameters affecting the capacity estimation of saline aquifers for CO_2 storage such as porosity, pressure, temperature, thickness, and depth. Porosity is one of the important rock properties for estimating the CO_2 storage capacity. The average porosity of Utsira is about 0.2112

(Norwegian Petroleum Directorate, 2019). Porosity samples in this study are represented by a Gaussian distribution with a standard deviation of 0.04. Aquifer initial conditions such as temperature and pressure influence the density of CO_2, and this impacts the storage capacity. Warmer aquifers decrease CO_2 density resulting in lower CO_2 storage volume. Pressure and temperature are calculated using depth (Allen, et al., 2018):

$$P = (\rho_w g z + P_s),$$

$$P_0 = P + \frac{\sum_i^n pv_i P_i}{\sum_i^n pv_i} \frac{d}{100},$$

$$T = T_b + \nabla T(z - z_b),$$

Where P_0 is initial hydrostatic pressure, ρ_w water density, g gravitational acceleration, z caprock depth, P_s surface pressure, pv pore volume, d is the deviation in percent, T is initial caprock temperature, T_b is seafloor temperature, $(z - z_b)$ is depth below seafloor, and ∇T is the thermal gradient in the vertical direction. Pressure samples are produced with a mean and standard deviation of [0,5]% deviation (d) which means a standard deviation of 12 bars as the model reference pressure is 80 bars. Also, the samples of temperature are produced with sampling the thermal gradients (∇T) with mean and standard deviation of [37.5, 3.36] °C/km.

According to the Norwegian Petroleum Directorate (NPD), the Utsira formation has an average depth of 900 m, a horizontal area of 110 km and 430 km in the vertical direction, with a storage efficiency factor of 4%.

MATLAB Reservoir Simulation Toolbox (MRST)

MRST is an open-source software implemented in MATLAB for reservoir simulation and modeling. It includes simulation tools, datasets from real aquifers such as Utsira, and examples. But it is also possible to create your own simulations with specific values (Lie, 2019). In this study, we used available datasets for the Utsira formation to quantify uncertainties' ranges and averages.

3.2 Decision Frame

In this study, we assume a company that has an opportunity to invest in a CO_2 injection and storage project at the Utsira formation. There are two alternatives $a \in \{0,1\}$, invest in the project $(a = 1)$, and do not invest $(a = 0)$ for the company. The goal is to maximize the Net Present Value (NPV). In this case, we assume the CO_2 storage is a commercial service of the company, and NPV is the profit of the company calculated from its revenue and costs. In other cases, the company that produces CO_2 emissions must pay a carbon tax. Carbon tax is like a charge that governments require CO_2 emitters to pay for each ton of greenhouse gas emissions they produce. In Norway, carbon tax started in 1991 (Jason, 2013). Therefore, the profit can be considered as a carbon tax reduction. The injection period in this study is considered 40 years with equal injection rates based on the estimated capacity. In addition, the trapping capacity of the formation and the capturing costs are uncertain.

3.2.1 CO_2 Storage Capacity Estimation and Uncertainties

As discussed by Aminu, et al. (2017), CO_2 storage capacity estimation is not easy. There are some methods suggested to calculate capacity such as the Carbon Sequestration Leadership Forum (CSLF). The CSLF model is a volumetric approach and is calculate as:

$$M_{CO_2} = AH\varphi\rho_{CO_2}(1 - S_{wirr})C_c$$

Where A is trap area, H average thickness, φ porosity, C_c capacity coefficient, S_{wirr} irreducible water saturation and $\rho_{CO_2}$is CO_2 density. $(1 - S_{wirr})C_c$ is equal to storage efficiency factor. CO_2 density has a non-linear relationship with temperature and pressure which we use as uncertainties in our model. In this study, we estimated the storage capacity of the Utsira formation with porosity, depth, thickness, temperature, and pressure of the rock as uncertainties.

According to the NPD, CO_2 that is being injected in Utsira is in supercritical liquid form. CO_2 density behavior needed to be implemented to account for the influence of temperature and pressure since they are important features in determining the CO_2 storage capacity. Bahadori, et al. (2009) has introduced a new model to predict CO_2 density as a function of temperature and pressure. In his suggested correlation, ρ, P, T , are density, pressure, and temperature respectively and are presented as:

$$\rho = \alpha + \beta T + \gamma T^2 + \theta T^3.$$

Where,

$$\alpha = A_1 + B_1 P + C_1 P^2 + D_1 P^3,$$

$$\beta = A_2 + B_2 P + C_2 P^2 + D_2 P^3,$$

$$\gamma = A_3 + B_3 P + C_3 P^2 + D_3 P^3,$$

$$\theta = A_4 + B_4 P + C_4 P^2 + D_4 P^3.$$

All coefficients A_1,… A_4, B_1,..., B_4, C_1,..., C_4, D_1,..., D_4 are constant but varies in 2 different pressure zones of 25-100 bar and 100-700 bar. In this study, we used this model to consider the effect of pressure and temperature as uncertainties.

3.2.2 Decision Frame and Influence Diagram

To define a decision frame, we determine NPV, uncertainties affecting it, and decision points. With these all, the influence diagram of the case study can be shown as below where uncertainties are temperature, pressure, thickness, porosity, depth, and storage cost, which all are impacting the NPV.

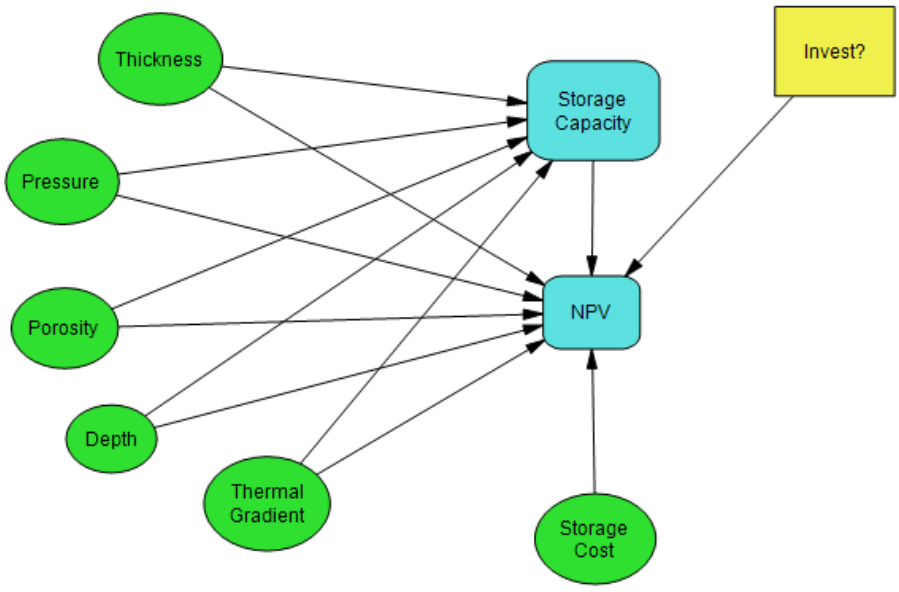


Figure 3.2 - Influence Diagram of CO_2 capacity estimation case study

To convert storage capacity to value and do the VOI calculations, price of CO_2 and total cost, capital expenditure (CAPEX) and storage cost, are needed. The annual profit is gained by average carbon price of the first quarter of 2021, which is \$38/tCO_2 (Quandl 2021), minus total cost. In Zero Emissions Platform 2011 document, CAPEX for an offshore saline aquifer without re-usable wells is \$330 million and cost of storage is \$[6, 20]/tCO_2 (Zero Emissions Platform, 2011). We assumed area of injection to be 100 km^2 and discount rate of 10% per year. Using CO_2 injection capacity (M_{CO_2}) and costs, NPV is calculated as:

$$NPV(x, a = 0) = 0,$$

$$NPV(x, a = 1) = \frac{revenue - cost}{(1+i)^n} - CAPEX$$

where,

$$revenue = \frac{\$38}{tCO_2} \times \frac{M_{CO_2}}{n},$$

$$cost = \$(cost_{offshore\ storage})/tCO_2 \times \frac{M_{CO_2}}{n},$$

$$n = discounted\ period, i = interest\ rate.$$

3.3 Workflow

After modeling uncertainties and choosing their distributions and ranges, we generate 100,000 random samples of each uncertainty, which we are going to use as inputs in the next step, the simulation-regression. After generating samples of uncertainties, we do choose the most important and dominant uncertainties in NPV value using a Tornado diagram. Then, we consider multiple regression methods to find the model best fitted to the dataset. In this study, we look at OLS, Piecewise OLS, KNN, XGB, RF, and Piecewise SVR as regression model possibilities. We use 10-fold cross-validation to find out which method is a better method for prediction and controlling the potential overfitting of the regression models, e.g., *k* neighbors in KNN. We use the value discretization method to compare the results of the regression with the correct VOI value, assuming as discussed in Chapter 2 that we have used enough values in the value discretization method to make it correct within the required number of decimal digits. After identifying the most suitable model to fit the dataset, we calculate the value of perfect and imperfect information. Then, we

move on and consider a sequential information gathering scheme to bring flexibility to the decision making and calculate the value of flexibility (VOF), and lastly, we implement sensitivity analyses on VOF.

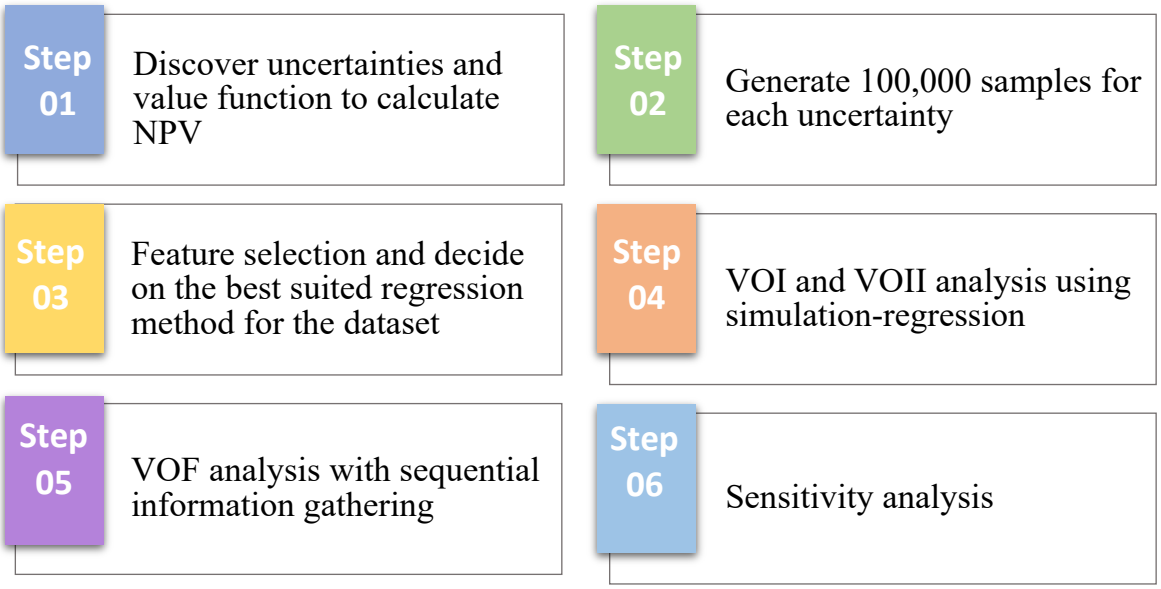


Figure 3.3 - Workflow of the CO_2 capacity estimation case study

3.3.1 Generate samples

To start with regression, after determining uncertainties, their value ranges, and value function, we generated 100,000 samples. We used vectorization from the Numpy library for samples and calculated the NPVs related to them. This approach reduced the running time of the code significantly. Time taken by list comprehension is 180 seconds per loop with 7 runs and 1 loop for each, and time taken by Numpy vectorize method is 6 seconds per loop with 7 runs and 1 loop for each.

3.3.2 Feature Selection

Feature selection techniques provide insight into the data and the model using sensitivity analysis on ranges of output value with change in one variable at a time (Bratvold and Begg, 2010), and are useful for dimensionality reduction. In this study, a Tornado diagram is implemented on the value function to find the features with the highest impact on the NPV as a feature selection

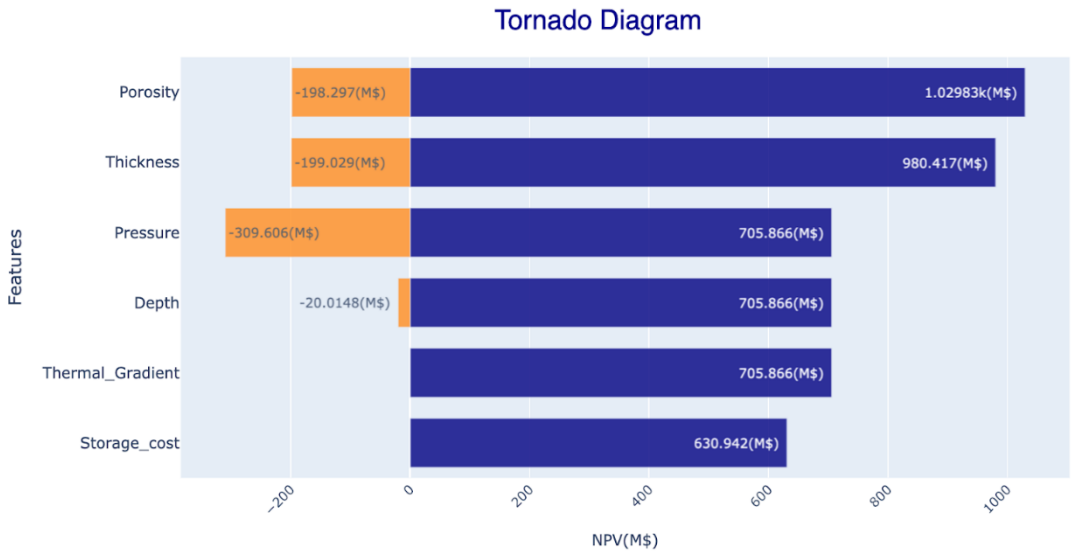


Figure 3.4 - Tornado Diagram, features vs NPV

method. This chart, shown in Figure 3.4, provides a means for assessing the NPV impact of each of the underlying variables. As seen in Figure 3.4, pressure, porosity, thickness, and depth are material as they can change the decision by influencing the NPV to give negative values. Moreover, as depth has only a very small chance of being material, we focus on the top three variables in the Tornado diagram - porosity, thickness, and pressure - in our analysis. Note that no conclusion of the VOI for the variables can be made based on the NPV impact indicated in the Tornado graph. The Tornado graph as shown here is a deterministic analysis tool as no probabilities have been assigned to the possible outcomes and, hence, VOI analysis cannot be conducted.

3.3.3 Regression Method Selection

In previous step, we selected thickness, pressure, and porosity as uncertain variables and will use the average values for the remaining variables. Then, R^2 scores of different regression methods

with these three uncertainties are compared. Table 3.1 includes VOI, run time, VOI percentage error, and R^2 scores for each regression. In addition, cross-validation on hyperparameters for each regression is done. Therefore, each method is at its best possible fit. We use value discretization results with 30 value points for each uncertainty as the correct answer which is also included in the table.

Table 3.1 - VOI analysis of multivariate regressions with uncertain thickness, pressure, and porosity

	Value Discretization 30 values	*OLS*	*OLS Piecewise 10 splits*	*XGB*	*KNN*	*RF*	*SVR Piecewise 10 splits*
EVwoI	353.44	353.20	353.20	353.20	353.20	353.20	353.20
EVwI	379.86	376.68	380.54	380.48	375.92	376.73	378.00
VOI	26.42	23.48	27.34	27.28	22.72	23.53	24.80
R^2 *scores*	-	0.91	0.99	0.99	0.87	0.97	0.98
VOI Error	-	11.1%	3.4%	3.2%	14%	10.9%	6.1%
Run time	12,108	0.3	13.5	422	0.5	2.7	917
Hyperparameter	-	-	-	7	100	90	50

XGB regression is chosen as the best model fitted in this case (uncertain thickness, pressure, porosity) based on its high R^2 score and the VOI calculated of 27.28 with a 3.2% error. However, there should be a balance between VOI calculation accuracy and running time. In some cases, there might be a limit in running time which must be considered while selecting the best regression model for a specific dataset. In this case, we assume we have no limitation in time and 422 seconds for our decision is reasonable for our case and also it is less than the value discretization method which is 12,108 seconds. Thus, XGB is selected as the best regression method for our dataset with 3 uncertainties.

3.3.4 VOII Analysis

The VOI analysis with perfect and imperfect information with $noise = \{10\%,\ 20\%,\ 30\%,\ 50\%,\ 70\%,\ 100\%\}$ is done on data for four information gathering schemes where we receive information on:

1. All three uncertainties (thickness, pressure, and porosity),

2. Information on thickness,
3. Information on pressure,
4. Information on both thickness and pressure.

Noise is implemented with a normal distribution and as a measurement error before regression, affecting the R^2 scores as shown in Figure 3.5.

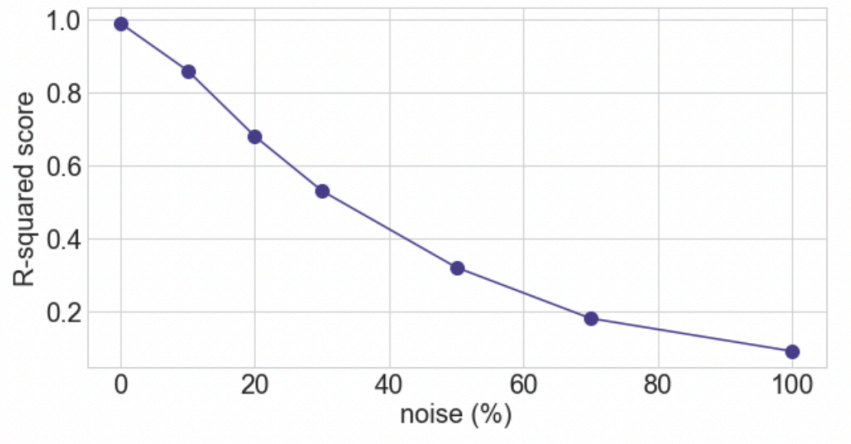


Figure 3.5 - The sensitivity analysis of noise vs R^2 score

XGB regression method is chosen due to its high R^2 score which has been calculated previously for the three uncertainties case in Table 3.1, and for two uncertainties case based on Table 2.2. OLS is selected for the case with only thickness as uncertainty based on Table 2.1, and SVR for uncertain pressure based on Figure 2.8. As mentioned earlier, the regression with

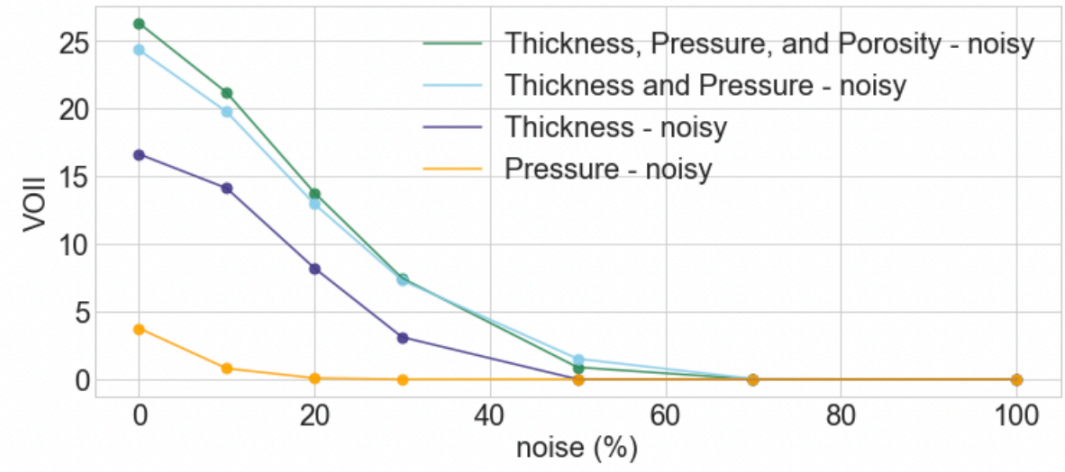


Figure 3.6 - The sensitivity analysis of VOII vs noise

imperfect information is repeated 100 times. As we can see from Figure 3.6, the VOI drops when noise is included in the information.

3.3.5 VOF Analysis

The VOF analysis for this study is done using uncertainties in thickness, pressure, and porosity. As mentioned earlier, in this example information is gathered sequentially which allows for flexibility and increases the value of information (Begg, et al. 2002). The VOF cannot be less than the static case with all information at once. Sequential information gathering can bring value when there are costs for available tests related to uncertainties and we split the decision for gathering information into multiple steps. In addition, in this study, we implemented a sensitivity analysis on test cost to observe its effect on VOF and VOI.

First, the static case (have all tests at once) is implemented with XGB regression since it is selected as the best regression for these three uncertainties in Table 3.1. This case has a value of free information of 27.28. The costs of the tests must be subtracted from this value to find the VOF with costs. Then, based on the sequential information gathering structure shown in Figures 2.19 and 2.20, there are multiple regressions needed in the process which are chosen based on the dataset. Cross-validation is done on hyperparameters of each regression method to find the optimal fitted model. Then, R^2 scores of all regression methods are compared to find the optimal method among them. These 2 steps are done in each step of the sequential information gathering process. There are 6 possible sequences for sequential information gathering with 3 available tests. The number of possible sequences grows with the number of tests available. With n information sources (tests), $n!$ sequences are possible. For each sequence, depending on the number of alternatives $a \in A$, and the number of features n, there will be $(a+2)n-2$ regressions required. Finally, 10 regressions are needed in our case study where two are 3-variates, four are 2-variates, and four are univariates. In this study, after the cross-validation steps mentioned, XGB for 3-variates with $max\ depth = 7$, Piecewise OLS for 2-variates, and OLS for univariates are selected based on their high R^2 scores and running time. R^2 scores of all regression methods at all steps resulting in these selections are presented in Appendix 3.

After selecting the best models fitted for each step in the sequential information gathering, a sensitivity analysis is done on the impact of tests cost on the VOF. Four different cost sets are used for information about thickness, pressure, and porosity. First, the test costs for thickness and pressure are kept constant at \$8 and \$2 million respectively. The sensitivity analysis is done on the cost of information about porosity with values of \$2, \$8, \$15, and \$40 million. The optimal sequence including the cost scenarios is thickness-pressure-porosity. In Figure 3.7 costs are represented with this sequence. For instance, a cost set of [8, 2, 15] is [thickness test cost of \$8, pressure test cost of \$2, porosity test cost of \$15] million. VOI of the static case is also $\$27.28\ million - (Total\ cost\ of\ all\ tests)$ since we are having information about all uncertainties at once. For instance, with [8, 2, 2], $VOI_{static} = 27.28 - (8 + 2 + 2) = 15.28$.

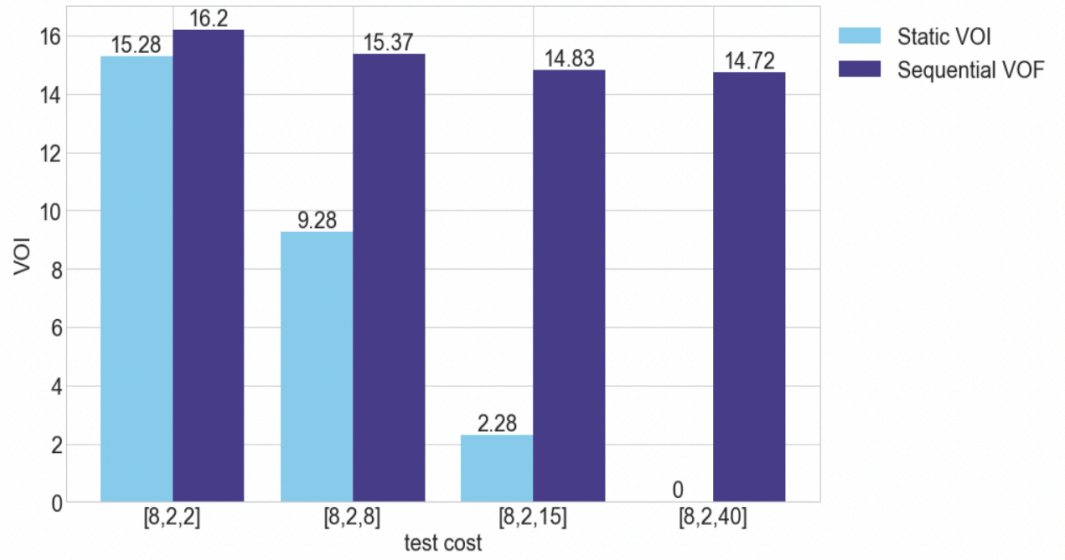


Figure 3.7 - VOF analysis with sensitivity analysis on tests costs

In Figure 3.7, different costs for porosity and its effect on decisions are shown. According to this figure and Bar plots of decisions in Figure 3.8,

1- VOF is higher than VOI with static information gathering.

2- Differences between VOF and static VOI increases with an increase in costs which means that with expensive tests, flexibility brings more value because the decision maker can say no to further information gathering when the cost of the information is higher than its value.

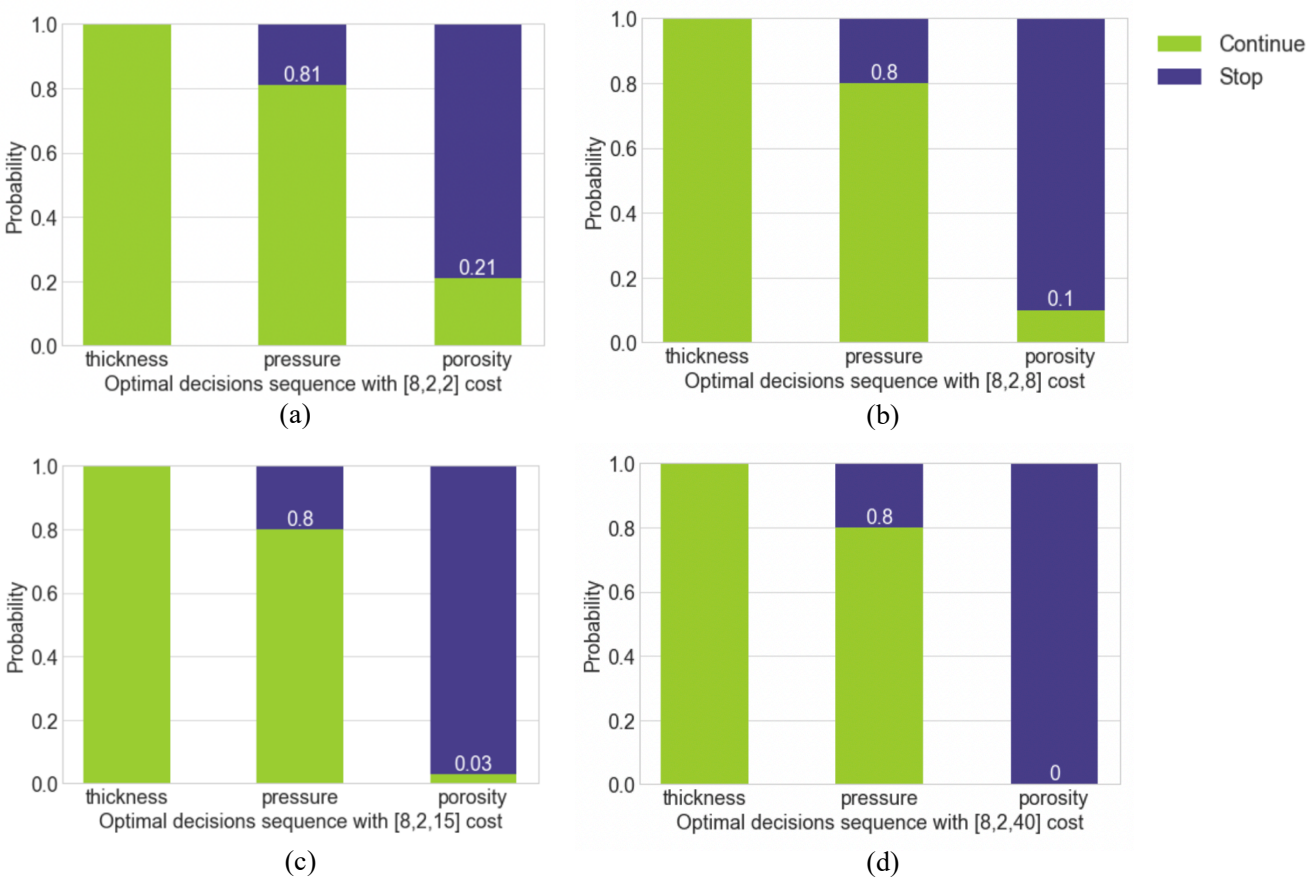


Figure 3.8 - The sensitivity analysis on tests costs vs decisions

VOF results can also show which information is having the most effect on NPV. For example, by comparing 2 different sequences with costs set of [2, 2, 2], for thickness, pressure, and porosity respectively, the best sequence is thickness-pressure-porosity, with decisions of 1- "having information" in all samples, 2- "continue to have more than one test" in 80% of samples, and 3- "have all three tests" in 20% of samples. But in another sequence with pressure-porosity-thickness, decisions are 1- "having information" in all samples, 2- "continue to have more than one test" in 95% of samples, and 3- "have all three tests" in 95% of samples. This means if we choose this sequence, with a probability of 95% we should continue with the 3rd test as well which is for thickness, and we should pay for all tests. This difference shows how thickness is effective in NPV and its information is bringing the highest value to the decision making.

Therefore, a sensitivity analysis is done on the cost of thickness to find at which cost of thickness test, the best sequences changes. As shown in Table 3.2, the best sequence of gathering information changes when the cost of information increases between $10 and $15 million.

Table 3.2 - Senitivity analysis on tests costs vs best information gathering sequence

Costs	*Best Sequence*			*VOF*
[2, 2, 2]	**Thickness**	Pressure	Porosity	22.21
[10, 2, 2]	**Thickness**	Pressure	Porosity	14.21
[13, 2, 2]	Pressure	**Thickness**	Porosity	11.38
[15, 2, 2]	Pressure	Porosity	**Thickness**	9.94

Results of the sequential information gathering sequence depend very much on the regression method for which overfitting and underfitting should be avoided. In this study, we also considered 2 different cases of sequential information gathering with:

1- Specifying the regression methods at each step and having a mixture of methods.

2- Conducting all the steps with only one regression method.

The first case is the case we have discussed up until now.

For the case with using only one regression method in sequential decision making, costs for thickness, pressure, and porosity are $2, $2, and $2 million respectively. We have calculated the VOF with these costs in the previous section as 22.21 with specifying regression methods for each step in Table 3.2. Now, we consider cases with not specifying a regression method for each step, and use only XGB, OLS, Piecewise OLS, and RF for all the steps in sequential information gathering. As mentioned, we have two 3-variate (3V), four 2-variate (2V), and four univariate (1V) regressions in our sequential information gathering case. The alternative "Not Invest" values in our case are zero, and the regressions for this alternative have the highest possible R^2 score of 1.00, no matter which regression method we choose. According to the sequential information gathering decision tree in Figure 2.19, one 3-variate, one 2-variate, and one univariate of 10 regressions needed are regressing the "Not Invest" alternative values. Thus, in this study, we do not include them in the regression method selections, which are colored red in Table 3.3.

In Figure 3.9 and Table 3.3, R^2 scores of each regression in the sequential information gathering is presented with test cost of $[2,2,2] million. The optimal information gathering in this sequence is thickness-pressure-porosity as shown in Table 3.2. "2V_3" in Figure 3.9 and Table 3.3 is the third 2-variate regression of the sequential information gathering, and these orders are the same order in solving the decision tree in Figure 2.19. As we can see from Figure 3.9 and Table 3.3, in mixing regression methods case R^2 scores of all regressions are higher than other cases, and as we discussed in Chapter 2, the higher the R^2 scores, the more accurate the result of the regression.

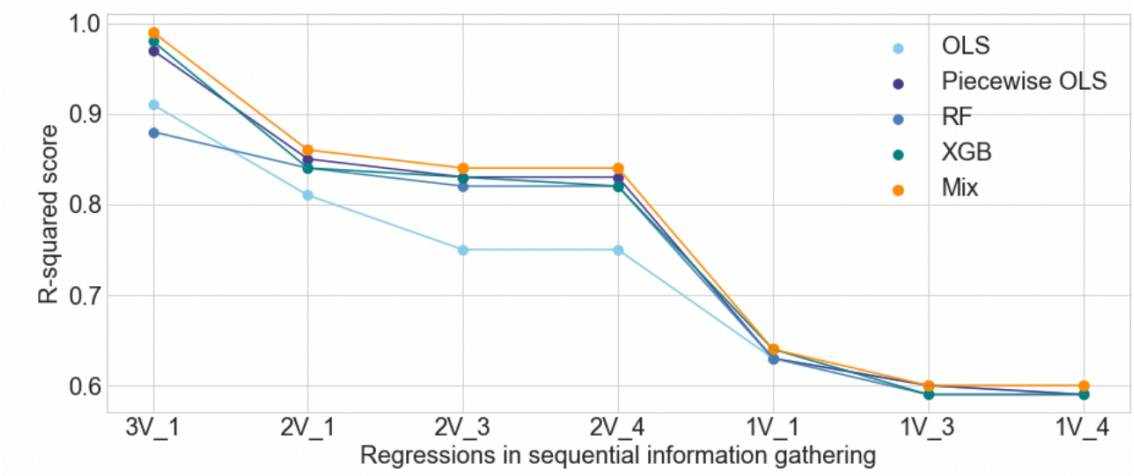


Figure 3.9 – Sequential information gathering with different regression method selection scenarios - cost $[2,2,2] million

Table 3.3 - R^2 scores of sequential information gathering - cost $[2,2,2] million

	Mix	*Piecewise OLS*	*OLS*	*RF*	*XGB*
3V_1	0.99	0.97	0.91	0.88	0.98
3V_2	1.00	1.00	1.00	1.00	1.00
2V_1	0.86	0.85	0.81	0.84	0.84
2V_2	1.00	1.00	1.00	1.00	1.00
2V_3	0.84	0.83	0.75	0.82	0.83
2V_4	0.84	0.83	0.75	0.82	0.82
1V_1	0.64	0.63	0.63	0.63	0.64
1V_2	1.00	1.00	1.00	1.00	1.00
1V_3	0.60	0.60	0.60	0.59	0.59
1V_4	0.60	0.59	0.59	0.59	0.59

In addition, going step by step in sequential information gathering, the number of features is decreasing, which as mentioned before, means the regressed model is getting underfitted and would perform poorly since it needs other uncertainties as well to fit a model, and therefore, the R^2 scores are getting worse step by step, which we can see from Figure 3.9.

In Table 3.4, the percentage errors of VOF calculation with different regression method selections are provided. VOF with specifying regression methods for each step (Mix of regressions) is the correct answer for comparison purposes. There will be some errors in VOF calculation when the regression method is not selected based on the data type of each step. In summary, regression methods for each step of the sequential information gathering must be selected in a way that they build the best model fitted.

Table 3.4 - VOF analysis with different regression method selection scenarios- cost $[2,2,2] million

	Mix	*All OLS*	*All Piecewise OLS*	*All XGB*	*All RF*
VOF	22.21	18.66	22.46	21.38	20.76
VOF Error	-	15.6%	1.1%	3.7%	6.5%

There are some critical points in costs that the optimal sequence changes in sequential information gathering. For example, $[14,2,2] is one of these critical points in our case. With costs between $12 and $14 million about thickness, the optimal sequence is pressure-thickness-porosity, and for costs more than $14 million, the optimal sequence changes to pressure-porosity-thickness. At critical points like this, we might have different results if we do not specify the optimal regression method for each step. Thus, we also consider another case with a cost set of [14, 2, 2], which results are in Table 3.5.

Table 3.5 - VOF analysis with different regression scenarios with cost set of $[14,2,2] million

	Optimal sequence	*VOF*	*VOF Error*
Mix	pressure – thickness - porosity	10.59	-
All OLS	pressure - porosity - thickness	8.99	15%
All Piecewise OLS	pressure – thickness - porosity	10.84	2.3%
All XGB	thickness - pressure - porosity	9.38	11.4%
All RF	pressure – thickness - porosity	9.22	12.9%

As shown in Table 3.5, in this case, when we used only XGB or OLS for all 10 regressions, the best sequence is different from the case where we specified regression methods for each step. Therefore, in critical points, where the optimal sequence changes, it is important to use the optimal regression method at each step of sequential information gathering to achieve accurate and correct results in VOF calculation and optimal sequence selection.

3.4 Summary

Simulation-regression approach can approximate the expected value with and without information to calculate VOI and its accuracy is highly dependent on the regression method, the number of samples, and the linearity or non-linearity of the relation between uncertainties and the value function. Therefore, *k*-fold cross-validation is a necessary step to choose the best model fitted and to control the hyperparameters of regression models to avoid overfitting and underfitting.

Each regression method has advantages and disadvantages based on the dataset. Simulation-regression approach can handle a large number of uncertainties in the model while this can be extremely complex with decision trees and Bayes' rule.

In cases with sequential information gathering, an optimal regression method must be chosen for each step to calculate VOF to ensure accuracy in VOF calculation and the optimal sequence selected. Lastly, simulation-regression approach has the potential to reduce the curse of dimensionality with an approximation of VOI relative to a decision tree calculated VOI.

Chapter 4 - Conclusions and Recommendations

In this research, we have proposed a workflow for using the simulation-regression method to assess VOI and VOF that can be implemented in many decision making situations including CO_2 storage capacity estimation. The workflow includes (1) identifying uncertainties, (2) generating samples, (3) selecting the most important features, (4) selecting the regression method suited to the data, and (5) approximating expected value with and without information to calculate the VOI.

In this study, we compared six different regression methods and showed how they can affect the VOI and VOF. We also illustrated the importance of selecting the best model to fit in order to achieve satisfactory accuracy in the calculations. We used value discretization as a means for comparison and to evaluate the advantages and disadvantages of the simulation-regression approach. However, there are many other machine learning regression methods, e.g., moving window, piecewise regression which we only implemented briefly on OLS and SVR regression, and neural networks that could also be implemented to find the best model for VOI analysis. We implemented the workflow for assessing the VOI for CO_2 storage capacity estimation with a small number of uncertainties but the simulation-regression approach is perhaps most powerful when working with a larger number of uncertainties.

The analysis and workflow can be implemented for VOI and VOF calculations in real-world investment situations with a large number of uncertainties for which classical methods like decision trees or value discretization are suffering from the curse of dimensionality. This analysis and workflow are not limited to this field and can be used in any decision making situation in all fields.

References

Quandl, 2021, ECX EUA Futures. Available at: https://www.quandl.com/data/CHRIS/ICE_C1, [Accessed April 2022].

Agami Reddy, T., 2011. Estimation of Linear Model Parameters Using Least Squares. In *Applied Data Analysis and Modeling for Energy Engineers and Scientists*. Boston, MA: Springer US, pp. 141–182.

Allen, R. et al., 2018. Using simplified methods to explore the impact of parameter uncertainty on CO2 storage estimates with application to the Norwegian Continental Shelf. *International journal of greenhouse gas control*, 75, pp.198–213., doi:10.1016/j.ijggc.2018.05.017.

Alyaev, S. et al., 2019. A decision support system for multi-target geosteering. *Journal of petroleum science & engineering*, 183, p.106381, doi: 10.1016/j.petrol.2019.106381.

Babajide Mustapha, I. Saeed, F., 2016. Bioactive Molecule Prediction Using Extreme Gradient Boosting., *Molecules* 21(8): 983, doi: 10.3390/molecules21080983.

Bahadori, A., et al., 2009. New correlations predict aqueous solubility and density of carbon dioxide. *International journal of greenhouse gas control,* 3(4), pp. 474-480.

Bakshi, C., 2020. Random Forest Regression., *Level Up Coding*, Available at: https://levelup.gitconnected.com/random-forest-regression-209c0f354c84., [Accessed 30 April 2022].

Begg, S., et al. 2002. The Value of Flexibility in Managing Uncertainty in Oil and Gas Investments., *Proceedings - SPE Annual Technical Conference and Exhibition.,* doi: 10.2118/77586-MS,

Bratvold, R. B., et al., 2010. Making good decisions. Richardson, Tex, Society of Petroleum Engineers.

Bratvold, R. B., et al., 2009. Value of Information in the Oil and Gas Industry: Past, Present, and Future., SPE reservoir evaluation & engineering, 12(4), pp. 630-638.

Breiman, L., 2001, Random Forests, *Machine Learning*, 45(1), p. 5–32, doi: 10.1023/A:1010933404324.

Chugh, A., 2020. MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared - Which Metric is Better? *Analytics Vidhya*. Available at: https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e, [Accessed April 2022].

Dutta, G., et al., 2019. Value of information analysis for subsurface energy resources applications., *Applied Energy*, 252, p.113436, doi: 10.1016/j.apenergy.2019.113436.

Eidsvik, J., et al. 2017. Simulation-regression approximations for value of information analysis of geophysical data. *Mathematical geosciences*, 49(4), pp. 467-491, doi:10.1007/s11004-017-9679-9.

Eidsvik, J., et al., 2017. Sequential information gathering schemes for spatial risk and decision analysis applications., *Stochastic environmental research and risk assessment,* 32(4), pp.1163-1177.

Eidsvik, J., et al., 2015. Value of Information in the Earth Sciences, *Integrating Spatial Modeling and Decision Analysis.* Cambridge, Cambridge: Cambridge University Press.

Friedman, J. H., 2001. Greedy function approximation: A gradient boosting machine., *The Annals of statistics,* 29(5): 1189-1232, doi: 10.1214/aos/1013203451.

Grootendorst, M., 2019, Validating your Machine Learning Model. *Towards Data Science,* Available at: https://towardsdatascience.com/validating-your-machine-learning-model-25b4c8643fb7, [Accessed May 2022].

Gupta, P. and N. K. Sehgal (2021). Introduction to machine learning in the cloud with Python : concepts and practices. Cham, Switzerland, Springer.

Kenton, W., 2021. Monte Carlo Simulation. Available at: https://www.investopedia.com/terms/m/montecarlosimulation.asp. [Accessed April 2022].

Lie, K.-A., 2019. An Introduction to Reservoir Simulation Using MATLAB/GNU Octave: User Guide for the MATLAB Reservoir Simulation Toolbox (MRST). Cambridge, *Cambridge University Press.* doi: 10.1017/9781108591416.

Natekin, A., Knoll, A., 2013. Gradient boosting machines, a tutorial. *Front Neurorobot,* 7, pp. 21-21, doi: 10.3389/fnbot.2013.00021.

Nikolaeiv, D., 2021. Overfitting and Underfitting Principle, *Towards data Science*, Available at: https://towardsdatascience.com/overfitting-and-underfitting-principles-ea8964d9c45c, [Accessed May 2022]

Nordbotten, J. M. and M. A. Celia, 2012. Geological storage of CO2; modeling approaches for large-scale simulation. *Reference & Research Book News,* Portland, Portland: Ringgold, Inc. 27(1).

Pedregosa, F., et al., 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research,* 12, pp. 2825-2830.

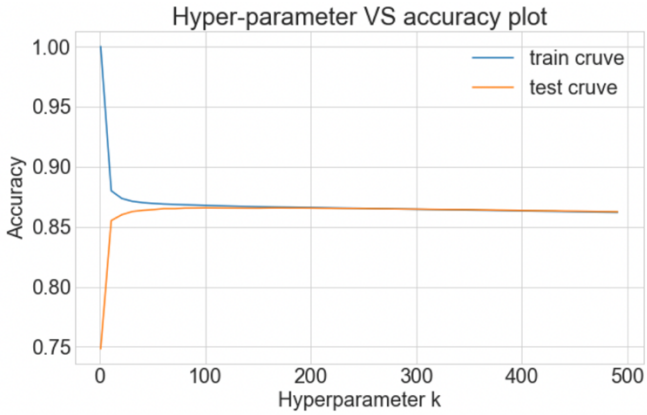
Zero Emissions Platform (Zep), 2011, The Costs of CO2 Storage: Post-demonstration CCS in the EU, *European Technology Platform for Zero Emission Fossil Fuel Power Plants*. Available at: https://zeroemissionsplatform.eu/document/the-costs-of-co2-storage. [Accessed April 2022].

Raj, A., 2020. Unlocking the True Power of Support Vector Regression - Using Support Vector Machine for Regression Problems. *Towards Data Science*. Available at: https://towardsdatascience.com/unlocking-the-true-power-of-support-vector-regression-847fd123a4a0. [Accessed April 2022].

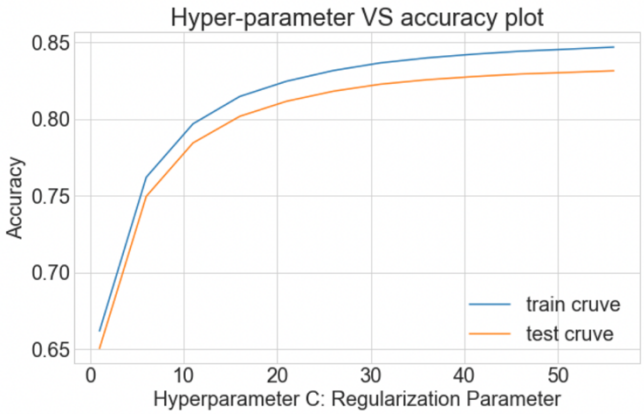
Appendix 1 – Cross-Validation on Hyperparameters

With 3 uncertainties and 3 features for regression.

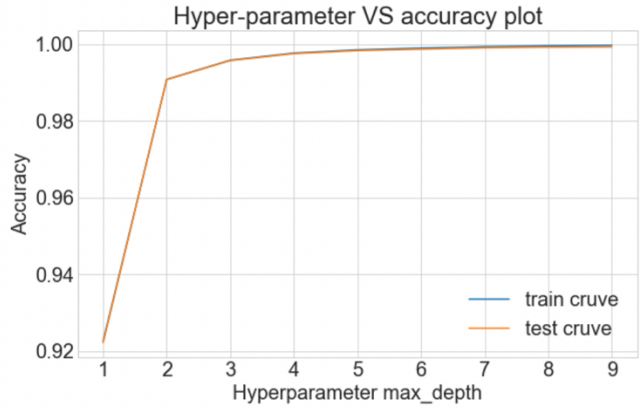
1- KNN cross-validation on hyperparameter k



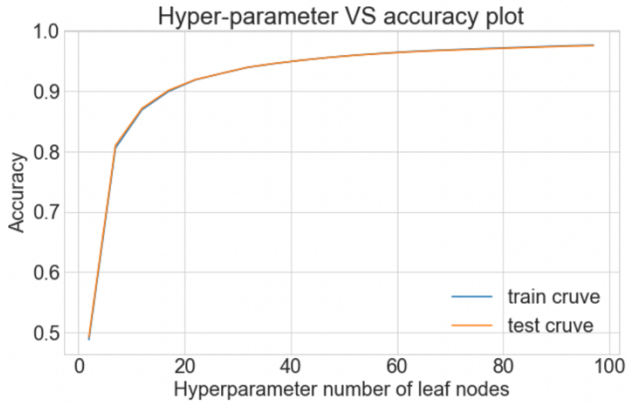
2 - SVR cross-validation on hyperparameter C



3 – XGB cross-validation on hyperparameter maximum depth

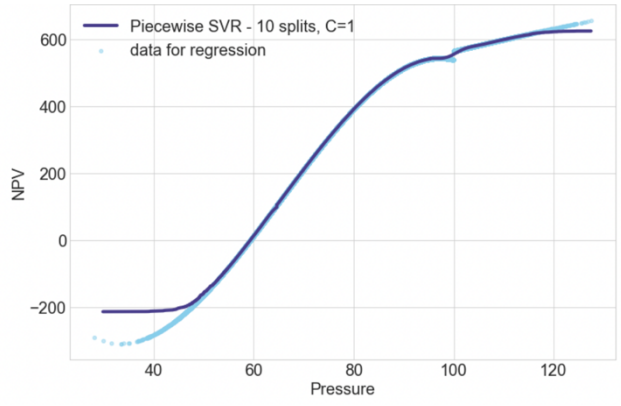


4 – RF cross-validation on hyperparameter maximum leaf nodes

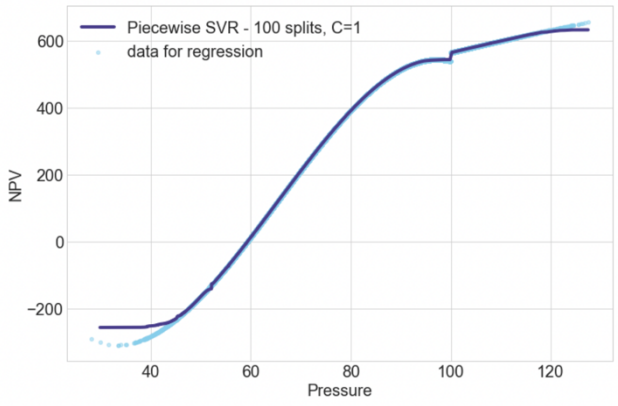


Appendix 2 – Sensitivity analysis on SVR and OLS performance using Piecewise Regression

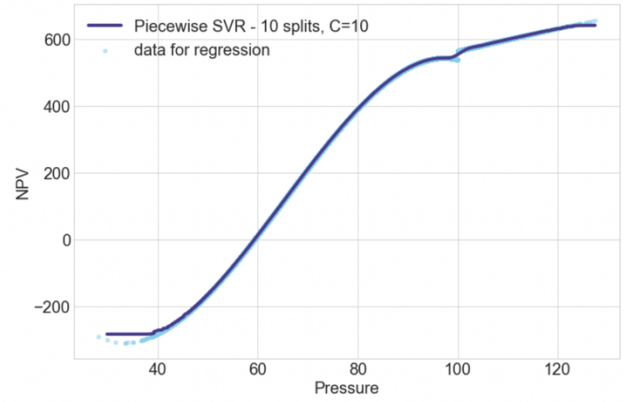
1 - 10 splits, C=1



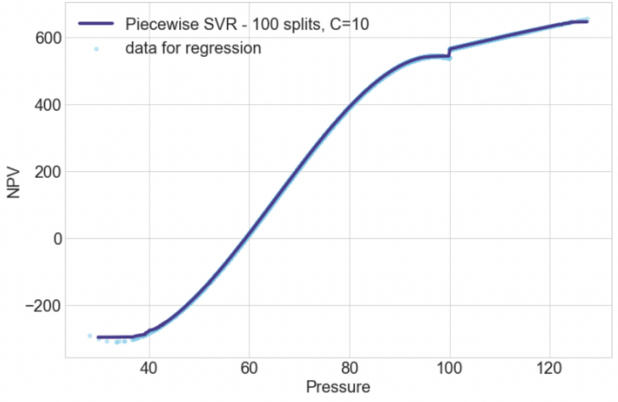
2 – 100 splits, C=1



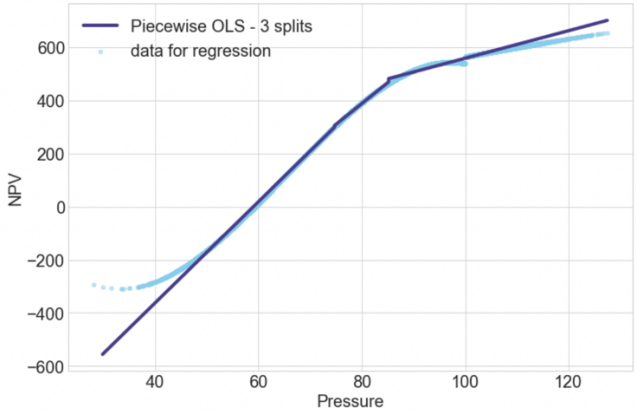
3 - 10 splits, C=10



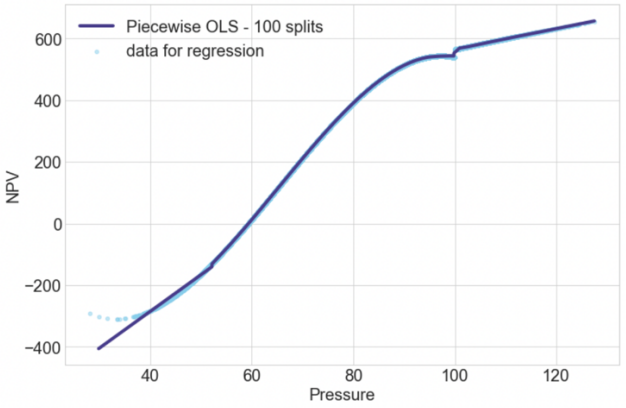
4 - 100 splits, C=10



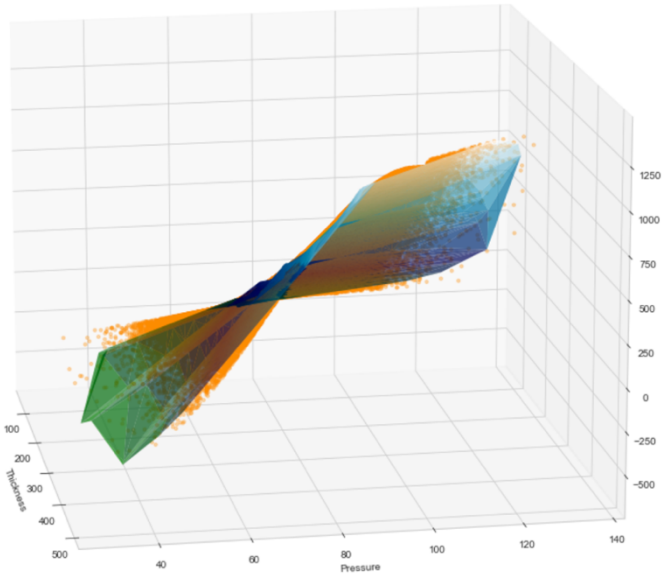
5 – 3 splits, OLS



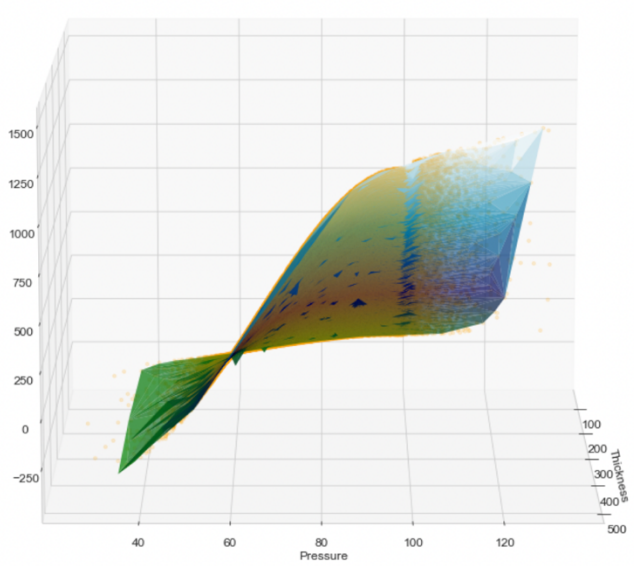
6 – 100 splits, OLS



7 – Multivariate 3 splits, OLS



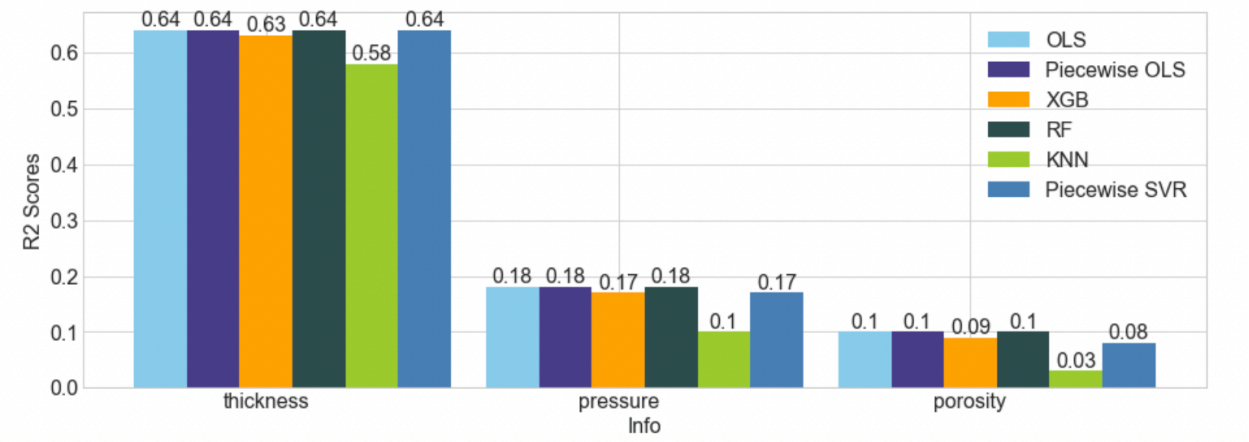
8 – Multivariate 100 splits, OLS



Appendix 3 – R^2 scores of Utsira case study

3 uncertainties – thickness, pressure, porosity

Info about one variable at a time– OLS is chosen



3 uncertainties – thickness, pressure, porosity

Info about two variables at a time – Piecewise OLS is chosen