PVE and SHAP

Application of SHAP and Machine Learning to PVE Quantitative Data

by

M.G.M. van de Ven

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Thursday August 24, 2023 at 16:00.

Student number:	4456505	
Project duration:	November, 2022 – August, 2023	
Thesis committee:	Dr. mr. N. Mouter,	TU Delft, first supervisor
	Dr. G. de Vries,	TU Delft, chair and second supervisor
	M. de Vries,	Populytics, external supervisor
	J. I. Hernández,	advisor

Preface

“I am a slow walker, but I never walk back”

- Abraham Lincoln

This thesis has one name under it, but the list of contributors is nearly endless. There are certain people who make the world a better place just by being in it. If you are reading this, you are probably one of those people to me. Thank you! I will never forget the help, patience and support you provided.

M.G.M. van de Ven
Delft, August 2023

Summary

Governments are faced with difficult policy dilemmas. Citizen trust in government decisions is crucial for a large support of government policies (Mangion & Frendo, 2022). To better involve citizen preferences in policies, governments use public participation practices (Arnstein, 2019). Examples of public participation practices are citizen forums or surveys among the population. These practices aim to provide insight in the preferences and opinions of citizens and involve citizens in the decision-making process.

A relatively novel approach of public participation is the Participatory Value Evaluation (PVE) in which a dilemma of a policymaker is provided to citizens (Mouter, Shortall, et al., 2021). In a PVE, citizens face a realistic choice task in which the policy dilemma is explained to the participant (Hernandez et al., 2023). Participants have to divide a set budget over several options. Citizens can for example choose several projects from a list of possible projects, but have to remain within budget. Or citizens are given a budget to divide over several spending options. The experiments result in datasets containing participants preferences and socio-economic characteristics of the participant.

A literature study of the state-of-the-art in PVE analysis revealed there are three main methods for analyzing PVE datasets: descriptive statistics, Latent Class Cluster Analysis (LCCA) and choice modelling. The descriptive statistics provide an aggregate overview of the dataset. The LCCA is often used to gain insight in different clusters of participants apparent in the dataset. This gives policymakers a feel for the subgroups in the society. Choice modelling is used to compute optimal portfolios. Choice modelling is time-consuming and prone to analyst assumptions, whereas LCCA insights are limited to identifying clusters in the data. A tool that can identify relations between participant features and their preferences can provide additional insights to policy makers and researchers.

An example of such a tool is machine learning. Machine learning is a technique that is able to find correlations between two sets of data. In analysing large datasets, machine learning can outperform human learning (Kühl et al., 2022).If machine learning is to be applied to PVEs, it could be used to predict choice task outcomes based on features of the participant. However, the opaque algorithms in machine learning can make it difficult for a human to understand how the results were produced, which can make it difficult to interpret. The field of Explainable AI has risen as a response to this issue: the aim of explainable AI methods is to give insight in the inner workings of the machine learning algorithm. An example we further research is SHAP (SHapley Additive exPlanations), a method to gain insight in machine learning models by explaining how each individual prediction is caused by (demographic) features (Lundberg & Lee, 2017).

SHAP might be able to reveal insights in the PVE datasets that are currently not found. However, this has not been academically tested. Therefore this thesis focuses on the implementation of SHAP in PVE analysis, under the following research question:

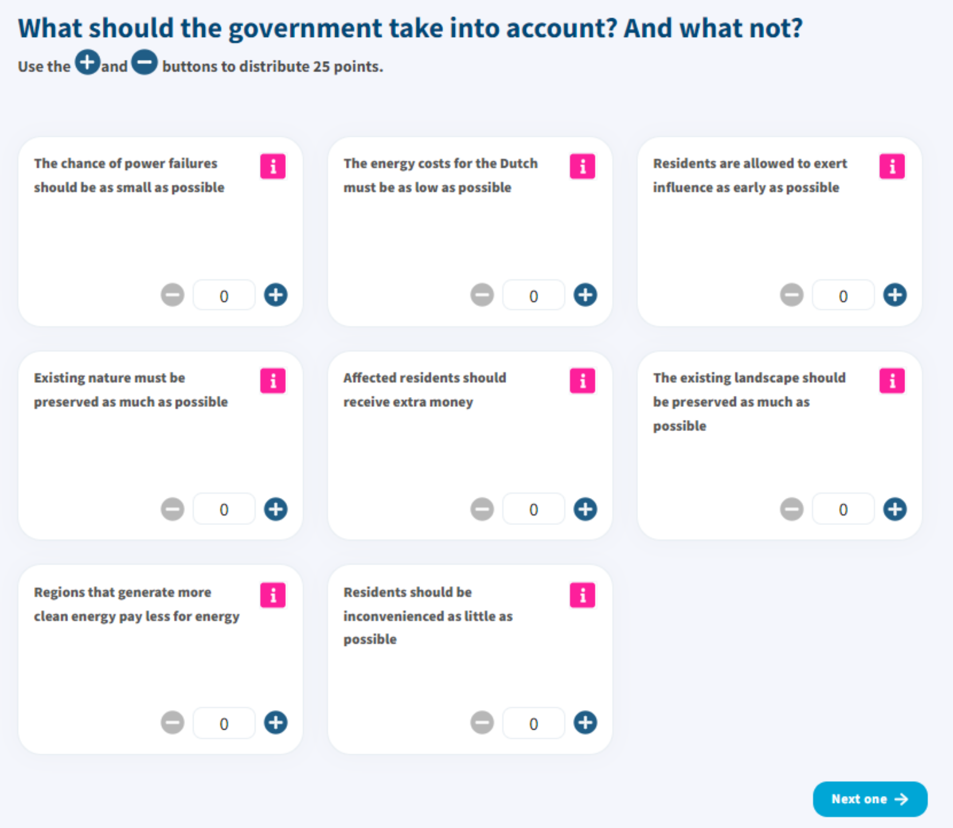
”What additional insights does machine learning with SHAP provide for PVE quantitative analysis compared to conventional methods?”

To answer the research question, a case study is performed by applying the SHAP method to the PVE of the National Programme Regional Energy Strategy (NP RES) [1]. The programme is aimed at investigating where and how the Dutch targets for renewable energy can be met. The PVE experiment of NP RES asked citizens to indicate which value-based statements they found most important. They could divide a budget of 25 points over the statements, as shown in Figure 1. The project was chosen for a case study since choice modelling was not suitable to analyze this type of PVE. The LCCA did not

[1] https://regionale-energiestrategie.nl/default.aspx

generate many suitable insights according to the analysts. Therefore, the NP RES dataset provides a compelling case to test the potential benefits of applying SHAP. SHAP is focussed at relating demographic factors with choice outputs, and is thus interesting to compare to LCCA mainly.

Figure 1: Wevaluate interface for NP RES PVE



What should the government take into account? And what not?

Use the ⊕ and ⊖ buttons to distribute 25 points.

The chance of power failures should be as small as possible — 0

The energy costs for the Dutch must be as low as possible — 0

Residents are allowed to exert influence as early as possible — 0

Existing nature must be preserved as much as possible — 0

Affected residents should receive extra money — 0

The existing landscape should be preserved as much as possible — 0

Regions that generate more clean energy pay less for energy — 0

Residents should be inconvenienced as little as possible — 0

Three machine learning models (XGBoost, Random Forest and an Artificial Neural Network) were applied to the PVE dataset, with the Random Forest model providing the best fit to the dataset. The Random Forest model was therefore chosen for analysis using SHAP.

When these results of SHAP analysis are compared to the results of the LCCA, it is apparent that SHAP provides more insights (26 versus 11). More is not necessarily better, but the insights from SHAP were more detailed than the LCCA insights. An example of an insights gathered using SHAP is that female participants award more points to choice options that favor the state of nature. An example of an insight from the LCCA is that gender is not significant in determining clusters of participants. SHAP is able to reveal patterns on a smaller scale than LCCA. The resulting insights are different to the results of the LCCA analysis. Many of the insights from SHAP analysis are not seen in the LCCA.

Overall, it can be concluded that applying SHAP results in new insights that were not found with other methods used on the NP RES PVE case. This study has shown that SHAP can be a relevant tool to gather insights about the PVE data and the differences among participants. It can gather individual effects of demographic variables on the choices participants make. Therefore it can lead to more and refined policy advice to governments.

The results of the SHAP analysis pose new questions. How is it possible that the LCCA and SHAP sometimes provide contrasting insights? And how reliable is the method of SHAP on PVE experiments? These questions can be addressed in further research. An important next step is to research how large a PVE dataset should be before SHAP provides stable results, as well as a sensitivity analysis to gain insight in the stability of the SHAP method under changes in the dataset, model or variables. Future research may improve the understanding of the reliability of the SHAP method.

PVE experiments are still in their infancy. The ability of SHAP to provide additional insights into PVE experiments within this thesis provides an incentive to further use SHAP in PVE experiments, including academic PVEs. SHAP predominantly provides insights into the relation between participant characteristics and their valuation of options in the choice task, allowing for the diversity of groups within the participant population to be directly addressed. This direct address broadens the range of results and does justice to the diversity of our society.

Contents

1 Introduction

The introduction of this thesis report will serve as a guideline for the rest of the report. In the introduction the scope for the thesis is set. Starting with an exploration of Participatory Value Evaluation (PVE), a research gap is delineated. To contribute to filling this knowledge gap, the central research question of the thesis is set, along with the research approach. Governments face dilemmas daily. When citizens do not trust governments, extra difficulties for governments arise (Mangion & Frendo, 2022). Especially during the COVID-19 pandemic, the topic of government trust became more and more researched (Devine et al., 2021). Governments wanted citizens to take COVID-19 vaccines, and stick to behavioral measures. Although it is difficult to measure trust, and the scientific community does not yet have a holistic understanding of government trust, it can be said that there is a relation between citizen adherence to measures and government trust (Devine et al., 2021). Less adherence to COVID measures, according to the same study, in turn led to higher mortality rates, revealing the importance of government trust.

1.1. Preference Elicitation

Due to the size of the population, it is often costly in terms of time to hear everybody's opinion and take it into account. Taking the opinion of citizens or stakeholders into account when making policies is called participation, there are many different forms to do this (Arnstein, 2019). It is important for policy support, quality enrichment and strengthens democracy (Edelenbos, 1999). A downside of participation is that it leads to an information explosion, that can be though to manage (Edelenbos, 1999). Most public deliberation methods like citizen forums take up a lot of time, which is a large disadvantage (Irvin & Stansbury, 2004).

A method that is often used to asses the desirability of publicly financed projects is the Cost-Benefit Analysis (Layard & Glaister, 1994). Methods like conjoint analysis, experimental auction and contingent valuation method are used to estimate the worth of non-monetary costs and benefits (Grunert et al., 2009). In these ”stated-preference” methods, citizens face a choice task. Boxall et al. describe them as ”methods that involve the elicitation of responses to predefined alternatives in the form of ratings, rankings or choice” (Boxall et al., 1996, p. 244). There are downsides to these methods which are not fully addressed. (Kahneman & Knetsch, 1992). An example of these downsides is that when the order of the experiment is changed, the results change as well. When aforementioned stated-preference methods are compared to each other, they often result in different results (Stevens et al., 2000). In the study by Stevens et al. differences up to 20 times in numbers were found. Standard stated choice experiments are thus associated with high biasses (Hanley et al., 2001).

1.2. PVE

A new method for preference elicitation of citizens was developed at TU Delft. It is called the Participatory Value Evaluation (PVE). In that method, participants are faced with the dilemma of the decision-maker (Mouter, Shortall, et al., 2021). They have to make choices, while being constrained in for

instance a budget.
The PVE method is first posed in a discussion paper in 2019 (Mouter et al., 2019). In the paper the flaws of the Cost-Benefit Analysis (CBA) method are discussed. Instead of a Willingness-to-Pay being evaluated, the authors build further on the Willingness-to-Allocate-Public-Budget. Instead of asking citizens how much they would pay privately for public projects, the citizens are asked how much they think the government should pay for public projects. The paper was released together with a further discussion paper about the economics of the PVE method (Dekker et al., 2019).

The essence of a PVE is that the citizen face a realistic experimental setting in which the policy dilemma is recreated (Hernandez et al., 2023). This dilemma is first brought to the essence, to ensure a citizen is able to perform the PVE in about 20 minutes (de Vries, Spruit, et al., 2022). For instance, a citizen can choose certain projects, but there is a budget limitation (Mouter et al., 2019). This forces citizens to explain what choices they would make, instead of being radically against an idea (de Vries, Spruit, et al., 2022). It is a way to give the silent middle a voice. Citizens give explanations to support their answers to the choice task. In the study by Mouter et al. 617 participants could divide 100 points over four strategies for a heat transition plan in which neighbourhoods are converted from using gas as heating source, to electricity (Mouter, Shortall, et al., 2021).

The method has since been further developed and used at both TU Delft and the company Populytics. It has been used to research which coronavirus measures Dutch citizens would prefer (Mouter, Jara, et al., 2022). Furthermore it has been used to investigate which measures to take to promote healthy body weight, energy transition and flood protection (Mouter, Koster, et al., 2021b; Mouter, Shortall, et al., 2021; Mulderij et al., 2021).

The method of PVE has undergone partial scrutiny. One of the things that has been tested is what the effects are of goal-dependent design of PVE, in which a differentiation of PVE designs is posed depending on the participation goal one wishes to achieve (Bouwmeester, 2021). More recently, the face validity of the method has been put to the test resulting in a framework to test face validity of specific PVE projects (Tuit, 2022). The testing of face validity involved investigating whether participants found the choice options to be genuine. While the study could not yet conclude definitively on the face validity of PVE, several recommendations for enhancing face validity were drawn for future PVE experiments.

The field of PVE is relatively young. Since it is so young, there are many open ends to the research. All research performed so far has a direct link with either TU Delft or the company Populytics.[1] One of the issues faced, is that there is hardly any research into analysis methods for PVE. It is relatively unknown which data analysis methods fit the PVE best. Current practises emerged rather spontaneously, according to the practitioners at TU Delft and Populytics. Since PVE can elicit preferences of citizens for academic research and government policy makers, insights from the PVE experiments can have important implications for government policies and academia. It is in the publics interest to draw correct conclusions from the PVE experiments. With not enough information about analysis methods, there is the risk that not all insights in the data are correctly extracted.

1.3. Analysis Methods

Examples of PVE analysis methods are descriptive statistics and LCCA (Latent Class Cluster Analysis) (de Vries, Spruit, et al., 2022), along with Choice Modelling (Bahamonde-Birke & Mouter, 2019). The descriptive statistics involve a general statistical overview of the results of the PVE experiment, such as average choices and standard deviations. The LCCA is used to find clusters of participants in the dataset and describe those clusters in terms of demographic characteristics to decision-makers (Geijsen et al., 2023). In choice modelling, relative utilities are computed to compare different portfolios of choice options on their utility. Hence, optimal portfolios of choice options can be deduced. An upcoming technique for data analysis is Machine Learning (ML). By using ML, an analyst does not need

[1] I am aware that in July and August 2023, during the closing stages of this thesis, two external academic publications were published using PVE as a method. Due to the timing, these two studies have not been considered in this thesis. The publications are about a PVE for climate goals, and its relevance and credibility (Hössinger et al., 2023; Juschten & Omann, 2023)

to make any assumptions on the data, as with modelling techniques (Buskirk et al., 2018). Machine Learning outperforms human learning in cases where lots of data is available (Kühl et al., 2022). PVE experiments result in large datasets, which makes machine learning a suitable method to use to reveal insights from the dataset that would not be revealed with conventional methods.

In a basket-based choice experiment (BBCE) researchers concluded that a barrier to the analysis was the enormous amount of combinations a participant could choose (Caputo & Lusk, 2022). In this BBCE participants could choose from 21 food options to construct a basket of one or more food options that they wished to purchase. Participants had to stay within a fixed budget. This experiment is very close to a PVE. As the amount of combinations of choice options in PVE and BBCE are so large, it becomes increasingly difficult for choice models to calculate the willingness to pay. Machine Learning can provide results many times faster than a choice model can. That makes machine learning suitable for these types of applications.

1.4. Machine Learning

The roles of machine learning and artificial intelligence are growing rapidly (Aggarwal et al., 2022). Artificial Intelligence is making a computer imitate human brains. It is aimed at performing complex tasks by computers, that traditionally would be done by humans (McCarthy et al., 2006). Machine Learning is a subset of Artificial Intelligence (MIT Sloan School of Management, 2019). Machine Learning aims to find patterns in data, and exploit the patterns to make predictions outside of the data. The difference with modelling is that in a model a human has to give directions about the expected shape and form of the real world structure. In a Machine Learning model, the computer finds patterns on its own, without being told what type of structures it should find. It is rising as a method in survey research, to analyze the data produced (Buskirk et al., 2018).

1.5. Explainable AI - SHAP

When a machine learning model is trained on a dataset, it is capable of making predictions based on inputs given to it. However, it gives no information on how it makes the predictions. Usually, the inner workings of a machine learning model remain opaque, making it intransparent and difficult to work with for policy makers. If a policy maker does not know what the algorithm bases its predictions on, it might be misinterpreted. An example in the past was that to determine the difference between a husky and a wolf, the machine learning model mainly used the background (Ribeiro et al., 2016). When there was snow, the image was classified as a husky, and vice versa. The model was not able to distinguish wolves and huskies accurately. This is relevant since at first the model seemed to be rather accurate at distinguishing huskies and wolves. However, it was innately not able to meaningfully determine the difference. This example shows the risk of using black box machine learning models, and at the same time shows the potential of explainable AI methods that can explain how machine learning models make their predictions.

To explain the black boxes of machine learning, the field of Explainable AI (XAI) has emerged. It can be used to help gain understanding of what the model bases its predictions on. Furthermore it can be used to grant more confidence in a model, when it bases its predictions on logical inputs.

One of the XAI techniques available is SHAP (SHapley Additive exPlanations) (Lundberg & Lee, 2017). It can assist in explaining the outcome of a model. SHAP values indicate the importance of a feature to the prediction of the model. It explains the difference between the model prediction when the feature is unknown, and the model prediction when the feature is known. As interpretability is important in the useful usage of machine learning, SHAP can contribute to a more reliable implementation of machine learning (Khosravi et al., 2022). Instead of creating a machine that can make predictions, SHAP can also explain these predictions, giving us insight in the data on which the machine is trained.

1.6. Research Gap, Question and Approach

PVE was founded to help gather decision-making information about the preferences of citizens. It is built as a method that takes into account preferences towards public policy (Mouter et al., 2019). Since the PVE method is relatively young, there are many aspects that are not yet fully researched. Performing PVE-evaluations results in large datasets, which contain information about the preferences of citizens. They also contain information about the participants. Analyzing the data can reveal citizen preferences, and how they are related to demographical factors. New methods of analyzing this data can deliver more valuable information for decision-making, which can give decision-makers in public policy a more effective method of taking citizen preferences into account. Not only can effective participation methods help in increasing citizen trust, they can assist in making more legitimate decisions as well (Creighton, 2005; Kim & Lee, 2012).

Machine learning is an upcoming method in survey data analysis (Buskirk et al., 2018). One of the main advantages is that no assumptions have to be made about the structure of the data. Other methods used on PVE data, such as LCCA and Choice Modelling do involve assumptions on the dataset. SHAP can be used to give predictions on individual profiles, making the possible results more zoomed in than the aggregate outcomes of LCCA and Choice Modelling (Lundberg & Lee, 2017). Therefore machine learning combined with SHAP can result in more insights from PVE datasets. Machine learning approaches with SHAP are thus interesting to analyze PVE experiments. However, the use of machine learning and SHAP on PVE data is not broadly researched yet, and thus there is no information on whether SHAP would contribute to more insights in undertanding citizen preferences. With this thesis, I intend to contribute to discovering whether SHAP is a method that can result in better analysis of PVE data. Therefore in my thesis I will use the following central research question:

“What additional insights does machine learning with SHAP provide for PVE quantitative analysis compared to conventional methods?”

1.7. Research Subquestions

To answer the main research question, multiple steps need to be taken. The first step is to find out how PVE-evaluations are currently analyzed. Which methods are already in use? What data is gathered? Which methods have been considered? Which methods are chosen? To find the answer to those questions a literature study will be performed on current PVE practices. More information about the choice for literature study as method is provided in chapter 2. This is summarized in the first subquestion:

1. How is quantitative PVE data currently analyzed?

To answer whether machine learning brings additional insights, a case study is set up. To test the practical appliance of machine learning, a case study is performed to (partially) answer the research questions. This case study is used as basis for the comparison of analysis methods throughout. In providing a practical case study for the application of machine learning to PVE data, we hope to find out about the potential extra insights the method can give. Using a case study might reveal problems, discussion points and practical connotations to the use of machine learning. More background on the rationale of using a case study as method is provided in chapter 2. The following subquestion is used for the case study:

2. Which results does Machine Learning provide on a PVE case study?

The aim of the case study is to reveal insights about the NP RES dataset. To know whether these insights are additional and useful, the results are compared to conventional methods under the following subquestion:

3. ”How do the results of machine learning compare to conventional PVE analysis methods?”

1.8. Choice for Case Study

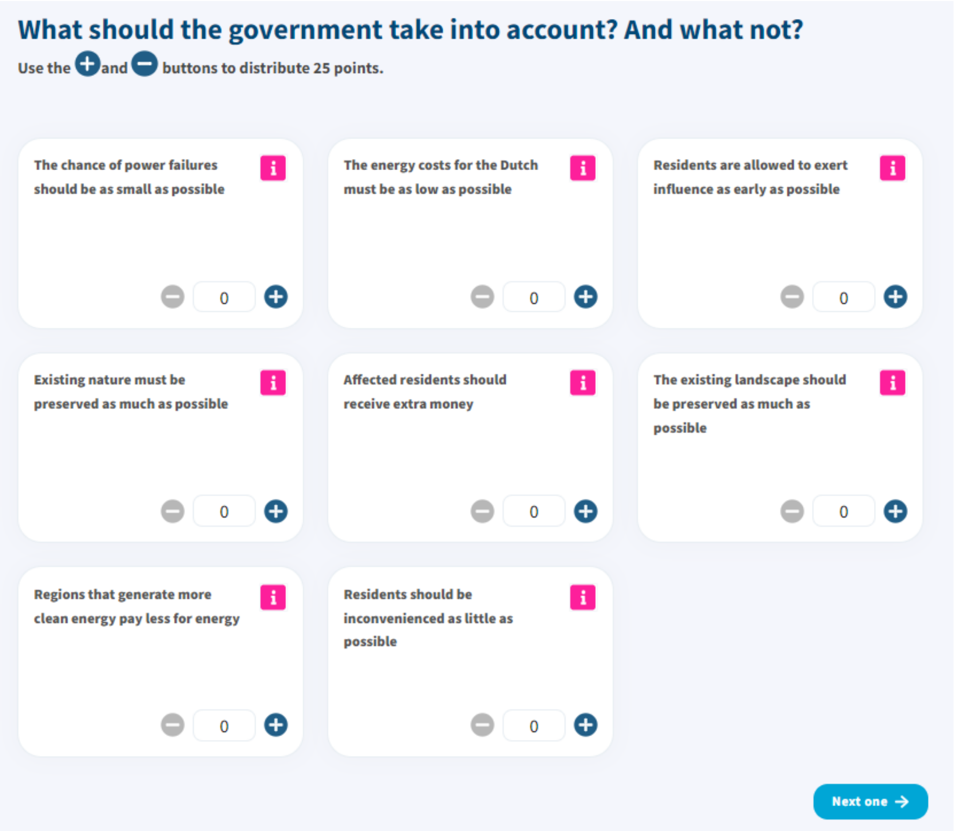
To choose a case study, staff of Populytics was asked which one of their datasets suits the research best. They reported that the case for the National Programme Regional Energystrategy (NP RES) was a viable option. It is a points-type PVE. In such a PVE, a participant can divide a set amount of points over different value statements. There is no choice modelling approach that suits this type of PVE, according to Populytics staff. The LCCA did, to their opinion, not reveal many insights. There were three clusters identified in the data, but the (demographic) characteristics of the clusters were very similar. The amount of insights about how different groups in society responded to the PVE was therefore, according to the analysts, limited. It is thus an interesting case to test the added value of using machine learning, and see if machine learning combined with SHAP is able to generate many interesting results.

1.9. NP RES Case Description

NP RES is a programme by the Dutch national government that aims to increase sustainable energy production in the Netherlands. It focuses on finding out how much renewable energy can be produced where in the Netherlands, and determining regional goals. These regional goals will then have to be met by the regions. To support the process, Populytics held a PVE experiment among Dutch residents, to provide information about which values were important to consider to Dutch citizens.

On the 6th of December, 2022, the PVE experiment opened. All 1534 valid respondents originate from a panel by Dynata, a panel company. In total there were 2843 participants in the dataset, of which many were invalid or could not be used. The cleaning process is described in Appendix C. Using a panel can assist in having a representative group of citizens as respondents. In an open experiment, this can not be controlled for. Respondents faced the task of awarding 25 points to value statements. The interface can be seen in Figure 1.1.

Figure 1.1: Wevaluate interface for NP RES PVE



What should the government take into account? And what not?

Use the ⊕ and ⊖ buttons to distribute 25 points.

The chance of power failures should be as small as possible

The energy costs for the Dutch must be as low as possible

Residents are allowed to exert influence as early as possible

Existing nature must be preserved as much as possible

Affected residents should receive extra money

The existing landscape should be preserved as much as possible

Regions that generate more clean energy pay less for energy

Residents should be inconvenienced as little as possible

The interface is originally in Dutch, since the entire experiment was held in Dutch. The webpage is automatically translated by Google Translate for the image. The Dutch/English translations used in the text of the report can be evaluated in Appendix B.

1.9.1. NP RES Dataset

The NP RES dataset contains many data columns, which we do not all need. What is needed for the machine learning analysis is the inputs and outputs that we wish to include in the analysis. Since the experiment was already performed, I had no influence on the questions asked in the PVE, so I am limited to the information in the dataset for the analysis. For the inputs, demographic variables about the respondents are chosen, as seen in Table 1.1. For the outputs, the choice task point allocations are chosen, as seen in Table 1.2. The goal of the machine learning process is to use observed variables about participants, such as demographic variables, and see if we can predict their choices. Later, the trained machine learning model can then be analysed with SHAP to reveal which demographic variables are important in the prediction of choice task. Therefrom we can deduce which variables the machine learning uses and what effect they have. The outputs are all choice task options, since those are the choices the machine learning model will try to predict. The inputs are all variables that are in the dataset that tell us something about the participant. We thus use as many inputs as reasonably possible, to give the machine learning full freedom in picking inputs to predict from. Feature selection (removing inputs) is mainly necessary if the runtime of a machine learning model is too large (Cai et al., 2018). Since the amount of inputs here is low enough for the model to run within a minute, there is no need to do a further selection. The machine learning process wil estimate influence of irrelevant inputs to approximately zero (Cai et al., 2018).

Question	Levels	Level Explanation
In this PVE we asked advise from Dutch citizens. We can also ask advise from experts. Which advise is most important to you?	0	The government should fully follow citizen advice
	1	The advice of citizens is more important than expert advise
	2	Both advices are equally important
	3	The advice of experts is more important than citizen advice
	4	The government should fully follow expert advice
What is your age?	0	Below 35 years old
	1	Between 35 and 65 years old
	2	Over 65 years old
What is the highest level of education you completed?	0	Primary school, VMBO, MBO-1, Year 1/2/3 of HAVO/VWO
	1	Year 4/5/6 of HAVO/VWO, MBO-2, MBO-3 or MBO-4
	2	HBO or University Bachelor, HBO or University Master
NaN, it was indirectly derived based on panel data	0	Less than 50,000 inhabitants
	1	Between 50,000 and 150,000 inhabitants
	2	More than 150,000 inhabitants
Which sentence describes your situation best?	0	Every month, I have a shortage of money
	1	Every month, I have enough money
	2	Every month, I have more than enough money
Which statement fits you best?	0	I am a man
	1	I am a woman

Table 1.1: Table of all inputs used for machine learning analysis

Codename	Variable	Full Statement
landscape	kt1_landscape	The existing landscape should be conserved as much as possible
influence	kt2_influence	Citizens are allowed to exert influence as soon as possible
failure	kt3_failure	The chance of grid failures should be as low as possible
costs	kt4_costs	Energy prices for Dutch citizens should be as low as possible
region	kt5_region	Regions that produce more clean energy, pay less for energy
nature	kt6_nature	The existing nature should be conserved as much as possible
nuisance	kt7_nuisance	Citizens should have the least amount of nuisance possible
compensation	kt8_compensation	Citizens that experience nuisance should be compensated with extra money

Table 1.2: Outputs for Machine Learning Prediction

1.10. Populytics

In this thesis, I am supported by the company Populytics. Populytics performs PVE-evaluations and is a spin-off of TU Delft. I can use their resources and support. Populytics is currently the only commercial company performing PVE-analyses. It emerged from the same TU Delft research group that performs the research on PVE's. It helps the government, provinces, municipalities and companies to elicit preferences [2].

The added value of performing my thesis both at TU Delft and Populytics is that I can use the real-world datasets of Populytics. Furthermore their employees have experience in using PVE as a practical tool in real life. Access to their expertise and support helps me better understand PVE and its applicance.

For this thesis, a case study is performed. The dataset used was generated by a Populytics project for the Dutch National Programme on Regional Energystrategy (NP RES). The NP RES aims to manage and divide the goals of renewable energy targets in the Netherlands.

1.11. Link with EPA Programme

The MSc Engineering and Policy Analysis is focussed on contributing to policies to address “Grand Challenges” [3]. According to their website those are “complex problems that involve many parties with conflicting interests” [4]. Low government trust is related to many grand challenges. If governments don't function well, their policies suffer. Dealing with challenges such as climate change, energy prices or the war in Ukraine can become problematic if the policies are not supported by Dutch citizens. In that sense, having a trustworthy government can be seen as a prerequisite to solving grand challenges.

The research question is focused on analysis methods. It is therefore both related to public policy and modelling/data analysis. The PVE experiments are in itself models of the dilemma the policy maker faces in reality.

1.12. Use of Software

In this section an overview of used software will be given. The software used is important information, since it gives a quick overview of reproducability. Furthermore it provides information on some of the dependencies of the thesis report.

The major software tool used in the project was Python. The version used throughout was Python 3.9.13. Python was chosen due to previous projects on the same subject being programmed in Python as well, therefore comparison can be done and lessons learned taken into account. Furthermore Python

[2] https://populytics.nl/wie-zijn-wij/

[3] https://www.tudelft.nl/onderwijs/opleidingen/masters/epa/msc-engineering-and-policy-analysis

[4] https://www.tudelft.nl/onderwijs/opleidingen/masters/epa/msc-engineering-and-policy-analysis

is a widespread and free programming language. It is therefore accessible for others to use. Several packages that are build for Python have been used in the project:

- Numpy (v1.23.5): basic module for handling data structures.
- Pandas (v1.5.2): module specifically for data analysis, with easy-to-use data editing and overviews.
- scikit-learn (v1.2.0): module for machine learning, used for both Random Forest and Neural Network predictions.
- xgboost (v1.7.3): module for XGBoost machine learning.
- shap (v0.41.0): module for Explainable AI Method SHAP.
- matplotlib (v3.7.0): module used for visualizations.
- seaborn (v0.12.2): module for advanced visualizations.

The platform used by Populytics to perform the PVE's is called Wevaluate. It is in-house software developed by Populytics. It is capable of performing PVE's. It provides a User Interface for citizens to fill in the PVE, and provides a backoffice for Populytics employees to create and customize the PVE. It thereafter exports the data. I have not used Wevaluate myself for this thesis, but without the software the case study that is performed could not be done.

1.13. Use of AI

In 2023, the use of Artificial Intelligence (AI) is starting to be more and more regular (Aggarwal et al., 2022). High school children already use programmes like ChatGPT to assist them in making homework assignments [5]. ChatGPT is a powerful chatbot trained using machine learning which is capable of answering questions in language indistinguishible from human language answers (Lund & Wang, 2023). It is powerful in performing natural language processing tasks, but has trouble performing logical reasoning tasks.

At TU Delft, the discussion about the use of ChatGPT and other AI software is still ongoing at the moment of writing this thesis [6]. TU Delft says ChatGPT can be used as a learning tool, but it has implications for teaching. For instance, practical assignments should be designed such that they are difficult to perform using just ChatGPT. Furthermore, students need to be educated on the risks and dangers of using such AI. In ethical guidelines surrounding AI, transparency is very frequently used (Jobin et al., 2019). To increase transparancy, in this section an overview of used AI is given. A rationale will be given for the use of AI, as well as a workflow description.

In this thesis ChatGPT has been used for feedback and reflection. A first use of ChatGPT was when I got stuck in coding with an error. Instead of using Google to find the answer to the problem, I have sometimes used ChatGPT to give feedback about the code and why it would not run. Often the answers given inspired the solution. Furthermore already written bits of text have been entered into ChatGPT to provide textual feedback, in a similar way as Grammarly would. This has so far been done for two paragraphs, in which the ChatGPT answers have inspired some textual changes. I have unfortunately not held an overview of which paragraphs these were. Given that these ways of using ChatGPT resemble asking others around for feedback on text or searching the internet for solutions to coding issues, these are not cited or referenced.

In this thesis ChatGPT has not been used to generate text or generate code for the thesis outside of the uses defined here:

- The LaTeX codes to insert the figures in Appendix A are written by ChatGPT based on the first one that I did manually. It saves time to let such a repetitive task be done by AI.

[5] https://nos.nl/artikel/2460020-chatgpt-glipt-langs-docenten-ik-gebruik-het-om-snel-huiswerk-te-maken

[6] https://www.tudelft.nl/2023/teaching-support/chatgpt-at-tu-delft

1.14. Definitions and Terms

Some definitions and terms will come back often throughout the report. If that is the case, an explanation is given in this section if necessary.

1.14.1. PVE Definitions

- Choice Task: a Choice Task is the quantitative choice a participant had to fill in during the PVE. It can be choosing projects within a budget or allocating a budget/points over statements.
- Qualitative PVE Data: the written answers provided by participants in the PVE. These are left out of the scope of this thesis.
- Quantitative PVE Data: the choice task answers of participants. These all results in numerical choices, suitable for analysis using quantitative methods.

1.14.2. Statistical Definitions

- Covariates: variables that are analyzed to see if they have any predictionary value for the outcome of the choice task, such as demographic variables.

1.14.3. Machine Learning Definitions

- Inputs: data fed to the algorithm which it will use to predict the outputs. In the context of this thesis, these are demographic variables.
- Features: same as inputs.
- Outputs: data fed to the algorith which it will try to predict using the inputs. In the context of this thesis, these are choice task outcomes.

1.15. Outline of Report

The thesis report continues with a description of the subresearch questions in support of the main research question, and an explanation of the methods used. This is given in chapter 2. Thereafter, the first research question investigating the conventional use of PVE analysis methods is answered in chapter 3. The results of the applied SHAP analysis to the case study are given in chapter 4. The comparison between conventional methods and SHAP is performed in chapter 5 and the results further discussed and concluded in chapter 6 and chapter 7.

2 Method

The goal of the thesis is to explore how machine learning with SHAP can be used to gain additional insights in PVE quantitative data analysis. The aim is to develop and apply machine learning models combined with SHAP to identify patterns and relationships within the data that may not be apparent using traditional statistical methods. Per subresearch question, an overview of the method used is given. The main research question is:

“What additional insights does machine learning with SHAP provide for PVE analysis compared to conventional methods?”

In chapter 1, three subquestions were defined. For each of these questions, the methods are discussed in this chapter. For each question, a rationale is given why this method is suitable and why I expect it to raise valuable insights.

2.1. Case Study

All three subquestions relate (partly) to a case study that is performed throughout. In a case study, a subset of the world is examined, encouraging researchers to gain a deep understanding of the case they study (Patricia Anne, 2008). A case study is different to an experimental design in that it studies a phenomenon in a natural environment, instead of a controlled environment (Crowe et al., 2011). The NP RES case is a project that was performed without my interference, so it can be seen as a natural experiment. Case studies are specifically good at researching implementation (Paparini et al., 2020). Since this thesis focusses on a novel analysis method application on PVE, implementation is of specific interest. The choice for the case study of NP RES is explained in section 1.8.

2.2. RQ1: How is quantitative PVE data currently analyzed?

The aim of the first research question is to give an overview of how PVE data is currently analyzed. By zooming in the current case study at hand, an image is created of what current insights look like.

2.2.1. Method Choice

The main method for answering this research question is a literature study, with a smaller role for personal communication with PVE analysts. A literature review can give a clear view of where a body of research is at the moment, and where it originated (Rozas & Klein, 2010). Most of the Populytics and TU Delft PVE research ends up either in papers or in reports. They are thus described in literature. The field is small enough to be able to take into account all publications. All publicized PVE reports are scraped to find which methods are already in use. These methods will be described based on literature, and benefits and downsides addressed. The process of determining an analysis method is evaluated, by asking practitioners in personal communication how they choose their methods. The results from the personal communication are treated as background information to provide insight in how these choices are currently made.

2.2.2. Method Elaboration

Firstly, all published papers and reports in which PVE was used as a method are searched for. For Populytics reports, these are found via their website[1]. For academic papers the search is performed by searching for the PVE-related papers with Niek Mouter as (co-)author. Another search has been performed using ”PVE” and ”Participatory Value Evaluation” as search terms, but did not result in more literature.[2] Their citation lists are scanned to evaluate whether any sources were missed.

In every paper or publication, the method section is evaluated to determine which method is used. An overview of all used methods will be given as result. The found methods are shortly described to give the reader an indication of how these methods work and what results they produce. From personal communication with practicioners, a short indication is given how the choices for analysis methods are made. For the case study, the used methods will be evaluated in more detail.

Insights can be defined as information that helps to understand what is currently happening (Awan et al., 2021). Learning new insights is about disovering relations in a problem that we did not see before ((Köhler, 1967)). It is partly a creative and exploratory process (Chowdhury, 2006). Therefore gathering insights is a process that can give different results when performed by different practitioners. Insights from the LCCA will be taken from the written text in the NP RES report.

The processing of literature, and the answer to the first research question are all described in chapter 3.

2.3. RQ2: Which results does Machine Learning provide on a PVE case study?

The aim of the second research question is to learn more about the practical application of machine learning on a case study, and gather results. By applying theoretical knowledge, practical barriers and restrictions arise. The theoretical use of machine learning is tested, to see whether it meets expectations. Performing the analysis gives insight in the time and effort needed to process data by machine learning.

2.4. Method Choice

The method for answering this research question a machine learning approach with SHAP to explain the results. The method of machine learning and SHAP was fixated by the scope and research gap of the thesis in chapter 1. However, there are several machine learning methods, and the application of SHAP requires explanation.

2.4.1. Types of Machine Learning

There exist a multitude of Machine Learning model types. These can be classified on many different aspects (Jordan & Mitchell, 2015).

An important difference between Machine Learning models is whether they are supervised or unsupervised (Ayodele, 2010). In supervised Machine Learning, the data is labeled. The algorithm trains itself on a dataset in which it knows what for each data point the correct answer should be. In unsupervised learning, the algorithm will try to find patterns without knowing how these should look. An example of

[1] https://populytics.nl/rapporten/

[2] I am aware that in July and August 2023, during the closing stages of this thesis, two external academic publications were published using PVE as a method. Due to the timing, these two studies have not been considered in this thesis. The publications are about a PVE for climate goals, and its relevance and credibility (Hössinger et al., 2023; Juschten & Omann, 2023)

unsupervised learning is the LCCA that is already regularly applied to PVE. Examples of supervised Machine Learning algorithms are Logical Regression, Support Vector Machines, K-means clustering, boosting, Decision Trees and Neural Networks (Ayodele, 2010).

2.4.2. Choice of Machine Learning algorithm

There is an abundance of types of Machine Learning methods that can be used for this thesis (Jordan & Mitchell, 2015). It is infeasible and impractical to use a wide variety of models on the data to see which is of best use. That is outside the scope of this thesis. The aim of the thesis is to examine whether machine learning with SHAP can be used to gain additional insights in PVE experiment data. Less focus is placed on optimizing these results fully.

The choice for Machine Learning algorithm is therefore based on practical reasons. Within TU Delft, there is experience in applying Neural Networks, Random Forest and Boosted learning to PVE data. A PhD candidate is working with those methods. So far, the results have not yet been publicized. The choice for these three algorithms ensured that I could use the PhD candidate to support me with the modelling, and provide feedback on the Python codes. These three algorithms will therefore be used as options on the NP RES data. All of these three methods are applied, after which their model fit is evaluated. The best-fitting model is used for further analysis.

2.4.3. Model Comparison Criteria

The three different models will all be ran, and the best one is selected for further analysis. The reasoning is that based on literature it is not always possible to determine which algorithm suits the data best (Jordan & Mitchell, 2015). A trial-and-error approach can reveal the best fit.

All three machine learning models are applied to the data. They are trained to predict the choice task outcome of a participant as accurate as possible, based on their demographic variables. We can compute how accurate the predictions are by comparing the predicted values by the models with the actual choice task results of participants. To compare the outcomes, the Negative Mean Squared Error (NMSE) is computed for all prediction models to judge which model is most suitable to the data. It is the negative of the Mean Squared Error (MSE). The MSE is the average of all squared differences between model prediction and actual points allocated by the participant. It is a metric of the fit between the model prediction and the dataset. A better fit, means that the model is better compared to other models. It is capable of providing more accurate predictions. The negative of the MSE is taken to make the number align with computational standards, in which a higher number indicates a better fit. Using Mean Squared Error as evaluation is reasonable for prediction models (Wallach & Goffinet, 1989).

2.4.4. Model descriptions

In this section the Machine Learning models are described. After a short description of its workings, the parameters used are reasoned. Advantages and disadvantages of every model are given. An explanation of the Python scripts to perform all methods is provided.

The first machine learning algorithm used on the NP RES dataset will be the XGBoost algorithm (Chen & Guestrin, 2016). It is based on tree boosting, which is often used. It is relatively good on sparse datasets, making it suitable to use on PVE data.

The XGBRegressor function from the xgboost Python package is used to run the XGBoost analysis. As an objective, it tries to minimize the squared error between the prediction and the training data. As evaluation metric it uses the RMSE (Root Mean Squared Error), since this is most commonly used for analyzing regression functions (Buskirk et al., 2018). To tune for the best parameters, the gamma, maximum depth of tree and minimum child weight are varied. The gamma is varied between 0, 1 and 2. The maximum depth is varied between 3, 5 and 7. The minimum child weight is varied between 1, 2, 3 and 5.

The second machine learning algorithm used is an artificial neural network (ANN). Instead of using decision trees, ANN's mimic the workings of the human brain. They are formed by assembling neurons into a network.

The ANN used in this thesis is the MLPRegressor from the scikit-learn package. It uses the Adam Optimization Algorithm as a solver. The validation fraction is 0.2, and the maximum amount of iterations set is 10000. After 6 iterations without a change, of reaching the maximum amount of iterations, the training is stopped. The hidden layer sizes are varied between 5, 10, 20, 30 neurons and two levels of 5, 10 and 20 neurons. The initial learning rate is varied between 0.1, 0.3 and 1. The activation is switched between a Rectified Linear Unit (ReLU) activation and a tanh activation. The batch size after which the network is updated is varied between 50, 100, 300 and ”auto”, in which the algorithm determines a suitable batch size itself.

Random Forest algorithms combine many decision trees to make predictions.

In the experiment, the RandomForestRegressor from scikit-learn is used to make the predictions. Several hyperparameters are varied. The first one is the number of estimators, which is varied between 100, 200 and 500. The maximum depth of decision trees is varied between 3, 5 and 7. The maximum amount of features the algorithm can use in every tree is set by the max_features parameter. It is alternating between a sqrt (square root of total variables) and log2 type.

2.4.5. SHAP

To explain how the best selected machine learning model makes its predictions, an Explainable AI (XAI) method is used. One of the emerging XAI methods is SHAP (SHapley Additive exPlanations) (Lundberg & Lee, 2017). SHAP can be used to explain how a model makes predictions. SHAP values give an explanation for the difference between the prediction for a respondent and the prediction if we would not know any of the demographic features for the respondent. Per input/feature, the difference is calculated between the output number with the respondents input value, and the average respondents input value. The difference in prediction is a Shapley value. A larger number indicates that an input is more important than other inputs. Whether a value is positive or negative determines whether an output is positively or negatively influenced by the input.

By plotting the SHAP values of all respondents, an analyst can spot the effects an input has. Nonlinear effects can be seen, which is a large benefit compared to other methods. Once all the values are known, certain specific respondents can be found and the SHAP values can tell which characteristics of the respondent influence its choice.

2.4.6. SHAP Setup

The SHAP setup in the scripts for this thesis is the SHAP Explainer feature from the shap Python library. The explainer computes all shap values, which can then be plotted in several different plots. For this thesis, mainly the Beeswarm Input Plots, Beeswarm Output Plots and SHAP Scatterplots are used.

The process of performing the case study and producing and analyzing the results gathered is described in chapter 4.

2.5. RQ3: How do the results of machine learning compare to conventional PVE analysis methods?

The fifth and final research question aims to provide boundaries for the use of machine learning in PVE. Furthermore the results of machine learning can be compared to the results gathered in other methods to provide information on the usefulness of machine learning in PVE data analysis.

2.6. Method Choice

The research question will be answered by gathering the insights of other methods on the NP RES case study. These can be compared to the results of the machine learning approach. For this, a comparative analysis is needed. The insights of both analysis methods will be compared on amount of insights, type of insights and inclusion of insights in the other method.

2.7. Method Elaboration

To compare the amount of insights, an overview table of insights per analysis method is created. The types of insights are compared by categorizing all individual insights in categories. The insights, once gathered, will be analyzed to see what types of insights they gather. The insights gathered from the analysis methods are strings of text, and thus qualitative data. These insights will be grouped to better compare the type of insights both methods provide. For this grouping process, the general framework by Pope et al. will be followed (Pope et al., 2000). The first step is to familiarize with the data by reading into it. Then the insights will be labeled with types. These are defined inductively, which means that the categories are made when an insight does not fit in a category that we already seen before. When the process is performed, the number of categories can be reduced by grouping similar categories into overarching ones. The categories are dependent on the goal of the researcher, and should be made such that they support that goal (Pope et al., 2000). The goal in this thesis for this analysis is to see whether the type of insights that analysis methods produce are different in nature to insights that SHAP produces. Are there types of insights that SHAP can find, that other methods can not? Or are there types of insights that can not be found using SHAP, but can be found with other methods?

It is interesting to see if the insights gathered from SHAP and conventional methods largely overlap or differ. Therefore systematically, for every SHAP insight it is determined whether it is in accordance, disagreement or anything inbetween with results from the conventional methods.

The comparison of machine learning results to other methods is described in chapter 5.

3

Conventional Methods for PVE Quantitative Data Analysis

After defining the method, the analysis for the first subquestion can be performed. The chapter aims to answer the following research question:

> How is quantitative PVE data currently analyzed?

The curren methods are first inventarised based on literature. Then used methods are explained. After short consultations with PVE analysts, a subsection is written on how the choice for an analysis method is currently made. The methods used in the NP RES case study are shown and discussed. The section is concluded with a conclusion in which the research question is answered.

3.1. PVE Analysis Methods

This subsection aims to provide an overview of the analysis methods that are used in current research and business. To achieve this objective, reports and papers are scraped. An explanation of all used analysis methods is given. The search resulted in the publications and reports in Table 3.1.

In the reports found in Table 3.1 several analysis methods can be found. From these methods, only the ones that fall in the definition of quantitative data analysis as defined in chapter 1 are taken into account. The search results in four mainly used analysis methods: general results visualization (descriptive statistics), subcategorical results visualization (descriptive statistics), Multiple Discrete-Continuous Extreme Value Modelling or Portfolio Modelling (choice modelling) and Latent Class Cluster Analysis (LCCA). An overview of methods used per source is given in Table 3.2. Per analysis method a description will be given of the method.

Project	Year	Type	Source
Windenergie Vijfheerenlanden	2022	Populytics Report	de Vries, Spruit, et al., 2022
Omgevingshuis Schiphol	2022	Populytics Report	Mouter, Geijsen, et al., 2022
Medische Rijgeschiktheid	2022	Populytics Report	Populytics, 2022
Langetermijnstrategie Coronabeleid	2022	Journal Article	Mouter, Jara, et al., 2022
Nationale Klimaatraadpleging	2021	Populytics Report	Mouter, van Beek, et al., 2021
Flood Risk	2021	Journal Article	Mouter, Koster, et al., 2021b
Healthy Body Weight	2021	Journal Article	Mulderij et al., 2021
Healthcare Disinvestment	2022	Journal Article	Rotteveel et al., 2022
COVID-19	2021	Journal Article	Mouter, Hernandez, et al., 2021
Flevoland	2022	Populytics Report	de Vries, Mouter, et al., 2022
Nationaal Programma Regionale Energiestrategieën	2023	Populytics Report	Geijsen et al., 2023
Amsterdam Traffic	2021	Journal Article	Mouter, Koster, et al., 2021a
Lelylijn	2023	Populytics Report	Mouter et al., 2023
Thermal Energy Utrecht	2021	Journal Article	Mouter, Shortall, et al., 2021

Table 3.1: Overview of PVE analysis reports used to determine current methods

Source	Type	Method(s)
de Vries, Spruit, et al., 2022	Report	Descriptive Statistics, LCCA
Mouter, Geijsen, et al., 2022	Report	Descriptive Statistics, LCCA
Populytics, 2022	Report	Descriptive Statistics, Choice Modelling
Mouter, Jara, et al., 2022	Article	Descriptive Statistics, Choice Modelling, LCCA
Mouter, van Beek, et al., 2021	Report	Descriptive Statistics, LCCA
Mouter, Koster, et al., 2021b	Article	Descriptive Statistics, Choice Modelling
Mulderij et al., 2021	Article	Choice Modelling
Rotteveel et al., 2022	Article	Choice Modelling
de Vries, Mouter, et al., 2022	Report	Descriptive Statistics, LCCA
Geijsen et al., 2023	Report	Descriptive Statistics, LCCA
Mouter, Koster, et al., 2021a	Article	Descriptive Statistics, Choice Modelling
Mouter et al., 2023	Report	Descriptive Statistics*, LCCA
Mouter, Shortall, et al., 2021	Article	Descriptive Statistics

Table 3.2: Overview of methods used for quantitative data analysis per PVE project. Note: the asterisk indicates that subcategorical results visualizations were used

3.1.1. Descriptive Statistics

Descriptive Statistics are general statistical measures that describe a dataset (Lee, 2020). It includes for instance averages, means and deviations. This is a form of standard and basic descriptions, that give an overseeable set of widely used statistics about a dataset. This means they are often easy to understand for non-specialists, and provide comprehensible information.

Descriptive statistics are used in almost every academic paper published about PVE so far. For instance, the method of descriptive statistics was used by Mouter et al. in their study on COVID-19 measures acceptability (Mouter, Jara, et al., 2022). The amount of respondents who found a certain measure to be acceptable was plotted per measure. It is a good example of a general statistic that describes the answers of a complete set of respondents.
In the projects as performed by Populytics, descriptive statistics is always used as a method. It is accessible and gives a very clear and general image of the average opinion of respondents on the choice task. An example of the descriptive statistics used by Populytics is the average amounts of points given by respondents on a choice task.

The insights that are gathered using descriptive statistics are mainly high-level insights about the average responses in the PVE choice task. Often this gives a ranking of how often a choice task option was chosen, or how many points were awarded. Decision-makers can thus see which measures or choices would be popular among the population. These insights are important decision-making information. The main benefit of descriptive statistics is that they are a relatively simple concept in statistical analysis (Lee, 2020). Therefore descriptive statistics are often easy to understand for policy makers, as well as the general public. A disadvantage is that they are not detailed. They do not give much information about the spread or diversity within results. The dataset usually is much richer than simple descriptive statistics can describe. Therefore additional analyses can be done in case more decision-making information is wished for by the policy maker.

Subgroup descriptive statistics are almost never seen in the literature study, except for a recent report of the Lelylijn (Mouter et al., 2023).

In general, descriptive statistics are always used when evaluating a PVE. They give a simple, yet effective summary of the respondents answers. They are time-effective, and can be easily communicated to non-specialists.

3.1.2. Latent Class Cluster Analysis

Latent Class Cluster Analysis (LCCA) is a method used to divide the set of respondents in a predetermined amount of groups. The aim of the LCCA is to create groups (clusters) in the data that have much homogeniety within a cluster, and a large amount of heterogeneity compared to other clusters.

LCCA takes the dataset and tries to fit a pre-set amount of clusters in the data. It uses observed variables to fit latent variables, hence the name. Usually, the amount of clusters is based on a trial and error method, with fit indicators used to find the amount of clusters that best represents the data. Since LCCA is rather flexible and good at handling uncertain data, it is often used as an analysis method in social studies (Lezhnina & Kismihók, 2022). A large downside is that the amount of clusters has to be set manually, and there is still debate on the best way of performing that iterative process.

An example of LCCA performed at Populytics was at the Vijfheerenlanden project (de Vries, Spruit, et al., 2022). Four groups were distuingished in the dataset, all having their own characteristics. One of the conclusions that could be drawn from the LCCA was that roughly half of the participants had an opinion very close to the average opinion. This gives decision makers a lively image of the groups present in the PVE respondents.

In academic works, the LCCA has been used as well to evaluate PVE data. An example is the application of LCCA in the long term COVID-19 strategy PVE (Mouter, Jara, et al., 2022). An example of conclusions drawn from the LCCA was that there are four groups that can be distinguished in their first scenario. One of these clusters mainly advises to use as many measures as possible, whereas there was also a cluster that wanted none of the measures. In the first cluster, women and elderly people were overrepresented. This gives valuable information about the spread of opinions in the sample.

The type of insights gathered using a LCCA analysis, is an overview of the groups present in the sample. Sometimes these results can show that certain groups within the society might need more attention or a different approach. It can diversify the view a decision maker has on the inhabitants opinion. Within these groups, demographic information is given when significant, to clarify the image of such a group.

The advantage of LCCA analysis is that it is relatively fast and easy to perform using software like Latent GOLD. Computing times are quick and the software is rather intuitive to work with for a trained analyst. The analysis times for the analysis are short, and the analysis gives a clear division of groups. The amount of groups is usually around three or four, which is still an insightful number of groups for a decision-maker to work with. A disadvantage is that you are structurally bound to the amount of groups you insert as analyst. There might in reality not be the same amount of groups in the dataset as inputted by the analyst. The LCCA can give the decision-maker the idea that certain clear-defined groups exist, which might not always do just to the diversity in the dataset.

Generally, an LCCA provides a few groups within the data, that can give the decision-maker a richer view on the diversity among respondents than by only checking averages. The groups can help the decision-maker sketch an image of what groups in society look like, and which groups might require a different approach or solution to the policy problem. The LCCA method is however rather inflexible, and sometimes does not do just to the diversity in a dataset.

3.1.3. Choice Modelling

In choice modelling, an attempt is made to reveal relative utilities, which can serve as a scale to measure which options that were available in the PVE are valued most by citizens. An attempt can then be made to construct a maximum utility portfolio, which would thus be the optimal choice. In order to construct a quantitative model, the analysts has to provide relevant interactions. The computer will then estimate the best parameters to fit the model to the data. The outputs are relative utilities.

The field of choice modelling stems from consumer modelling. It originated in economics, and was posed by McFadden to help observe consumer utilities (McFadden, 1974). Choice modelling is a well-documented and researched field of economics, in which there are handbooks available to anyone who wants to get involved with choice modelling (Hess & Daly, 2014). The method has matured, and models have been developped to apply to PVE experiments (Bahamonde-Birke & Mouter, 2019).

Choice modelling has been used as a method in multiple scientific publications in which PVE was the method. An example of this is Mulderij's research on promoting healthy body weight (Mulderij et al., 2021). In here, the Multiple Discrete-Continuous Extreme Value (MDCEV) model as proposed by Bahamonde was used (Bahamonde-Birke & Mouter, 2019). It was concluded that promoting fruit and vegetables and sports vouchers were popular among respondents.

Using choice modelling, optimal portfolios can be computed. This is important information to decision-makers, helping them choose which policy options to choose. Furthermore the method is robust and grounded in economic theories. It is observed that most scientific publications with PVE as method use choice modelling in their publication to provide detailed information on preferences of respondents.

In general, choice modelling provides robust results for the preferences of respondents. A downside of the method is the time it takes for models to converge, which can be up to several hours for larger models. Furthermore it is relatively difficult for analysts to work with the method, as it requires specialist skills. Another downside is that the analyst has to identify the relevant interactions beforehand (van Cranenburgh et al., 2022). This creates biases that might negatively influence the models reliability. Recent papers have suggested using machine learning to assist in the specification of the models, to decrease the modellers bias (Hernandez et al., 2023).

3.2. Reasoning behind Analysis Method Choice

How is the choice of method currently made? The answer depends on the type of analysis that has to be conducted. When performing an academic research, different choices are made then when the research is based in commercial companies. Both situations are described below. The commercial PVE description is based on a personal conversation (e-mail) with a Populytics analyst. Since the e-mail was in Dutch and a reaction to some questions, the contents are translated to English, interpreted and made into a full story by me.

In commercial PVE usage, there are currently two prerequisites for the choice in analysis method. First of all, expertise and experience with an analysis method plays an important role. If none of the employees has affinity with a certain method, a method is difficult to implement. It would take time to train the employee on a new method, which costs a company money in the short term. If a method is promising, this effort will be made on the long term. The other prerequisite is time available. If a method takes up more time, it will only be performed if the costs are covered by the client. If enough additional insights can be gained from the analysis, this will be done. The expected match with data and number of insights produced therefore play a role. Finally, the type of data plays a role. One PVE is not the same as the other. Sometimes they might have varying attribute levels, which requires more advanced techniques than non-varying attribute levels. Descriptive statistics in generally are always performed, although not always the same subtype of descriptive statistics. Populytics almost always performs an LCCA. They do this as it has delivered valuable insights to policy makers in the past, and they have employees who regularly run LCCA's, lowering the threshold to perform them. MDCEV analyses are usually performed when attribute values are varying.

In the academic literature on PVE experiments, it can be observed that descriptive statistics are always used as an analysis method. Furthermore MDCEV or Portfolio Choice Models are used in almost all cases to support the preference elicitation. LCCA is used less often in academic literature. It seems that for academic experiments, modelling is more often used than in business-related experiments. Choice modelling takes a relatively large amount of time, making it less suitable for business use. The difference between academic usage of methods, and business usage seems to stem from the importance of time and costs in business. Methods that take more time, are less favourable for use in business.

3.3. Case Study Methods

For the case study, the used methods are evaluated in more detail. Two analysis methods have been performed so far on the NP RES case (Geijsen et al., 2023). These are descriptive statistics and LCCA. The first one produces general descriptive statistics, they are the direct aggregate results of the PVE experiment. The LCCA is used to find clusters and describe those clusters in terms of demographic characteristics to decision-makers (Geijsen et al., 2023). As the LCCA is a further analysis on the data provided by PVE, instead of a description, the results will be more interesting to compare to SHAP. Extra focus will thus be on gathering the insights of the LCCA in the case study.

The first analysis method used on the NP RES case was the descriptive statistics (Geijsen et al., 2023). The descriptive analysis resulted in a relative ranking of importance of the value-based statements in the NP RES PVE. The average amount of points is plotted per choice option, as well as the distribution in points allocation by participants. The ranking is as follows, in which the first entry is the most chosen statement:

1. Energy prices for Dutch citizens should be as low as possible
2. The existing nature should be conserved as much as possible
3. The existing landscape should be conserved as much as possible
3. The chance of grid failures should be as low as possible
4. Citizens should have the least amount of nuisance possible
5. Citizens that experience nuisance should be compensated with extra money
5. Regions that produce more clean energy, pay less for energy
5. Citizens are allowed to exert influence as soon as possible

Furthermore the conclusion is drawn that the diversity and spread in citizen choices is large (Geijsen et al., 2023). For example, although the energy prices statement is chosen most, still 24% of respondents thinks it is the least important choice option.

The LCCA showed three clusters of respondents. The largest cluster, consisting of roughly 60 percent of respondents, was close to the average respondent. They can be seen as the majority group. Two other clusters that consisted of 20 percent of the respondents were identified. The first cluster was characterized mainly by picking nature and landscape options more often. The average age in this cluster was higher. The last cluster consisted of respondents with very high scores on cost-related choice options. Low-educated respondents were overrepresented in this cluster.

The full results of insights are taken from the Populytics report (Geijsen et al., 2023). They are indicated in Table 3.3.

Insight from LCCA
59% of participants is in a group that closely resembles the average points division
In the average cluster, young participants and people from large municipalities are overrespresented
Gender and financial health are not significant for the clustering
A cluster exists of 21% of participants which is very nature and landscape focussed
Older participants are in the nature-cluster more often
Participants in average-sized municipalities are more often in the nature-cluster
Average-educated citizens are more often in the nature-cluster
A cluster exists of 20% of participants who value low costs, high reliability, low nuisance and high compensation for nuisance.
In the citizen-focussed cluster, low-educated people are more prevalent
In the citizen-focussed cluster, citizens in small municipalities are overrepresented
In the citizen-focussed cluster, participants between 35 and 65 years old are overrepresented

Table 3.3: Insights resulting from LCCA analysis

3.4. Conclusion

The research question that this section aimed to answer is:

> How is quantitative PVE data currently analyzed?

The answer to the question is that in nearly all cases descriptive statistics are used as an analysis method, as can be seen in Table 3.2. Furthermore Latent Class Cluster Analysis (LCCA) is almost always used when experiments are performed in business environments, whereas Choice Modelling is almost always used when experiments are performed in academic environments. In only one case has LCCA been used in academic articles. In only one case has choice modelling been used in a Populytics Report. Subgroup analysis using descriptive statistics has only been used for the Lelylijn report.

In the NP RES case study, descriptive statistics and LCCA have so far been applied. They result in a relative ranking of importance of the choice task options, and in many insights about the relation between clusters of respondents and their demographic characteristics.

4

Results of Machine Learning on NP RES Case Study

The aim of the thesis is to find out if additional insights are gained in PVE analysis when SHAP is applied. To put this hypothesis to test, Machine Learning algorithms have been applied to the NP RES dataset. This has been performed following the methods described in chapter 2. The current chapter aims to answer the following research question:

> ”Which results does Machine Learning provide on a PVE case study?”

4.1. Model Selection Results

As described in subsection 2.4.4, three machine learning algorithms are applied to the NP RES dataset. This is a practical way of comparing results of different algorithms. As explained in subsection 2.4.3, the three applied machine learning models are compared on their mean squared error on the full dataset. The results are described in this section.

The comparison code compares two things. First it compares the three machine learning algorithms internally with different hyperparameters. Then it compares the best instances of all three models with each other. This gives two results: a best instance for every model, and a best Negative Mean Squared Error (NMSE). The higher the NMSE, the better a model performs. In Table 4.1, the results are given.

Algorithm	Hyperparameter	Best instance	NMSE
XGBoost	gamma	2	-8,08
	max_depth	3	
	min_child_weight	5	
Random Forest	max_depth	3	-7,66
	max_features	sqrt	
	n_estimators	100	
Artificial Neural Network	activation	relu	-7,67
	batch_size	300	
	hidden_layer_sizes	30	
	learning_rate_init	0,1	

Table 4.1: Result of algorithm hyperparameter tuning

From the hyperparameter tuning and comparison it can be concluded that the Artificial Neural Network and the Random Forest models outperform the XGBoost model. Although the differences between the ANN and RF model are small, the Random Forest model is chosen as it has the best score. All results presented in the remainder of this chapter stem from the Random Forest model. The hyperparameters used are the optimal ones coming from the hyperparameter tuning.

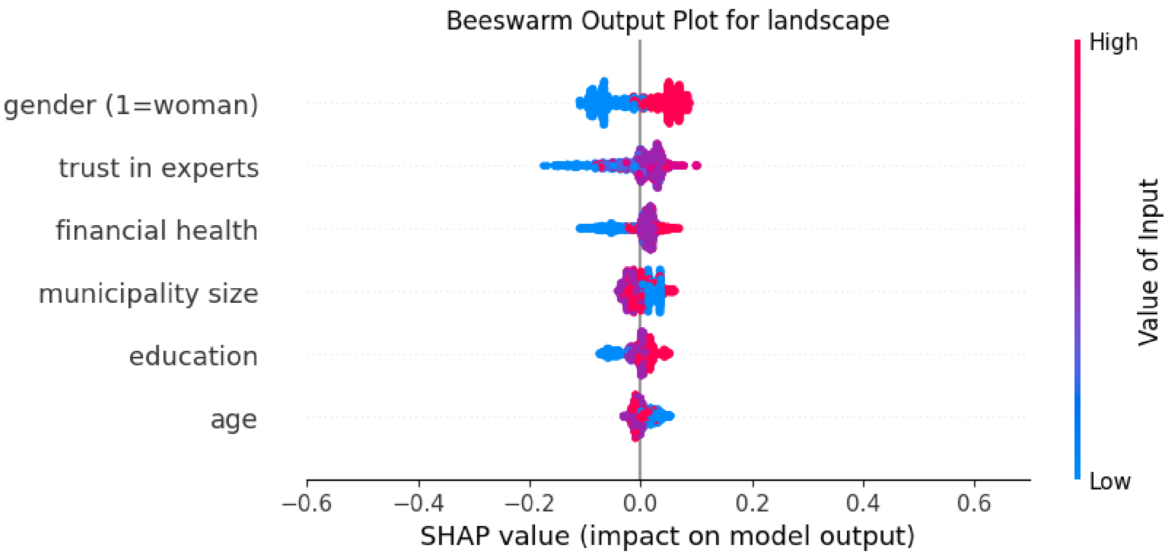
4.2. Interpretation of SHAP Results

Results from a SHAP-analysis are not always easy or logical to analyse. Background information is needed to be able to correctly interpret the resulting plots. In this section all resulting plots from the SHAP-analysis are discussed and an explanation for interpretation is given. This helps the reader in understanding the nature of the results.

4.2.1. Beeswarm Output Plots

There is one Beeswarm Output Plot (BOP) per output of the Machine Learning model. The outputs are the points given to every choice task option in the PVE. One of the outputs is for instance the amount of points given to the first choice option about landscape (”the existing landscape should be conserved as much as possible”). For this output, the BOP can be plotted, as seen in Figure 4.1.

Figure 4.1: Beeswarm Output Plot for landscape choice option



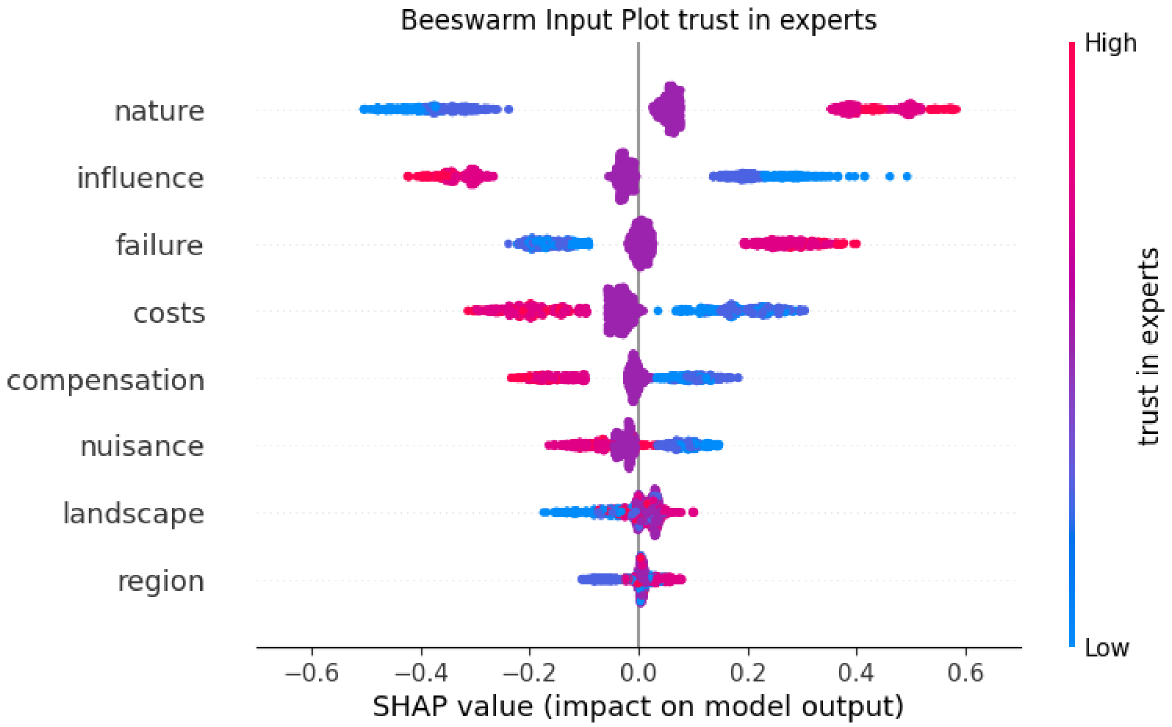
The six inputs of the Machine Learning models, being the answer to six demographic questions, are all plotted. For every respondent, the impact of an input (demographic variable) on the model output is given. If an input had a large negative or positive effect on the choice made by the respondent, a high absolute SHAP value is noted on the x-axis. If a SHAP value is high and positive, this means that said input had a large positive effect on the amount of points given by the respondent. One thing to note is that this effect is not a real-world effect, but an effect predicted by the machine learning model. The input that comes first in the output plot, is the one with the highest average effect on the model output. Thus, this input is more important for the choice of respondents than inputs ranked lower. Every dot on the chart is one respondent. For every respondent, the position on the x-axis gives its SHAP-value. The color of the dot gives the input value for the respondent.

As this can be quite difficult, we describe one example. From Figure 4.1, we can see that gender is the top ranking input. This means that on average, gender has the highest importance in determining a respondents choice. If a dot is blue, the respondent is male, if the dot is red, the respondent is female. Most red dots are to the right of the 0-line. Therefore, for almost all respondents, the fact that they are female, made the model predict that they would give more points to this output option. Thus it can be concluded that females have a tendency to give more points to the landscape-statement. From the spread, we can see that the distinction between male and female is quite sharp. Furthermore, the concentrations of both male and female SHAP-values are quite far from the 0-axis, indicating that the respondents with higher values are not outliers.

4.2.2. Beeswarm Input Plots

There is one Beeswarm Input Plot (BIP) per input of the Machine Learning model. The inputs are the answers to demographic questions, such as gender or age. For one input, its effects on all outputs is given. One of the inputs is to which extent expert opinion or citizen opinion should play a role in the government decisions. A high value (red dot) indicates a high trust in expert opinion. A low value (blue dot) indicates a low trust in expert opinion. For the advice-input, the BIP is given in Figure 4.2.

Figure 4.2: Beeswarm Input Plot for advice parameter



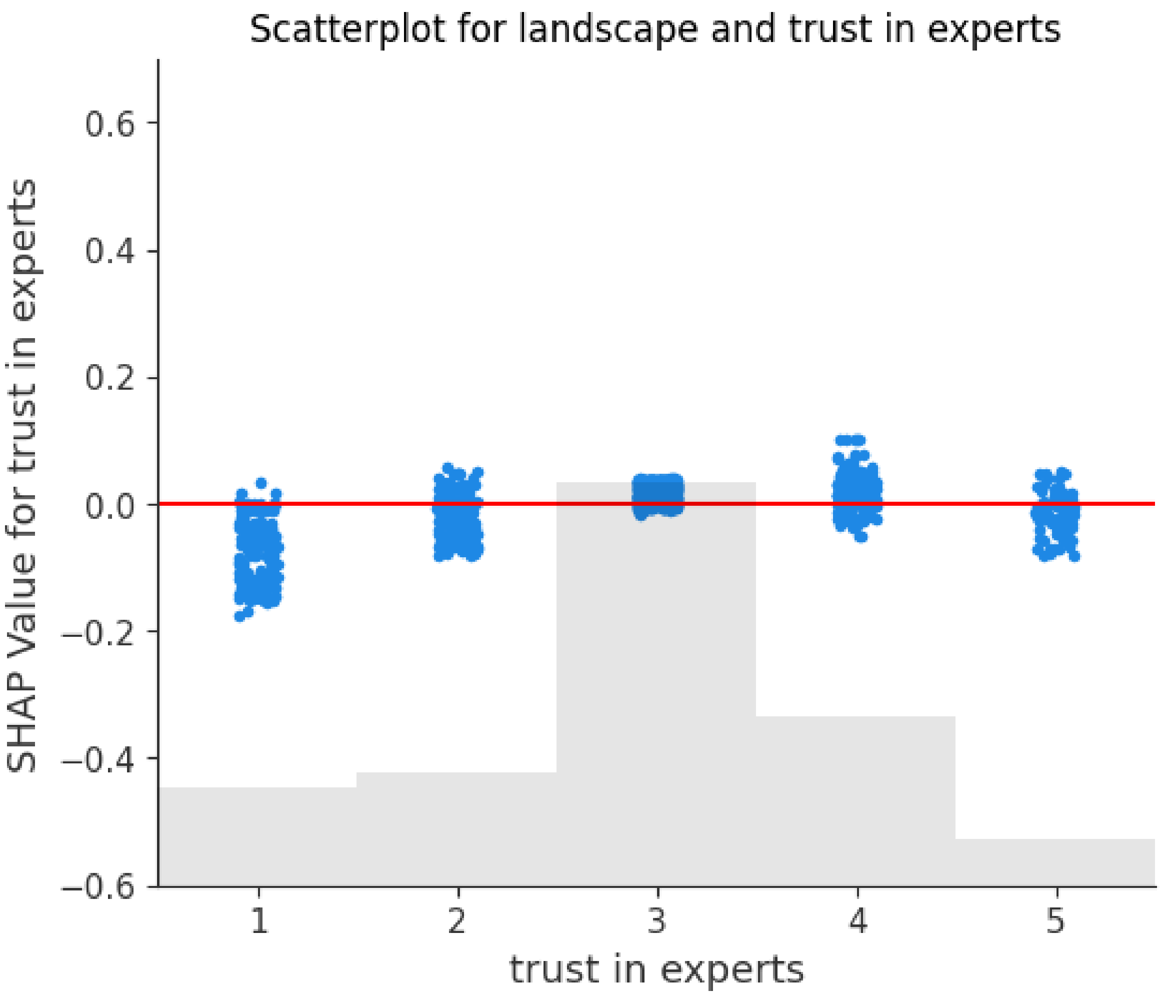
In the BIP, all 8 model outputs, being the choice task options, are plotted. For every respondent, one dot is plotted on all eight output lines. The distance from the 0-axis indicates the size of the SHAP-value per respondent. If a SHAP value is high and positive, this means that said input had a large positive effect on the amount of points given by the respondent. One thing to note is that this effect is not a real-world effect, but an effect predicted by the Machine Learning model. The output that comes first in the input plot, is the one on which the input had highest effects. Thus, this input is more important for the choice of respondents for the top output, than for the others. Every dot on the chart is one respondent. For every respondent, the position on the x-axis gives its SHAP-value. The color of the dot gives the input value for the respondent.

As this can be quite difficult, we describe one example. From Figure 4.2, we can see that the nature choice option (”the existing nature should be conserved as much as possible”) is most influenced by the advice-input. This means that whether or not someone thinks the government should rely on expert or citizen opinion is a good predictor of someones choice for nature. Respondents with low trust in experts and high trust in citizens opinion (low, and thus blue values), generally find nature less important and reward less points to the nature choice option. From the spreads we can see that for most respondents in the middle of the advice-scale, there is only a small effect on the choice for nature compared to the average respondent. Respondents who are on the extremes of the advice-scale, have much higher absolute SHAP-values. For most choice tasks, there is a cluster of respondents who said citizen and expert opinion are equally important in the middle of the chart. The blue and red dots, with respondents on the sides of the advice-scales, generally are on opposite sides of the BIP. We can conclude that respondents who indicate that we should trust expert opinion more, give more points to nature, and preventing failure risks. They give less points to local influence, costs and compensation options. For the nuisance, landscape and region option, the differences between clusters start to get rather small, so those are less relevant to analyze.

4.2.3. SHAP Scatterplots

Besides the input and output plots, there are also SHAP Scatterplots that plot the SHAP-values per combination of input and output. This can be plotted for every combination of input and output, 48 in total. An example is the SHAP Scatterplot of the effect of the advice-parameter (input, trust in experts or trust in citizens) on the landscape choice option (output). This example can be seen in Figure 4.3.

Figure 4.3: SHAP Scatterplot for advice parameter and landscape choice option



The SHAP Scatterplot, plots the SHAP-value for every respondent. It gives an overview of the effect of the input-parameter on the output choice option for every respondent. Every respondent (dot in the chart) has a unique value. If that value is low, it means that the amount of points given by the respondent to the choice option, was negatively influenced by the demographic variable, and vice versa. If the SHAP value has a high absolute value, the effect is stronger than when it has a low absolute value. For respondents on the 0-line, the effect of the demographic variable on the output choice was negligible. There is a grey bar chart in the SHAP Scatterplot. It gives an overview of how many respondents are in every vertical column. If the grey bar chart is highest, that means most respondents are in that category.

As this can be abstract, an example is provided. We can analyze the SHAP Scatterplot given in Figure 4.3. By looking at the grey bar charts, one can see that most respondents are in the category that thinks expert and citizen advise should be equally taken into account. Respondents in categories 1 or 2 think citizen advise is more important. Respondents in categories 4 and 5 believe expert advise is more important. This specific chart does not give a very clear pattern. There does not seem to be a large influence of the advice-parameter on the landscape choice option. Only people who think we should fully rely on citizen advice tend to have a lower SHAP-value, and are thus less likely to assert points to the landscape choice option. Another observation is that the categories with less respondents, have a wider spread of SHAP Values. Category 3 has the most respondents, and a more crisp SHAP distribution. As there are more people in this category, the machine learning algorithm could give a better prediction of the effect between input and output. The results in categories 1, 2, 4 and 5 are therefore less reliable than the ones in category 3.

4.3. SHAP Results

The full results are provided in this section, along with observations and conclusions that can be drawn from the resulting plots.

4.3.1. Resulting Plots

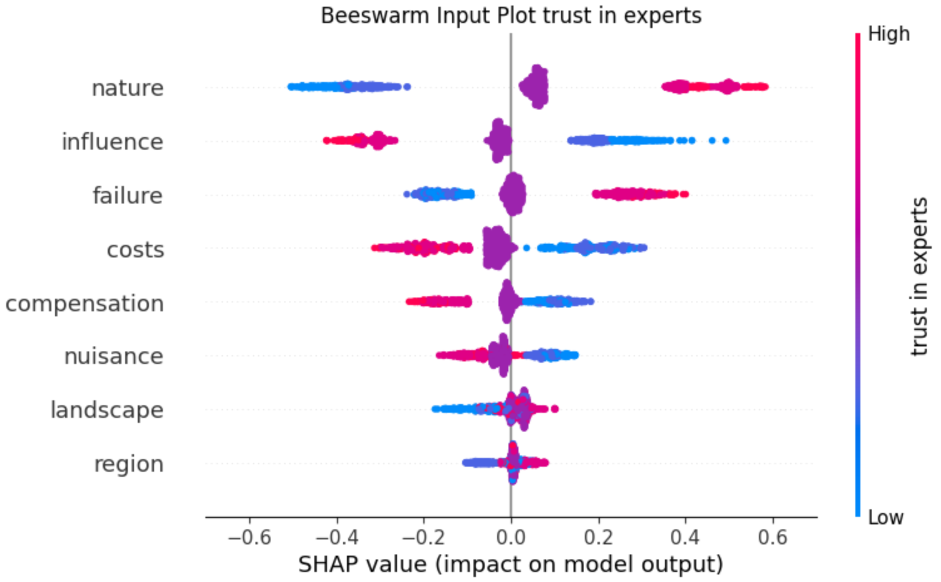
The result of the SHAP analysis is a total of 62 charts. To analyse all of these is time-consuming. The 48 SHAP Scatterplots contain no information that is not given in the input or output plots. However, it is zoomed in on one specifici input-output combination. These charts will thus be used to clarify interesting observations from the BIP and BOP charts if necessary. The principal analysis is done by analyzing the BIP and BOP charts. In this section these charts are analyzed one by one. If a clarification is needed, the SHAP Scatterplot will be used for further investigation.

All plots can be found in an Appendix A

4.3.2. Observations from Plots

In this subsection, all Beeswarm Input Plots (BIP) and Beeswarm Output Plots (BOP) of the SHAP algorithm are presented and analyzed. The SHAP Scatterplots are used for clarification when necessary. Firstly, all BIP's are given.

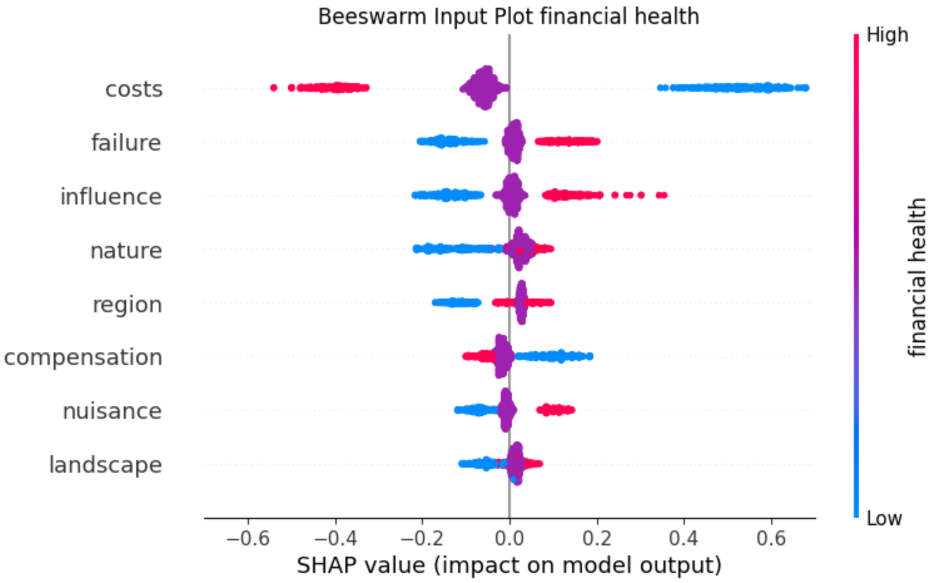
Figure 4.4: Beeswarm Input Plot for advice parameter



The first BIP is the chart of the advice-input. Respondents with a low score (blue) value citizen advise over expert advise. Respondents with a high score (red) value expert advise over citizen advise. The purple respondents are in the middle.

From Figure 4.4 it can be seen that citizens who rely on experts choose the nature option relatively often. A possible explanation is that experts often explain that the state of our nature is deteriorating, where this is difficult to see with your own eyes. Citizens who value expert advise more also value decreasing grid failure risks. Contrary, people with high trust in citizen opinion value local influence and lower energy costs. What is interesting to see is that category 1 (lightest blue, full reliance on citizens) and category 2 (more emphasis on citizen advise than expert advise) barely differ. The same can be said for categories 4 and 5. As these groups are relatively small already, it can be advised to combine them into one category for future machine learning analyses to provide more reliable predictions.

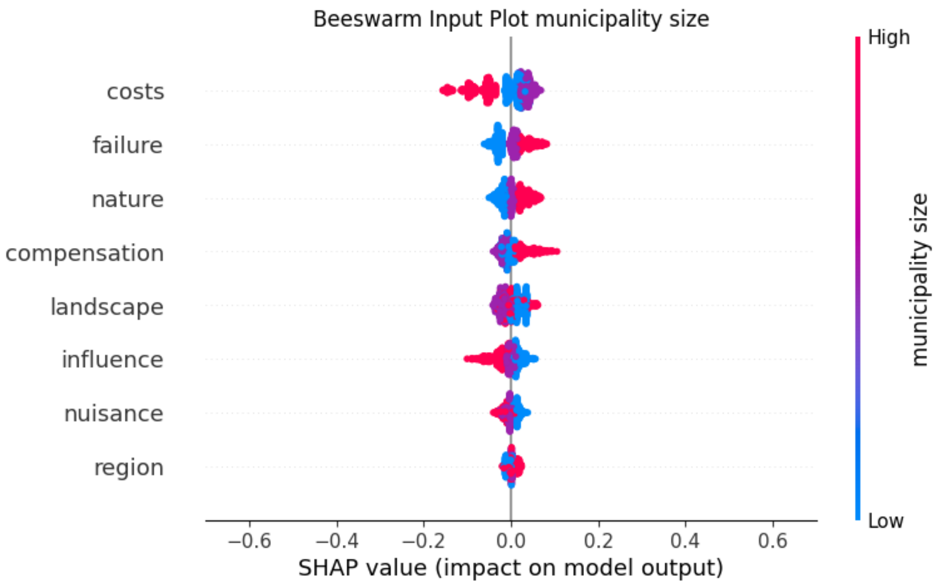
Figure 4.5: Beeswarm Input Plot for financial parameter



The second BIP is the chart for the financial input. Respondents with a low score indicate that they are in a poor financial situation. Respondents with a high score have more than enough money. Most respondents are in the middle category.

From Figure 4.5 it can be concluded that the effect of financial situation on the energy choice option is very strong. Respondents in worse financial situations value a low energy price highly. That is a rather intuitive result, and thus gives us confidence in the SHAP analysis. Respondents in better financial situations want more local influence and a reduction of grid failure risk than other respondents. People in worse financial situations give less points to nature, regional compensation, local influence and failure risks. They give more points to local compensation for nuisance. They seem to value the financial choice options more than citizens in better financial situations.

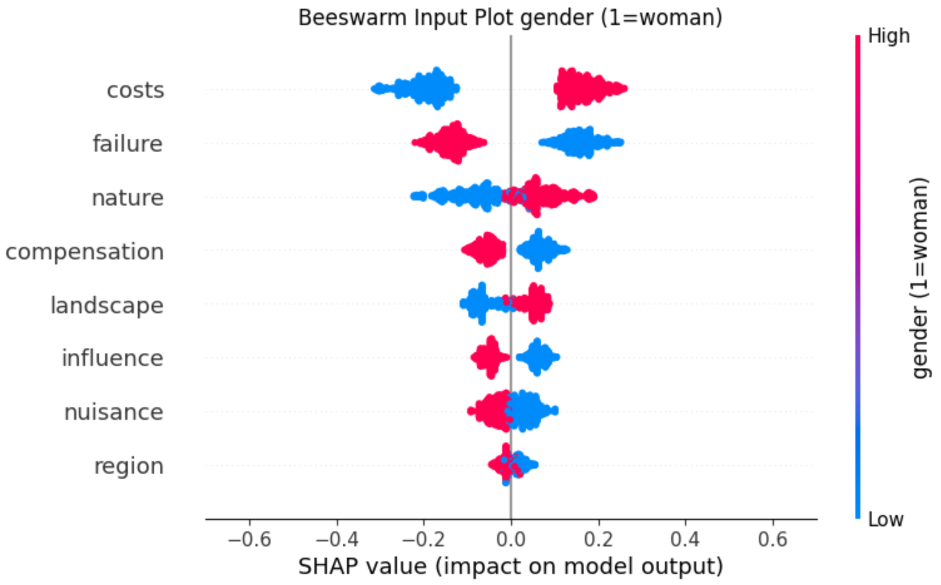
Figure 4.6: Beeswarm Input Plot for municipal parameter



The third BIP (Figure 4.6) contains the effect of the size of a respondents municipality on its choices. A larger municipality is a higher value. The red respondents live in the largest municipalities, the blue respondents in the smallest.

The effect of municipality size on most choice options is relatively small. This is interesting, since often discussion in the Netherlands are held about the gap between the urbanized Randstad and the rest of the country. However in this PVE, there is almost no effect of municipality size (which tend to be larger in the Randstad) on the choices of respondents. This data thus does not show a large gap. The largest effect of municipality size was on costs. People who live in large municipalities give less points to energy costs.

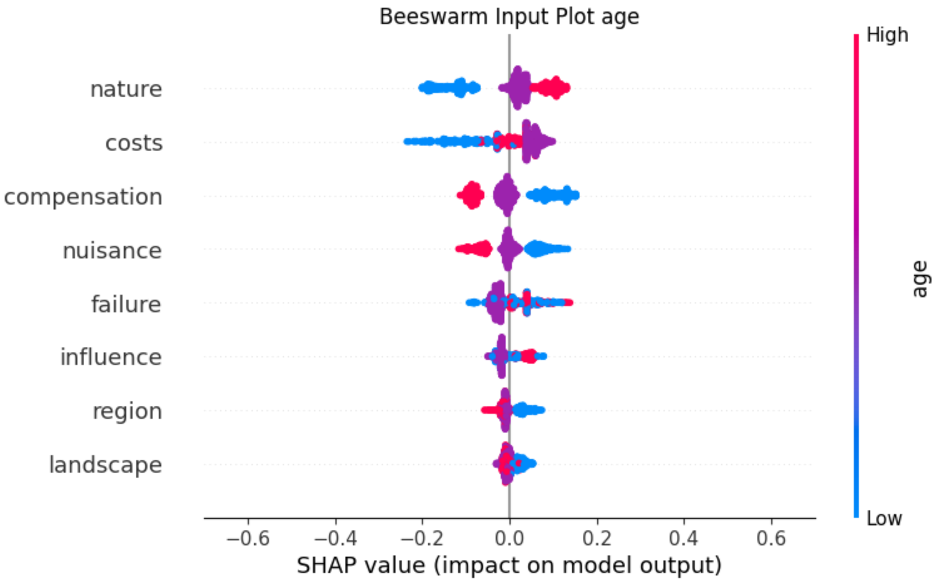
Figure 4.7: Beeswarm Input Plot for gender parameter



The fourth BIP plots the effect of gender on the choice task. Red respondents are women, blue respondents are men. All other categories were too small to make predictions on and were thus left out of the machine learning algorithm.

From Figure 4.7, it can be concluded that gender has quite a large influence on the choice made by a respondent. Women value low energy costs more than men, but men value grid reliability more. Furthermore women give more points to the nature choice option.

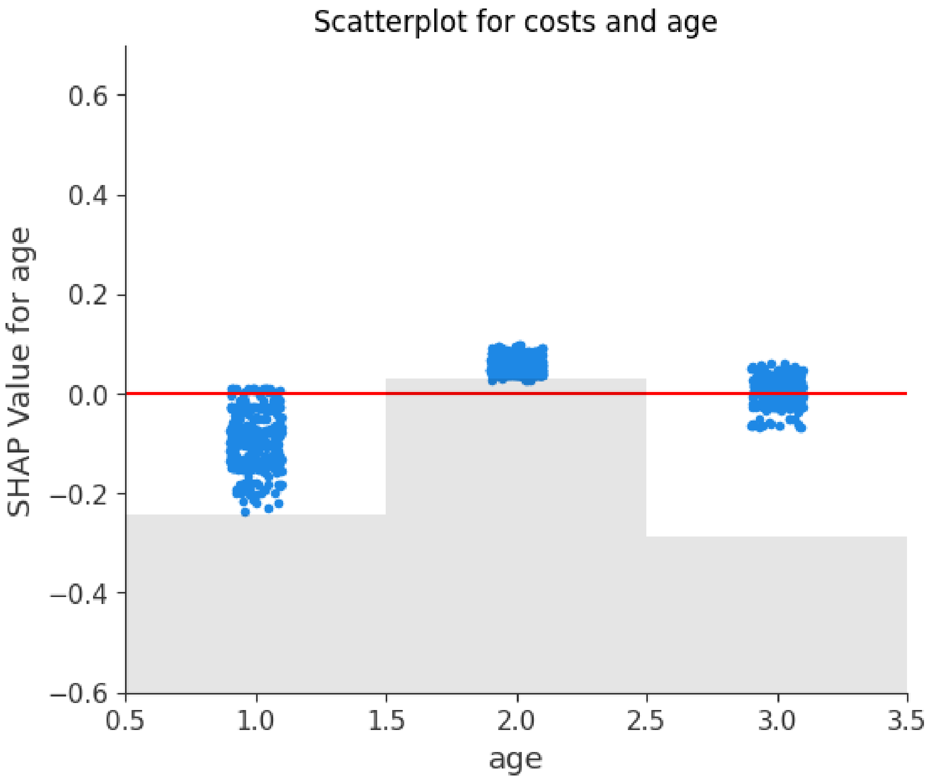
Figure 4.8: Beeswarm Input Plot for age parameter



The fifth BIP gives the effect of age on the choice task. Higher values (red respondents) indicate higher age, and vice versa.

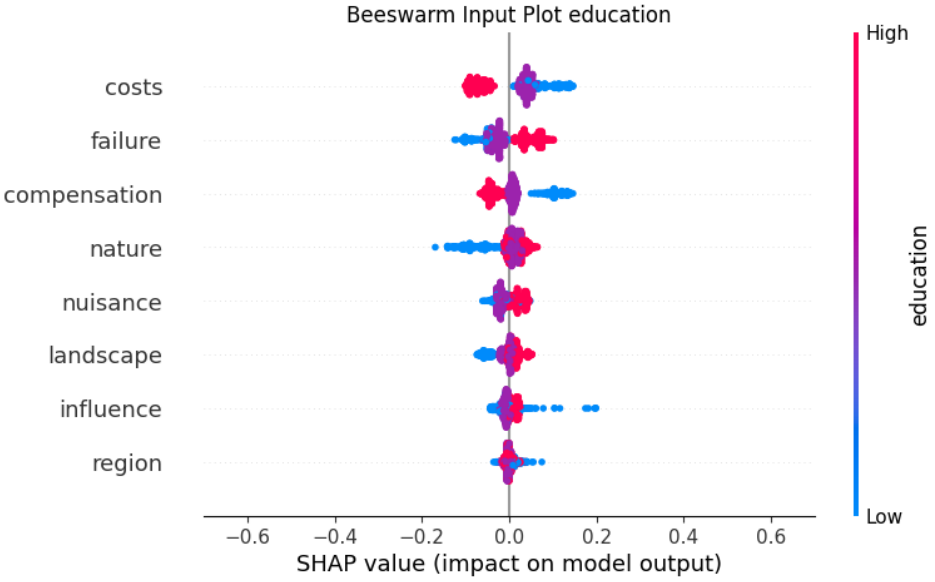
From Figure 4.8 it can be concluded that young people give less points to nature. To young people, compensation for nuisance and nuisance itself are more important than to older people. The effect of age on the energy cost choice task seems to be difficult to explain based on this chart only, therefore the SHAP Scatterplot is plotted in Figure 4.9.

Figure 4.9: SHAP Scatterplot for age parameter and cost choice option



From the SHAP Scatterplot in Figure 4.9 it can be seen that young respondents almost all have negative SHAP values. Thus their young age predicts that they give less points to energy costs. However, respondents in the middle age category are predicted to value energy costs relatively more than respondents in the oldest age category.

Figure 4.10: Beeswarm Input Plot for education parameter

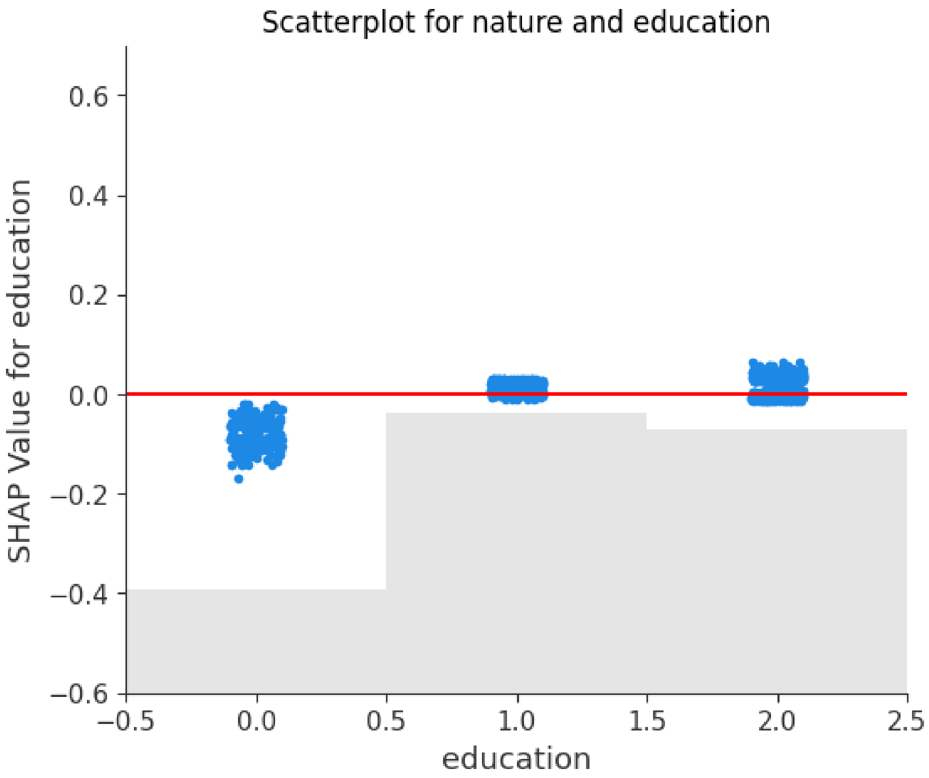


The last BIP plots the effect of education level on the prediction of the machine learning algorithm. If a respondent has a high value (red dot) it means it has had high education.

From Figure 4.10 it can be concluded that educational level is not one of the strongest influences on the choice made by respondents. One of the effects that can be seen is that high-educated respondents give less points to the energy price level choice option. They tend to award more points to the reliability of the grid.

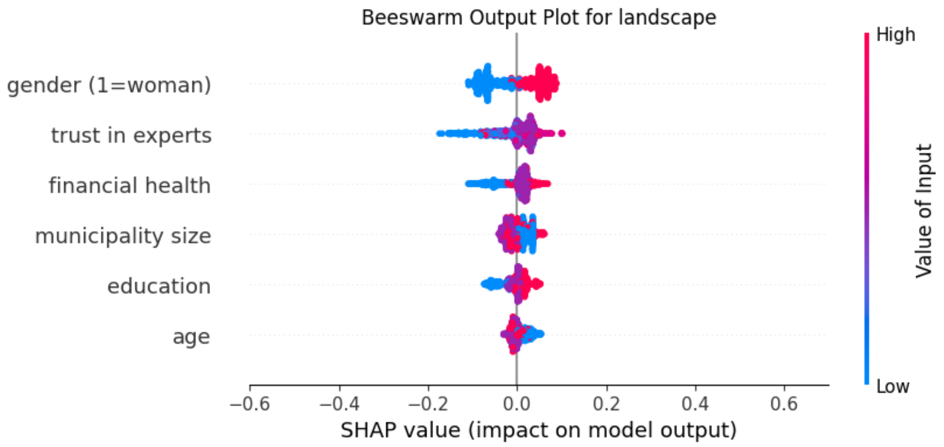
In many of the beeswarms, it appears that there are nonlinear effects at play. To zoom in, the scatterplot of education to nature is evaluated.

Figure 4.11: SHAP Scatterplot for education parameter and nature choice option



As can be seen in Figure 4.11, the amount of respondents in the low-educated category is low. That is an explanation for the sparsity and diversity in SHAP values for this category. The larger categories are more coherent, and show no relevant differences between groups.

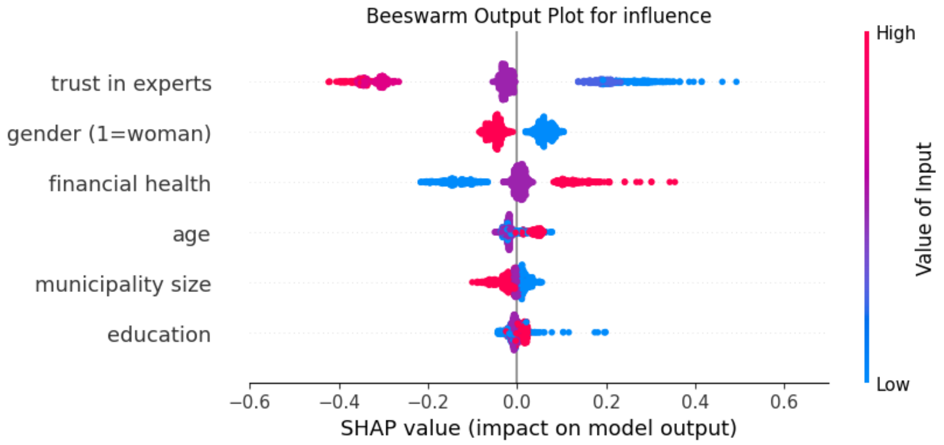
Figure 4.12: Beeswarm Output Plot for landscape choice option



The first analyzed BOP is that of the landscape choice option.

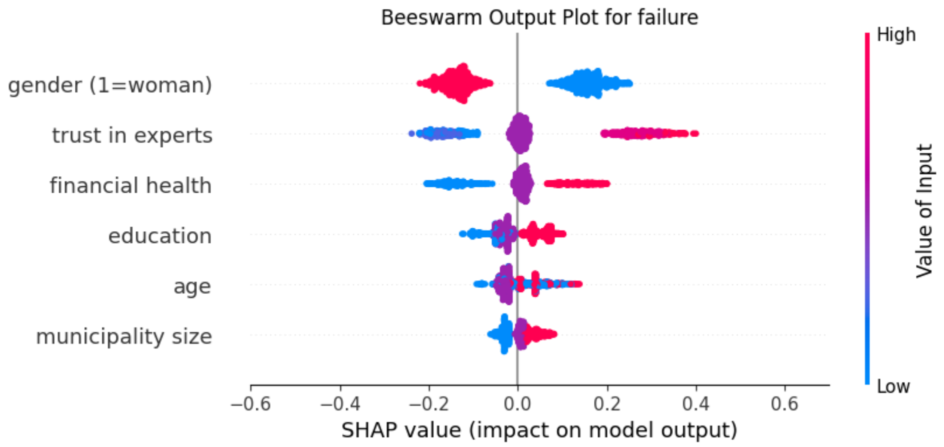
As can be seen in Figure 4.12, the diversity in choices for the landscape option is mainly determined by gender, in which female respondents choose landscape more often. Trust in experts increases the chance that a respondents awards points to this choice option. Most SHAP values are relatively low, thus the effect of the measured demographic variables on the choice for the landscape choice option is weak.

Figure 4.13: Beeswarm Output Plot for influence choice option



The second analyzed BOP (Figure 4.13) is that of the influence choice option. Respondents who give many points to this option, believe that it is important that local residents have much influence in the process. The main predictor for whether respondents pick this option is their trust in experts compared to citizens. That is an intuitive result, providing confidence in the model. Respondents who believe citizens opinion is more important than expert opinion want local citizens to have more influence in the decision-making process. Male respondents choose this option more often as well, as do people with better financial situations.

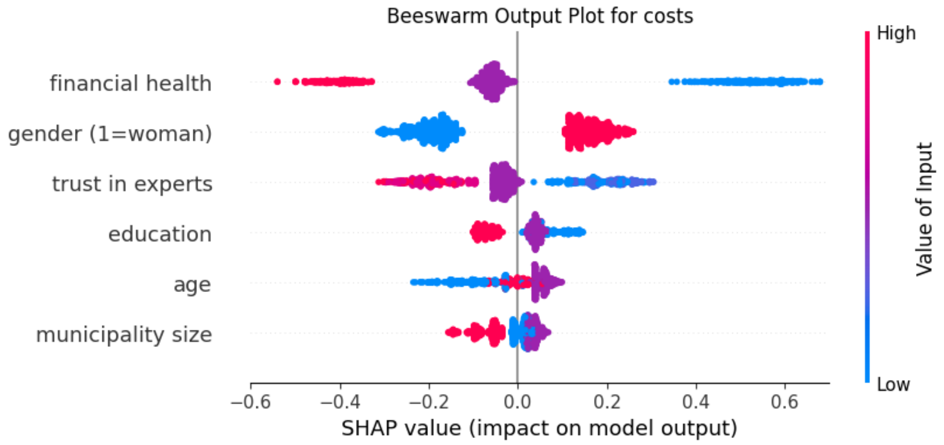
Figure 4.14: Beeswarm Output Plot for failure choice option



The next BOP is that of the failure choice option, indicating that reliability of the grid should be as high as possible.

The main predictor for whether this option is chosen is gender, along with trust in experts, as can be seen in Figure 4.14. Male respondents more often choose this choice option, as do people with high trust in experts and a good financial situation.

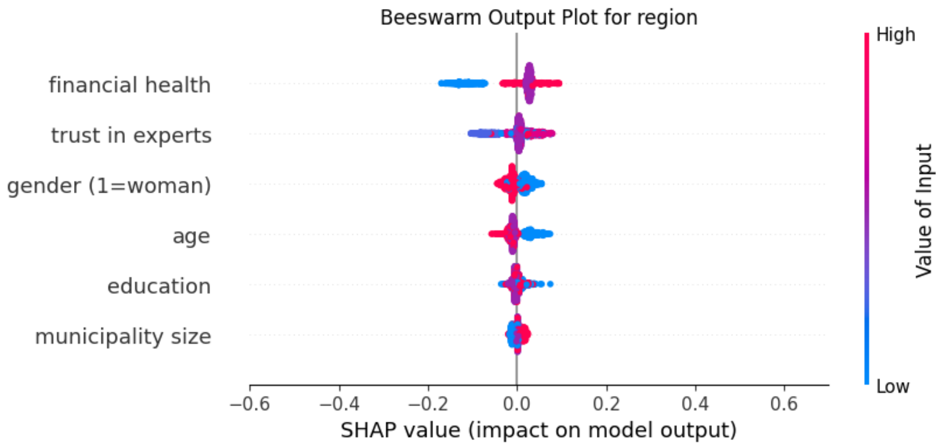
Figure 4.15: Beeswarm Output Plot for landscape cost option



The BOP of energy price is the BOP with the most significant effects, as can be seen in Figure 4.15. SHAP Values in this chart are the highest of all BOP's. That means the demographic variables in the PVE have a high influence on whether respondents choose this choice option.

Respondents in worse financial situations award more points to the energy price option. For them, energy prices are more important. Although intuitive, it gives confidence that this relation is found in the data. Furthermore female respondents, as well as respondents who believe we should rely on citizen advise more than expert advise give more points to this choice option.

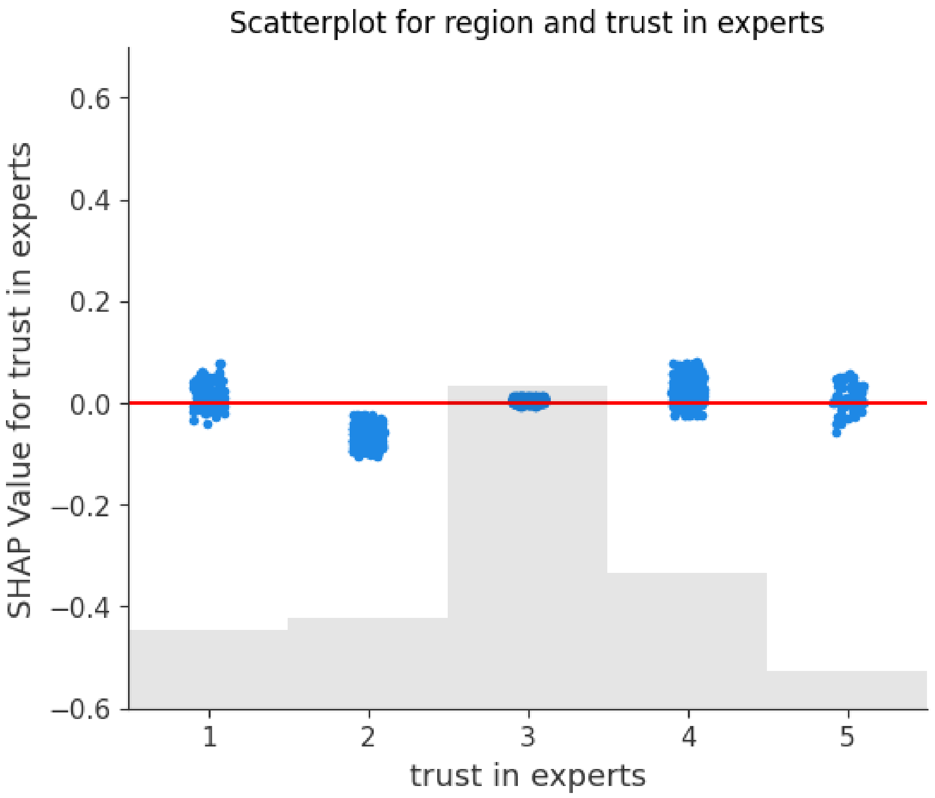
Figure 4.16: Beeswarm Output Plot for region choice option



The next BOP, seen in Figure 4.16 explains the choice option predictions for whether regions who will produce more renewable energy should profit more from the profits made. People in worse financial situations choose this option less often than respondents in good or okay financial situations.

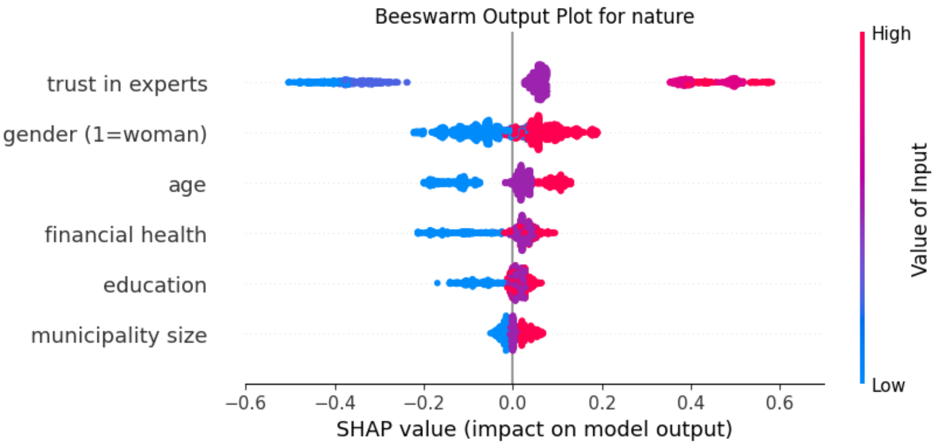
What is striking in this BOP is that the effect of municipality size on this choice option is negligible. Smaller municipalities are often more rural and have more space for renewable energy. However, this does not mean that their inhabitants choose this choice option more often. The SHAP beeswarm of the effect of the advice-parameter on the region choice option is unclear. As it is ranked second highest in this BOP, it is evaluated in more detail.

Figure 4.17: SHAP Scatterplot for advice parameter and region choice option



As can be seen in Figure 4.17, the large spread in SHAP values does not reveal a strong pattern. Rather, it seems to be due to the small categories with few respondents that there is a lot of uncertainty here. No relevant policy conclusions can be drawn.

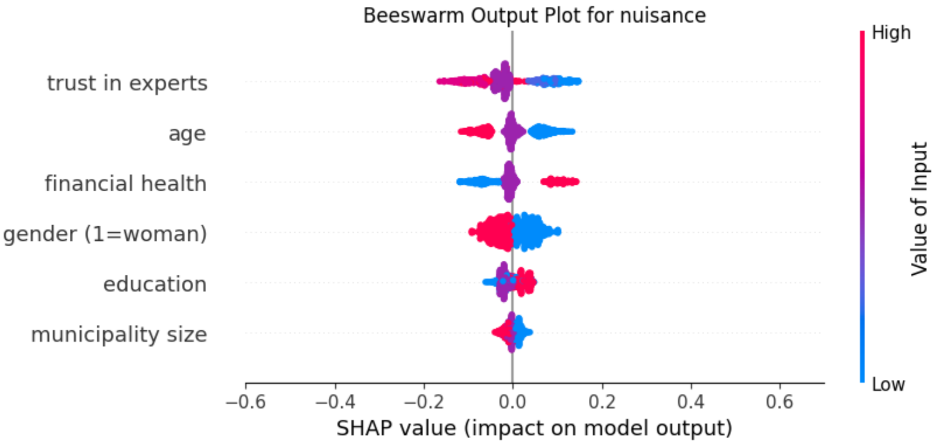
Figure 4.18: Beeswarm Output Plot for nature choice option



The next BOP plot explains the prediction of the model for the nature choice option.

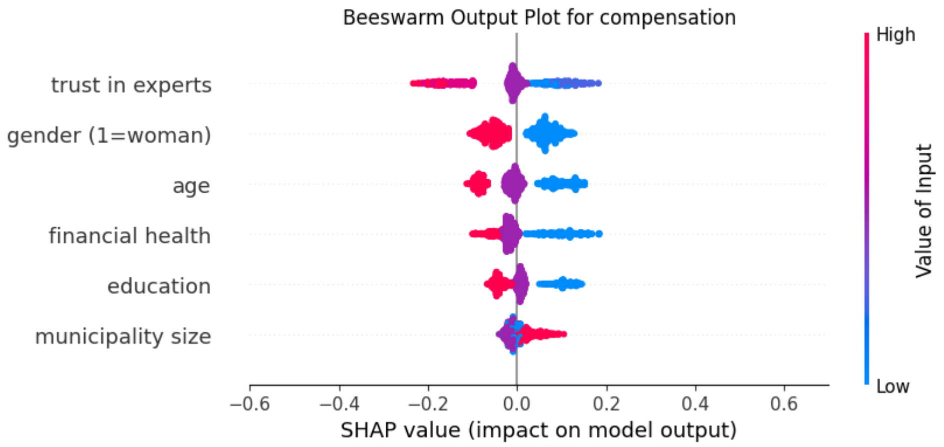
The results can be seen in Figure 4.18. The biggest predictor for the amount of points awarded to the nature choice option is whether a respondent values citizen advice over expert advise or the other way around. Respondents who rely on expert advise award more points to the nature choice option. This effect was seen in the BIP for advice as well, and has already been discussed there. All other effects have been evaluated in the respective BIP's as well.

Figure 4.19: Beeswarm Output Plot for nuisance choice option



The BOP for nuisance, as seen in Figure 4.19, explains which factors predict the amount of points awarded by a respondent to the choice option describing nuisance to be as low as possible. All influences are rather small. Respondents who rely more on citizen advise, younger respondents, male respondents and richer respondents award more points to this choice option.

Figure 4.20: Beeswarm Output Plot for compensation choice option



The BOP for compensation, as given in Figure 4.20, explains the predicting factors for points given to the compensation choice option. The compensation choice option is that citizens who have nuisance due to the renewable energy production, should be compensated. Respondents who rely on citizen advise rather than expert advise value compensation more. The same holds for male respondents, and younger respondents. These results overlap with the nuisance criterion. However, a good financial situation predicted more points to nuisance reduction, but less points to compensation.

4.4. Summary of Results

The chapter provides many insights that can be retrieved using SHAP on the NP RES dataset. The graphs need some knowledge of SHAP to interpret, and the text is several pages long. To summarize the results, a table with all the insights from this chapter is provided in Table 4.2

Insight from SHAP
Citizens who trust experts award more points to the nature choice option
Citizens who trust experts award more points to reducing grid failure risks
Citizens who trust citizen advice of expert advice award more points to local influence
Citizens who trust citizen advice of expert advice award more points to low energy costs
Citizens in a poor financial situation value low energy costs highly
Citizens in a better financial situation want more local influence
Citizens in a better financial situation award more points to reduction of grid failure risks
Citizens in a poor financial situation give less points to nature
Citizens in a poor financial situation give less points to regional compensation
Citizens in a poor financial situation give more points to local compensation for nuisance
There is no large effect of the size of a citizens municipality on the choice task
Women value low energy costs more than men
Men value high grid reliability higher than women
Women award more points to the nature choice option
Young citizens award less points to nature
Young citizens award more points to compensation for nuisance
Young citizens award more points to reducing nuisance
Young citizens award less points to energy costs
Middle-aged citizens award more points to energy costs
Education level is not of strong influence on the choice tasks
The diversity in choice for the landscape option is best predicted by gender
The main predictor for choosing for local influence is trust in experts: low trust is associated with a preference for local influence
Differences in choice for regional compensation are not well explained by the demographic variables.
Choosing for nuisance reduction is more often chosen among citizens with low trust in experts
Reducing nuisance and compensation nuisance are more often chosen by the same profile of respondents: low trust in experts, male and young.
A difference between reducing nuisance and compensating nuisance is explained by financial situation: financially healthier citizens choose more often for reducing nuisance.

Table 4.2: Insights resulting from SHAP Analysis

5 Comparison of Results from Machine Learning and Conventional Methods

Results of the descriptive statistics and LCCA on the NP RES dataset have been provided in chapter 3, and results of the SHAP have been provided in chapter 4. Therefore the results of these methods can be compared in order to answer the subresearch question:

”How do the results of machine learning compare to conventional PVE analysis methods?”

5.1. Comparison of SHAP and LCCA

It is interesting to contrast SHAP results to LCCA. LCCA provides groups of respondents that are as homogenous within the group, while being as heterogenous as possible compared to other groups. It can give insight in which large groups exist in the data sample. SHAP focusses on individual inputs, rather than a group analysis. SHAP will give information about which demographic factors are important for a respondents choices, whereas LCCA provides clusters of respondents with all factors. LCCA makes a more visible and clear image of groups in society for decision makers. SHAP provides a bulk of results for a lot of individual categories. In that sense, SHAP provides richer information, but is more difficult to comprehend. It can not give a clear picture as LCCA does. The results of SHAP and LCCA are compared on three criteria: amount of insights, type of insights and coherency of results.

First, the amount of insights can be compared. As seen in Table 4.2, there are 26 insights mentioned. The LCCA resulted in 11 insights. More is not necessarily better, but it gives an indication about the amount of information that can be gained from the analysis. Both analyses and reports have not been written by the same analyst. I am the analyst of the SHAP, whereas a Populytics analyst has done performed the LCCA. Therefore a quantitative comparison between amount of insights is not so precise. However, the difference between amount of insights is quite large. One of the explanations for SHAP resulting in more insights is the amount of significant factors. LCCA has given only age, education level and municipality size as significant. These factors are then computed for all three clusters. SHAP computes relations between all six demographic variables and for all eight outputs. Since SHAP generates much more data than an LCCA, it will likely often produce more and more detailed insights.

Secondly, the type of results can be compared. By following the framework defined in chapter 2, the categories are built. Since the amount of data was rather small (less than one page of text), steps to reduce the amount of categories were not necessary. The initial categorization resulted in three categories of insights in the LCCA, and two in the SHAP insights. In the LCCA, the category ”Significance of variable” was apparent. This was merged with the similar ”Influence of demographic variable on choice” from the SHAP results. They both described whether or not a variable was important to the choice of participants. The results of LCCA and SHAP with the type of results can be evaluated in Table 5.1 and Table 5.2.

One of the insight categories in the LCCA results is ”Cluster in data” type. These insights indicate that there is some group or cluster of participants in the data that have some homogeneity within the group. These data types can not be found in the SHAP method. Then there are insights with the type ”relative influence of demographics on choice”, which explain that some demographic characteristic is correlated with a certain choice. These type of results are seen by both analytical methods. Finally, there are insights of type ”influence of demographic variable on choice”. These insights tell whether or not a certain demographic characteristic is relevant or significant for the choice task outcomes. These types of insights emerge both in LCCA and in SHAP. The typification for LCCA and SHAP respectively can be evaluated in Table 5.1 and Table 5.2. The comparison of types can be evaluated in Table 5.3.

Insight from LCCA	**Type**
59% of participants is in a group that closely resembles the average points division	Cluster in data
In the average cluster, young participants and people from large municipalities are overrespresented	Relative influence of demographics on choice
Gender and financial health are not significant for the clustering	Influence of demographic variable on choice
A cluster exists of 21% of participants which is very nature and landscape focussed	Cluster in data
Older participants are in the nature-cluster more often	Relative influence of demographics on choice
Participants in average-sized municipalities are more often in the nature-cluster	Relative influence of demographics on choice
Average-educated citizens are more often in the nature-cluster	Relative influence of demographics on choice
A cluster exists of 20% of participants who value low costs, high reliability, low nuisance and high compensation for nuisance.	Cluster in data
In the citizen-focussed cluster, low-educated people are more prevalent	Relative influence of demographics on choice
In the citizen-focussed cluster, citizens in small municipalities are overrepresented	Relative influence of demographics on choice
In the citizen-focussed cluster, participants between 35 and 65 years old are overrepresented	Relative influence of demographics on choice

Table 5.1: All insights gathered using LCCA on the NP RES dataset

Insight from SHAP	**Type**
Citizens who trust experts award more points to the nature choice option	Relative influence of demographics on choice
Citizens who trust experts award more points to reducing grid failure risks	Relative influence of demographics on choice
Citizens who trust citizen advice of expert advice award more points to local influence	Relative influence of demographics on choice
Citizens who trust citizen advice of expert advice award more points to low energy costs	Relative influence of demographics on choice
Citizens in a poor financial situation value low energy costs highly	Relative influence of demographics on choice
Citizens in a better financial situation want more local influence	Relative influence of demographics on choice
Citizens in a better financial situation award more points to reduction of grid failure risks	Relative influence of demographics on choice
Citizens in a poor financial situation give less points to nature	Relative influence of demographics on choice
Citizens in a poor financial situation give less points to regional compensation	Relative influence of demographics on choice
Citizens in a poor financial situation give more points to local compensation for nuisance	Relative influence of demographics on choice
There is no large effect of the size of a citizens municipality on the choice task	Influence of demographic variable on choice
Women value low energy costs more than men	Relative influence of demographics on choice
Men value high grid reliability higher than women	Relative influence of demographics on choice
Women award more points to the nature choice option	Relative influence of demographics on choice
Young citizens award less points to nature	Relative influence of demographics on choice
Young citizens award more points to compensation for nuisance	Relative influence of demographics on choice
Young citizens award more points to reducing nuisance	Relative influence of demographics on choice
Young citizens award less points to energy costs	Relative influence of demographics on choice
Middle-aged citizens award more points to energy costs	Relative influence of demographics on choice
Education level is not of strong influence on the choice tasks	Influence of demographic variable on choice
The diversity in choice for the landscape option is best predicted by gender	Influence of demographic variable on choice
The main predictor for choosing for local influence is trust in experts: low trust is associated with a preference for local influence	Influence of demographic variable on choice
Differences in choice for regional compensation are not well explained by the demographic variables.	Influence of demographic variable on choice
Choosing for nuisance reduction is more often chosen among citizens with low trust in experts	Relative influence of demographics on choice
Reducing nuisance and compensation nuisance are more often chosen by the same profile of respondents: low trust in experts, male and young.	Relative influence of demographics on choice
A difference between reducing nuisance and compensating nuisance is explained by financial situation: financially healthier citizens choose more often for reducing nuisance.	Relative influence of demographics on choice

Table 5.2: All insights about the NP RES dataset retrieved using SHAP

Type	**SHAP**	**LCCA**
Cluster in data		X
Relative influence of demographics on choice	X	X
Influence of demographic variable on choice	X	X

Table 5.3: Types of insights apparent in NP RES results of SHAP and LCCA

The third way of comparing SHAP and LCCA is by looking at the insights of both analysis methods. Insights can be supported by insights from the other method, insights can be in contrast to insights from the other method, and insights can be unfound in the other method. An overview is given in Table 5.4.

Insight from SHAP	Comparison LCCA
Citizens who trust experts award more points to the nature choice option	Not tested in LCCA
Citizens who trust experts award more points to reducing grid failure risks	Not tested in LCCA
Citizens who trust citizen advice of expert advice award more points to local influence	Not tested in LCCA
Citizens who trust citizen advice of expert advice award more points to low energy costs	Not tested in LCCA
Citizens in a poor financial situation value low energy costs highly	Not significant in LCCA
Citizens in a better financial situation want more local influence	Not significant in LCCA
Citizens in a better financial situation award more points to reduction of grid failure risks	Not significant in LCCA
Citizens in a poor financial situation give less points to nature	Not significant in LCCA
Citizens in a poor financial situation give less points to regional compensation	Not significant in LCCA
Citizens in a poor financial situation give more points to local compensation for nuisance	Not significant in LCCA
There is no large effect of the size of a citizens municipality on the choice task	In contrast to LCCA
Women value low energy costs more than men	Not significant in LCCA
Men value high grid reliability higher than women	Not significant in LCCA
Women award more points to the nature choice option	Not significant in LCCA
Young citizens award less points to nature	Supports LCCA
Young citizens award more points to compensation for nuisance	In contrast to LCCA
Young citizens award more points to reducing nuisance	In contrast to LCCA
Young citizens award less points to energy costs	Supports LCCA
Middle-aged citizens award more points to energy costs	Supports LCCA
Education level is not of strong influence on the choice tasks	In contrast to LCCA
The diversity in choice for the landscape option is best predicted by gender	In contrast to LCCA
The main predictor for choosing for local influence is trust in experts: low trust is associated with a preference for local influence	Not tested in LCCA
Differences in choice for regional compensation are not well explained by the demographic variables.	Supports LCCA
Choosing for nuisance reduction is more often chosen among citizens with low trust in experts	Not tested in LCCA
Reducing nuisance and compensation nuisance are more often chosen by the same profile of respondents: low trust in experts, male and young.	Not tested in LCCA
A difference between reducing nuisance and compensating nuisance is explained by financial situation: financially healthier citizens choose more often for reducing nuisance.	Not significant in LCCA

Table 5.4: Insights from SHAP analysis on NP RES and their relation to insights in LCCA

Type	Count
Not tested in LCCA	7
Not significant in LCCA	10
Supports LCCA	4
In contrast to LCCA	5

Table 5.5: Appearance of every type of relation of SHAP insight to LCCA insight in Table 5.4

The result of counting how often a comparison type appears in the SHAP insights table is given in Table 5.5. First of all, some of the insights could not be tested, as they were not included in the LCCA design. In order to do a fair comparison of methods, in the future such included inputs should be the same to avoid comparing apples to oranges. Ten out of 26 SHAP insights are about demographic variables that were not significant in the LCCA. Both analysis methods have entirely different demographic variables that they find to be significant (LCCA) or find to be of large influence (SHAP). There seems to be a difference in relevance of a demographic variable for making clusters compared to explaining individual choice task scores. Of the other insights, four were in accordance with LCCA, and could thus have been found by only doing an LCCA. Five of the insights however, are in direct contrast to insights from LCCA. Most of these have to do with whether or not a demographic variable is of influence to the choice task outcomes.

5.2. Comparison SHAP and Choice Modelling

Something that SHAP analysis can make clear, and other methods like choice modelling have more difficulty with, is nonlinear effects. By looking at the SHAP plots, an analyst can visually spot nonlinear behaviour. An example is the effect of age on the amount of points given to energy price reduction. Younger respondents valued it least, but middle-aged respondents valued it most. The oldest respondents were somewhere in between. That is a result that linear methods would not be able to spot. Since choice modelling is not applied to the NP RES case (and can not be applied as there is no suitable model) results can not be compared.

5.3. Conclusion

The chapter started with the question how the SHAP results compare to the results of other methods. The results of SHAP are of a completely different type than the results of the descriptive statistics, and therefore it is difficult to compare them. All results that SHAP provided however, were not present in the descriptive statistics of the NP RES case and it can thus be said that using SHAP in addition to descriptive statistics generates more insights than when only descriptive statistics are used.

In comparison to LCCA, the SHAP method provided more results (26 versus 11), of which only four were also apparent in the LCCA analysis. Five of the insights resulting from SHAP are in direct contrast to results from the LCCA analysis. Ten insights were not significant in the LCCA, whereas they appeared as the strongest factors in SHAP. Seven of the insights in SHAP were not tested in the LCCA, as one SHAP input was not used as LCCA input (trust in experts).

6 Discussion

Using the SHAP method on the NP RES PVE dataset resulted in additional insights. The results are discussed and interpreted in this discussion, along with the societal and scientific relevance of the findings and limits that apply to the results. Several recommendations, both for stakeholders and future research are given.

6.1. Results Interpretation

In chapter 3, the conclusion was drawn that descriptive statistics are usually used to interpret results, with choice modelling (academic) and LCCA (Populytics) often being used as additional methods to increase understanding of the utilities and participants clusters in the data. A downside of these methods is that they require analyst assumptions for the analysis. This in turn created extra room for errors. Which methods are useful to use on PVE datasets is largely unresearched and practices emerged rather spontaneously. There is thus room for extra analysis tools in the PVE research toolkit.

SHAP proved to provide additional insights in the data in chapter 4. An important sidenote to the results is that machine learning algorithms, as well as SHAP analyses, do not give statistical significance numbers. It is therefore difficult to judge the reliability of results at times. However, there are several qualitative and quantitative tests and criteria that can be used to say something about the reliability of results.

When evaluating data in statistical methods, it is important that the sample is representative. A machine learning algorithm weighs respondents based on their demographics. A SHAP analysis gives an overview per demographic category, and does not take averages over multiple categories. However, if there is a demographic category over which an analysis should be representative, but it is not given as a parameter to the machine learning algorithm, this statement does not hold. The machine learning algorithm can only weigh over categories for which it has information.

Data in a dataset can be weighted. In such a process, more value is attached to underrepresented respondents than to overrepresented respondents (Kalton & Flores Cervantes, 2003). This process compensates for when the participant group is not representative of the population. There are signs in literature that using unweighted survey data for machine learning algorithms can result in an overestimation of the reliability of results (MacNell et al., 2023). The data used for this analysis was unweighted, therefore decreasing reliability of results. In future analyses, preferably weighted data should be used.

Sample size is important for SHAP analysis. Every data category should have enough respondents for the algorithm to make accurate predictions. When this is not the case, the SHAP values will start to show a large heterogeneity, and a very wide spread in SHAP values will be seen. This is the case for several categories in the NP RES dataset.

An example of a data category that does not have a lot of input, is the education level ”low-educated”.

It has 209 entries, accounting for 16% of the participants. In a fully representative sample, this should have been around 25% [1]. However, the problem is not necessarily the relative amount of low-educated participants, but the absolute number. The 209 entries result in quite heterogeneous predictions by the machine learning model, and thus are not enough for stable predictions. This compromises the benefits of the SHAP analysis. In this case, no relevant conclusions can be drawn on the relation between low education and the choices made in the choice task.

Another example of data categories with low amounts of inputs are the advice categories. There were five categories in which participants could place themselves, ranging from ”we should rely fully on expert advise” to ”we should rely fully on citizen advise”. The SHAP values of advice categories 1 and 2 (full reliance on citizen advise & more reliance on citizen advise compared to expert advise), as well as 4 and 5 (full reliance on expert advise & more reliance on expert advise compared to citizen advise), are roughly the same in every category. This means that in future research, for the SHAP analysis, it is better to merge those categories. This gives more respondents per category, allowing the machine learning algorithms to make better predictions with less respondents. The risk of merging two categories is that you can no longer differentiate between respondents in for instance category 1 and 2. However, if there is no difference between the answers these citizens give, merging can provide larger categories. The larger a category, the more stable the predictions of the machine learning algorithm.

Another way of analyzing the reliability of a SHAP analysis is by performing a sensitivity analysis. A sensitivity analysis is defined as ”a method to determine the robustness of an assessment by examining the extent to which results are affected by changes in methods, models, values of unmeasured variables, or assumptions.” If small things are changed, and the results of the SHAP analysis start to change dramatically, that indicates low reliability.

One of the options to perform a sensitivity analysis is by changing the machine learning algorithm. A Random Forest model was used for the described SHAP analysis. However, during the comparison of algorithms steps, the Artificial Neural Network model performed almost as good as the Random Forest model. The results of both of these models can be compared to see whether they provide the same results. This has not been performed for this study and is thus a recommendation for further research.

Another option to perform sensitivity analysis is by changing the random state of the machine learning algorithm. A different random state changes the selection of data for the training of the model, and the testing of the model. If another random state results in different results, the method is unreliable. The same can be said for large results changes if the dataset is altered slightly. By comparing the results of the self-cleaned dataset and the Populytics-cleaned dataset, this information can be partially obtained. Another method is by deleting 20 percent of your dataset and seeing whether the results are still roughly the same, or removing one input from the model. All of these analyses have not been performed, and are thus recommendations for future research.

An unexpected finding is that the results of SHAP and LCCA showed very little overlap. Even more, many insights gained using SHAP were insignificant in LCCA, or even in contrast to LCCA results. This raises questions about the validity of results for both methods. Partly these differences can be explained by their different variables. LCCA used five demographic variables in its analysis (gender, education level, age, municipal size and financial situation). SHAP used one extra variable compared to the LCCA, which is whether policies should be based off citizen advise or expert advise or anything inbetween. This category was included in the SHAP analysis as the approach was to use as many demographic variables as we had in the dataset. In hindsight to compare the results of SHAP and LCCA more accurately, it is better to use the same demographic variables in both methods. Unfortunately, by the time I realized this, there was not enough time to run this comparison. Extra research is recommended into the consistency of results of both analysis methods to explain the differences and draw conclusions on the reliability of both methods.

When comparing the results of SHAP and LCCA, the qualitative framework by Pope et al. was followed (Pope et al., 2000). One of the recommendations of the authors of the framework is that quality

[1] https://www.ocwincijfers.nl/sectoren/onderwijs-algemeen/hoogst-behaald-onderwijsniveau

of categorization might be improved by having multiple researchers perform the categorization. This has not been performed, and could enhance the robustness of the categorization. The conclusion that was drawn from the comparison was mainly that SHAP cannot provide information about apparent clusters in the data as LCCA can do. After scanning the insights of SHAP once more, none of the insights is close to something that can be categorized as clustering to me. It does not seem that having another analyst would thus change that conclusion of the research.

6.2. General Remarks about the Use of SHAP

SHAP provides information on the effect of an input parameter on the output choice task. Those effects are relative effects, and not absolute. A higher (positive) SHAP value means that because a participant is in a certain demographic group, the machine learning model predicts that the participant gives more points to a certain choice option. The fact that middle-aged respondents have higher SHAP values for the energy price choice option, does however not necessarily mean that they have awarded it more absolute points than younger or older respondents. It just means that the fact that they are in that age category, has contributed positively to the prediction of their amount of points. If the age category was unspecified for the participant, the prediction of points would be lower. This age group can still theoretically consist of people who on average gave less points to the energy price option than other respondents. That would be due to other inputs that are correlated with age. It is a strength of SHAP that it isolates the effects of several correlated inputs, but at the same time it makes the results prone to misinterpretation. Handling SHAP analysis results requires specific knowledge on the way SHAP works.

It is interesting to see how many of the results that SHAP produced could have been obtained by simply doing a scatterplot analysis with coloured categories or correlation analysis. This can reveal more about the added value of SHAP.

6.3. Societal and Scientific Relevance

Several results from the results plots for the case study provide valuable information for policy makers involved in NP RES. The strongest effects seen in the data are that people in worse financial situations value low energy prices more. To predict the amount of points given to the nature choice option, the best predictor is whether a respondent values citizen or expert advise higher. Respondents who trust on expert advise give more points to the nature choice option. Those two effects are strong and reliable.

Other findings were that municipality size is hardly of influence for any of the choice tasks. Gender is an important predictor for respondents choice, along with financial situation and whether a respondent values expert or citizens advise higher.

These additional insights generated by SHAP provide information that was not gathered using other analysis methods. This means that SHAP can be a relevant tool for PVE analysis. PVE experiments are performed for mainly public institutions to elicitate preferences of citizens. If these experiments can be better analyzed by using SHAP, the result can lead to more robust and accepted decision-making.

One application of SHAP is not researched in this thesis: using SHAP to provide predictions of how a demographic group would divide points. With the results of the machine learning model and SHAP, it is possible to give a prediction for what for example a 40-year old woman with high trust in government and a high-education is likely to choose. This provides a lot of flexibility for decision-makers to ask: what about this group? How would this group likely respond in the PVE? That way a decision-maker can address the heterogeneity of opinions in the participants better than with current methods. This application is not tested due to time-constraints in my thesis, and it is thus a recommendation for further research.

Scientifically, PVE experiments are relatively new. As SHAP has resulted in additional insights and lessons learned from application in this thesis, a first step is taken into using SHAP for PVE experiments more. This enhances the options of analysis for these experiments, opening possibilities for academic PVE's to be analyzed with SHAP. SHAP gives predominantly insights in how different subgroups of participants value the options in the choice task. It opens the door for more refined academic results in which the heterogeneity in participants and demographic groups is better addressed. This leads to a broader range of results, doing justice to the diversity apparent in the society.

6.4. Boundaries and Limits

The SHAP analysis for this thesis has been performed on a PVE dataset of Populytics for a project for NP RES (National Programme Regional Energy Strategy). Results are dependant on this specific case study, and it is interesting to look into other datasets in the future.

The aim of the thesis was to investigate whether SHAP can bring additional insights in PVE analysis. Less focus was placed on the insights themselves and their reliability. There are many design choices to be made, such as which machine learning algorithm to use, which variables to include and how to clean the dataset. As the applicance of SHAP to PVE is new, there are no standards and comparison is difficult. Before all insights can be used with confidence, more research about the stability, sensitivity and reliability of the SHAP applicance on PVE data is required.

The drawing of insights from LCCA and SHAP is a manual and human process. This means the analyst is an important source of diversity and errors in results. Both methods did not have the same analyst in this thesis, and no cross-validation was performed on the insight generating process. This implies that it might be possible that results could be different when a different analyst would have performed one of the analyses.

6.5. Recommendations

In the thesis SHAP is applied to the NP RES case study. Decision-makers of NP RES can use the insights gathered by this thesis to further enhance their policies and better grasp the heterogeneity in participants opinion.

For analysts of companies performing PVE experiments, the results can show what insights SHAP can give in PVE analysis. They can consider using SHAP when there is a specific interest in revealing diversity of opinions in a PVE experiment. Currently, the LCCA is used for that reason, but SHAP reveals many insights that were not revealed by the LCCA.

There are several recommendations for future research that can be drawn from the thesis:

- A sensitivity analysis can be performed by switching the algorithm from Random Forest to Artificial Neural Network. These two performed almost as good in terms of model fit. Comparing whether they give the same results can give valuable information about the reliability of using SHAP.
- A sensitivity analysis can be performed by switching the dataset. There are many options to do this, by changing the random seed, making manual subdatasets or using a dataset with different cleaning assumptions. This can reveal how sensitive the results are to small changes in the dataset, indicating reliability.
- The added value of using SHAP depends on whether there are methods that can achieve many of the same results the SHAP analysis delivers. SHAP gives different results compared to LCCA and descriptive statistics on the NP RES case study. However, there are many more methods that could possibly be used on PVE that might have the same results with less effort. An example is category-coloured scatterplots, or using descriptive statistics for subgroup analysis.

A factor that makes the results of the thesis difficult to work with, is the counterintuitive way in which SHAP works. The results produced need explanation about the method, and are not intuitive to decision-makers and even analysts. An important concept for the interpretation of results is that SHAP provides an explanation for the machine learning predictions. SHAP does not analyse the respondent data itself. SHAP thus gives insight in what the machine learning makes predictions on, and not on any direct effects observed in the dataset. It does not give an explanation for variability in the respondent answers, but for variability in the machine predictions. If the machine learning algorithm is a bad predictor, the SHAP analysis will not give useful results. The SHAP results thus depend strongly on the machine learning model. In future projects, mainly for analysts working with decision-makers, it is key to clearly explain the workings of SHAP to get to a correct interpretation.

7 Conclusions

Throughout the thesis, the aim was to investigate whether SHAP analysis is beneficial to PVE analysis, and whether it gains additional insights. During application of SHAP to the NP RES case study, many conclusions could be drawn regarding the added value of SHAP.

There were three subquestions that together would provide the information required for a conclusion on the main research question. These subquestions are:

1. How is quantitative PVE data currently analyzed?
2. Which results does machine learning with SHAP provide on a PVE case study?
3. How do the results of machine learning compare to conventional PVE analysis methods?

In chapter 3 the current practices of PVE analysis were explained. Descriptive statistics are the standard method, and are used in virtually every PVE experiment. Choice modelling is the golden standard in academic PVE research. It is grounded in economic theory and calculates the utility functions for the choice task. However, it is not always possible to apply choice modelling to PVE, since it is rather inflexible to changes in the PVE experiment. Furthermore it requires a trained professional to make assumptions to instate the model. Latent Class Cluster Analysis (LCCA) is sometimes used in academic research, and almost always used in Populytics reports. The clustering of respondents gives decision makers a vivid picture of the subgroups available in the data. However, the results it produces are aggregate and the amount of clusters has to be predefined by the analyst. There is thus room for improvement, in which SHAP can play a role.

The comparison of SHAP results with LCCA results, showed that SHAP was able to gather more insights than LCCA. The level of detail of SHAP insights is greater than those of LCCA, explaining the difference in amount of results. The full results are provided in chapter 4. A worrying result is that the results from LCCA and SHAP in many cases contrast each other. Not only are some insights of SHAP contrary to insights of LCCA, both methods have a radically different evaluation of significance or relevance of demographic variables. This raises questions, which are further adressed in chapter 6. SHAP as used in this thesis is not capable of retrieving insight types that are not possible with LCCA.

The thesis aims to answer the following central research question:

”What additional insights does machine learning with SHAP provide for PVE quantitative analysis compared to conventional methods?”

The answer is that SHAP can gather detailed results about the individual effects of demographic variables on the choice task. Compared to conventional PVE analysis methods, it is flexible, requires less analyst assumptions and delivers a wide variety of results. It gathered more insights than the LCCA.

The largest question raised by the thesis is how consistent and robust the findings of SHAP are. Further research on the stability of the SHAP results can grant more insight into the usefulness of SHAP as data analysis method for PVE. The main suggestion is to perform sensitivity analyses to see how large outputs differ when the inputs are changed. If the results turn out to be very volatile, they are less reliable. Furthermore the amount of participants can be varied to see when SHAP results start to converge. This can lead to a recommendation on how many participants are the minimum for succesful SHAP applicance.

The thesis points out that SHAP is a promising method for explaining heterogeneity in the citizens responses to PVE. Using machine learning and artificial intelligence can extract crucial insights from data that would otherwise be missed (Subasi, 2020). SHAP can thus be a vital tool in the toolkit of the future PVE analyst.

A

Appendix A: SHAP full results

This appendix provides all graphs resulting from the SHAP analysis.

A.1. Beeswarm Input Plots

Figure A.1: Beeswarm Input Plot for advice parameter

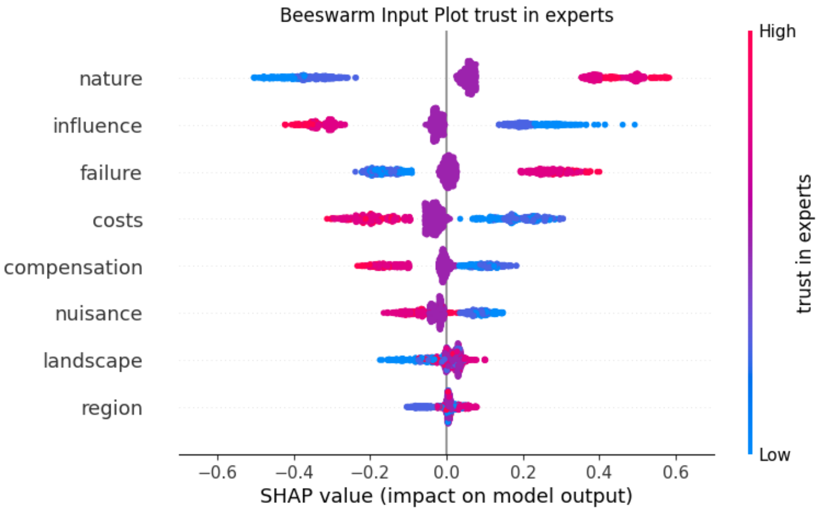


Figure A.2: Beeswarm Input Plot for financial parameter

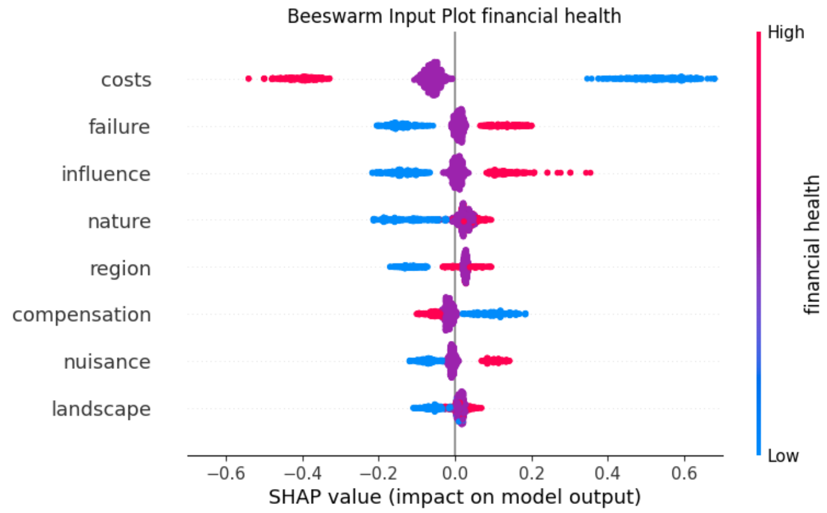


Figure A.3: Beeswarm Input Plot for municipality parameter

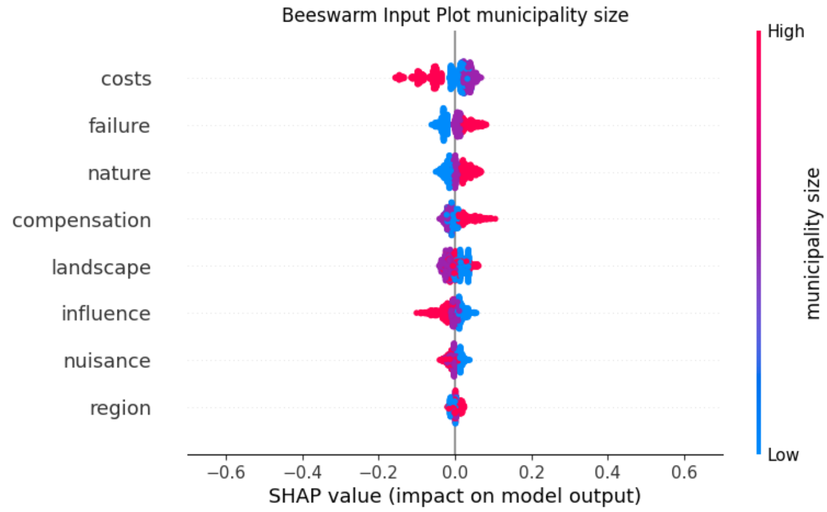


Figure A.4: Beeswarm Input Plot for gender parameter

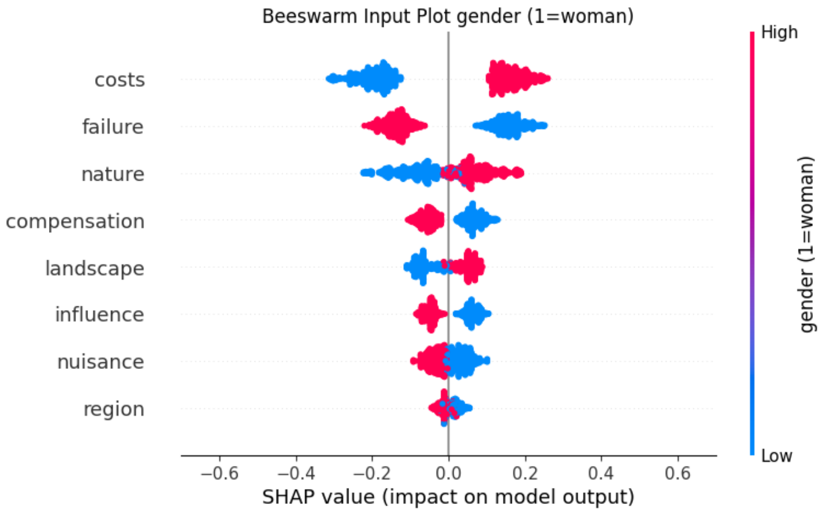


Figure A.5: Beeswarm Input Plot for age parameter

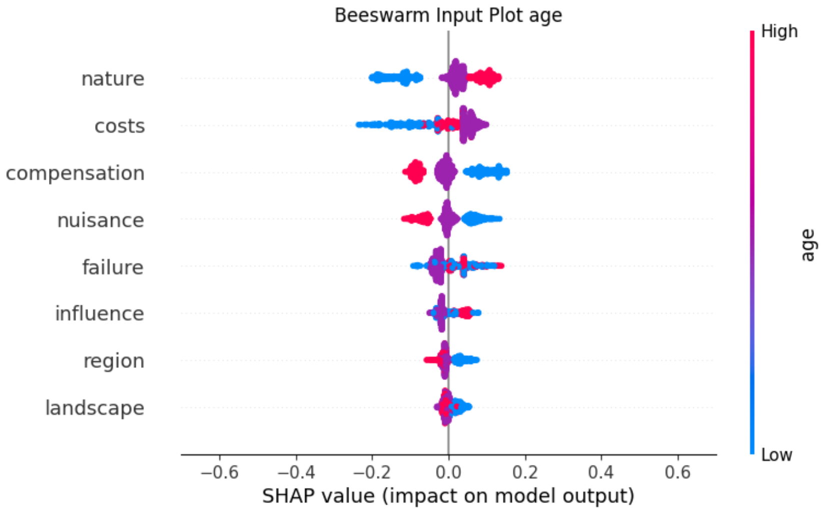
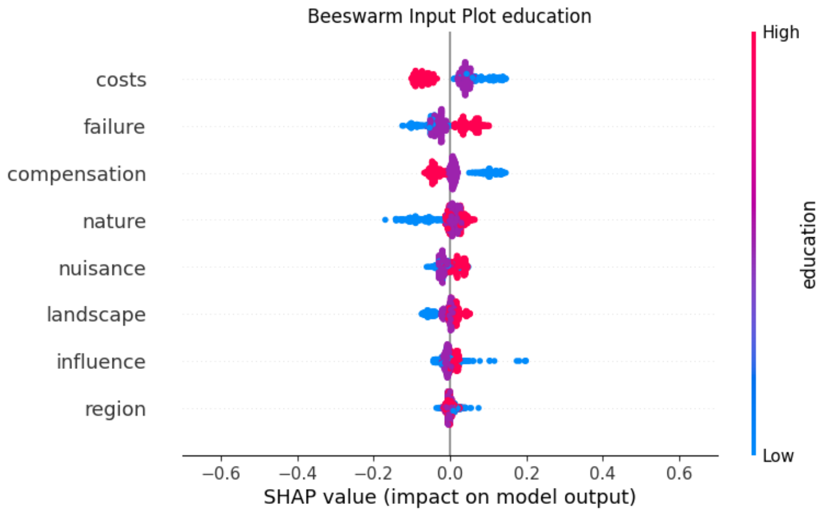


Figure A.6: Beeswarm Input Plot for education parameter



A.2. Beeswarm Output Plots

Figure A.7: Beeswarm Output Plot for landscape choice option

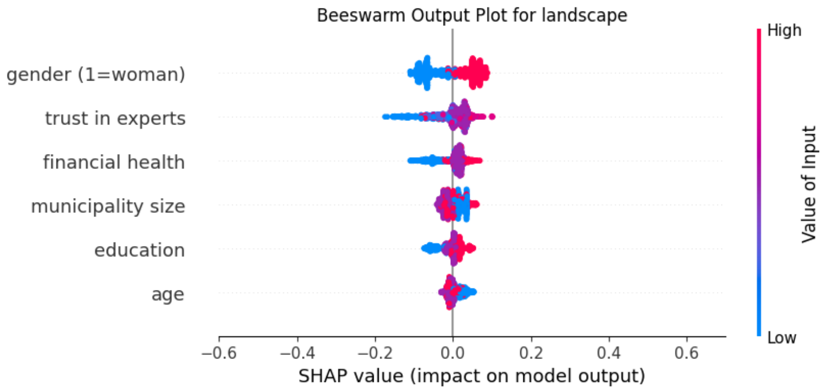


Figure A.8: Beeswarm Output Plot for influence choice option

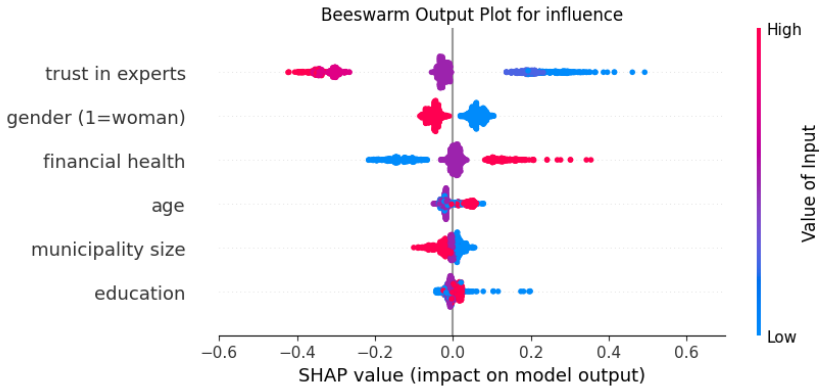


Figure A.9: Beeswarm Output Plot for failure choice option

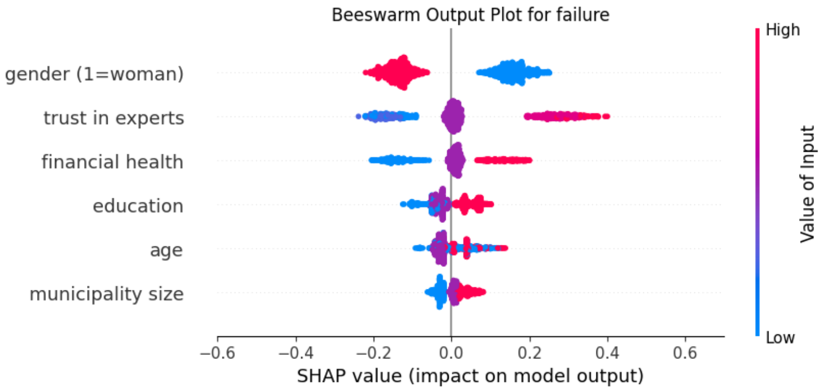


Figure A.10: Beeswarm Output Plot for costs choice option

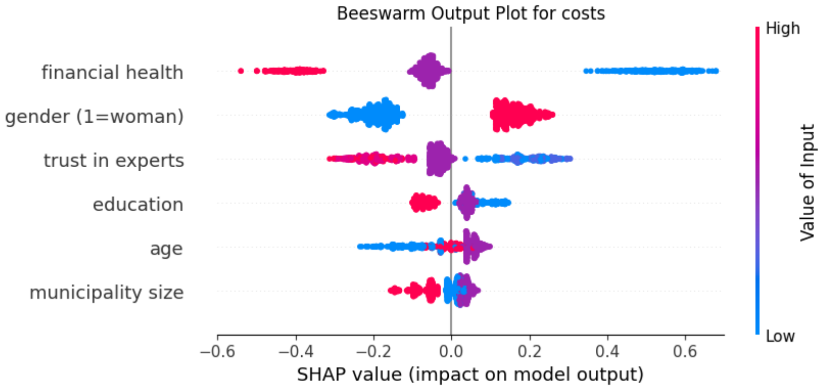


Figure A.11: Beeswarm Output Plot for region choice option

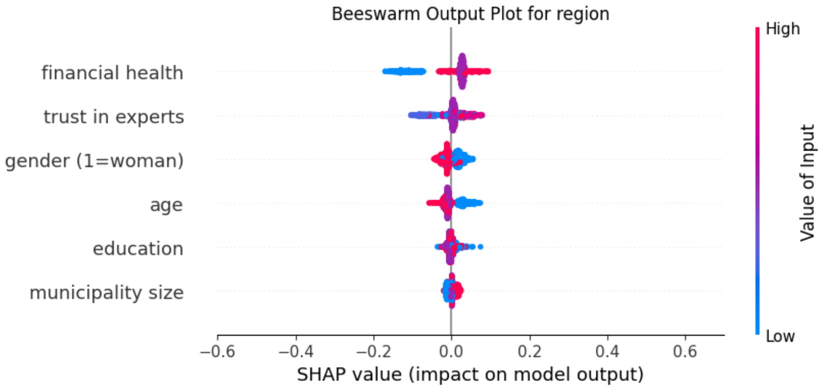


Figure A.12: Beeswarm Output Plot for nature choice option

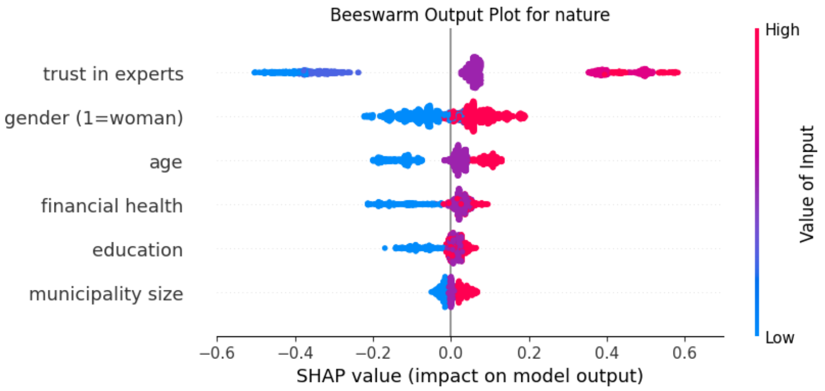


Figure A.13: Beeswarm Output Plot for nuisance choice option

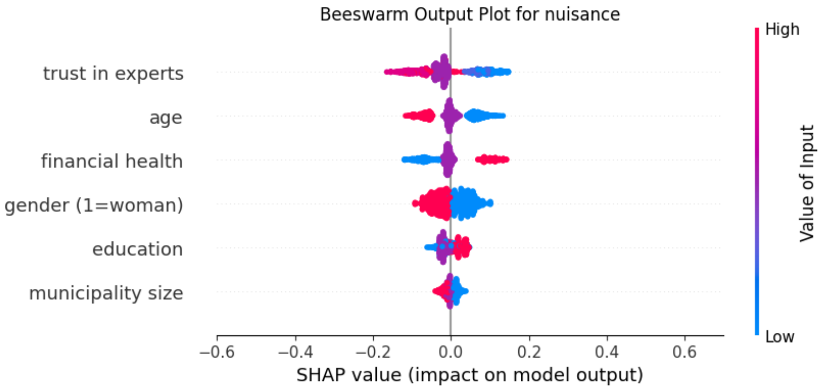
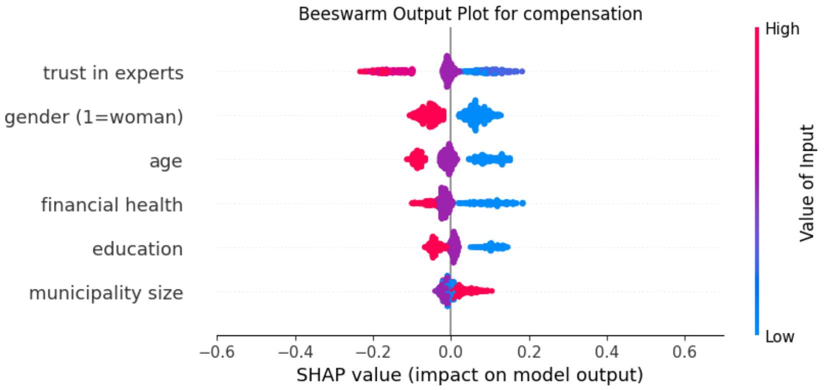


Figure A.14: Beeswarm Output Plot for compensation choice option



A.3. SHAP Scatterplots

Figure A.15: SHAP Scatterplot for advice parameter and landscape choice option

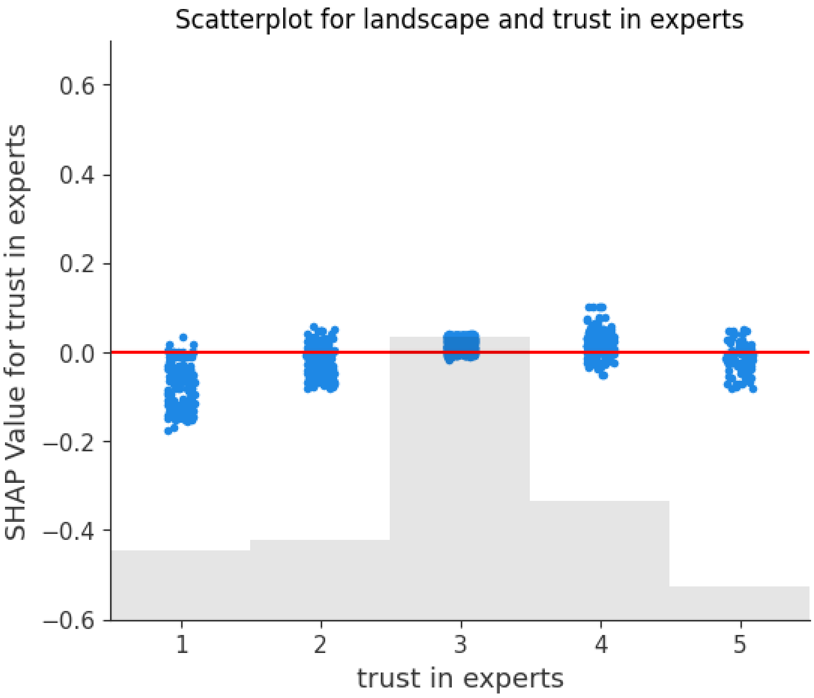


Figure A.16: SHAP Scatterplot for advice parameter and influence choice option

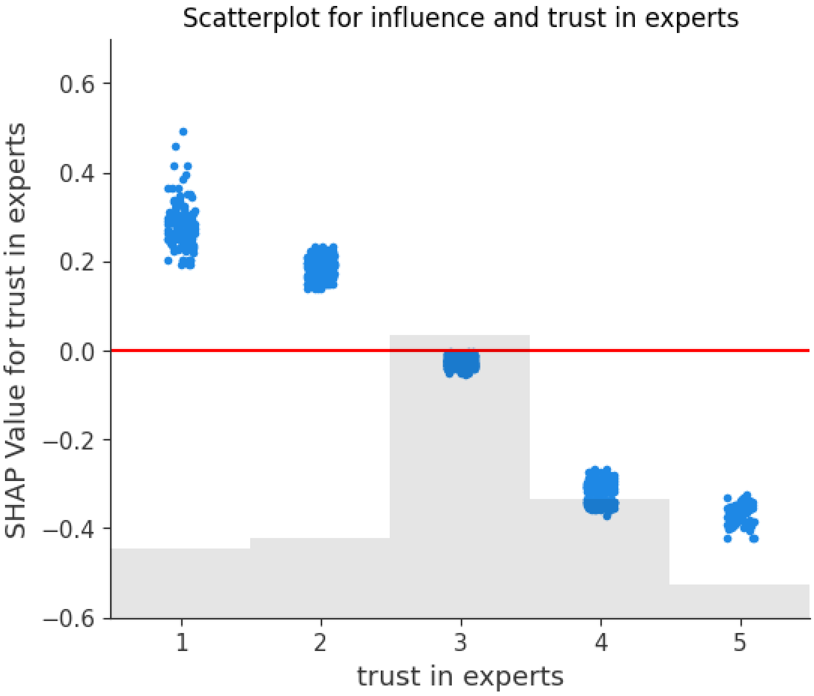


Figure A.17: SHAP Scatterplot for advice parameter and failure choice option

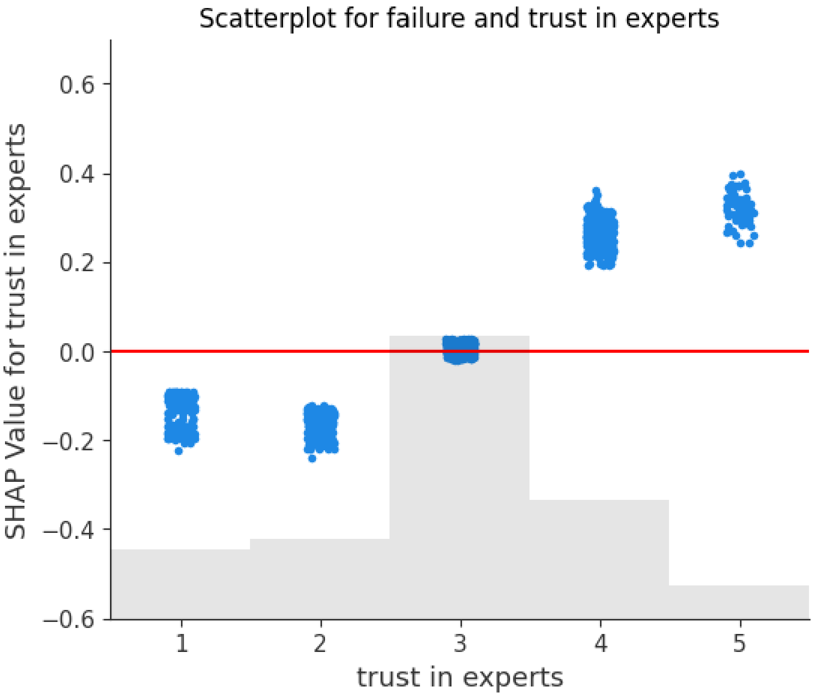


Figure A.18: SHAP Scatterplot for advice parameter and costs choice option

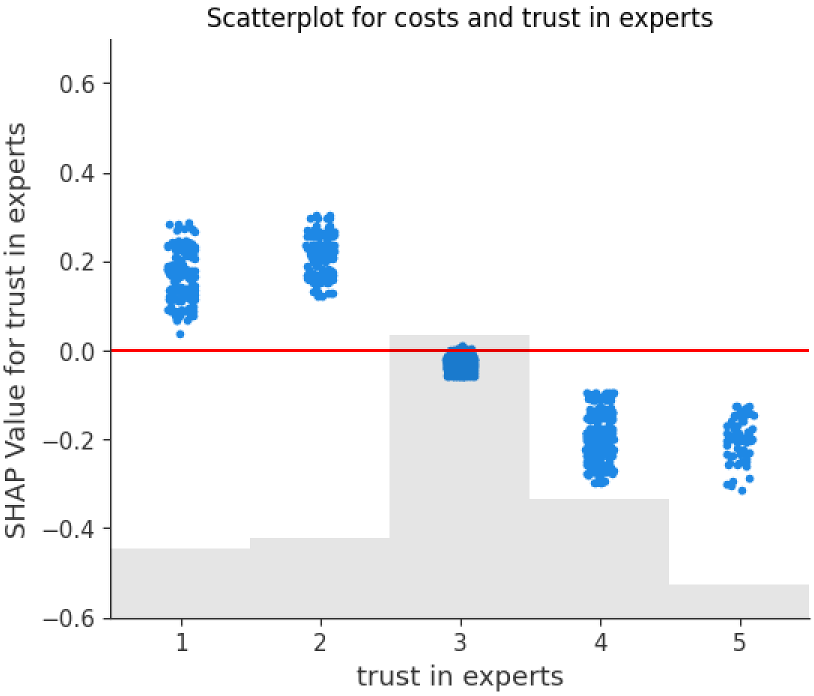


Figure A.19: SHAP Scatterplot for advice parameter and region choice option

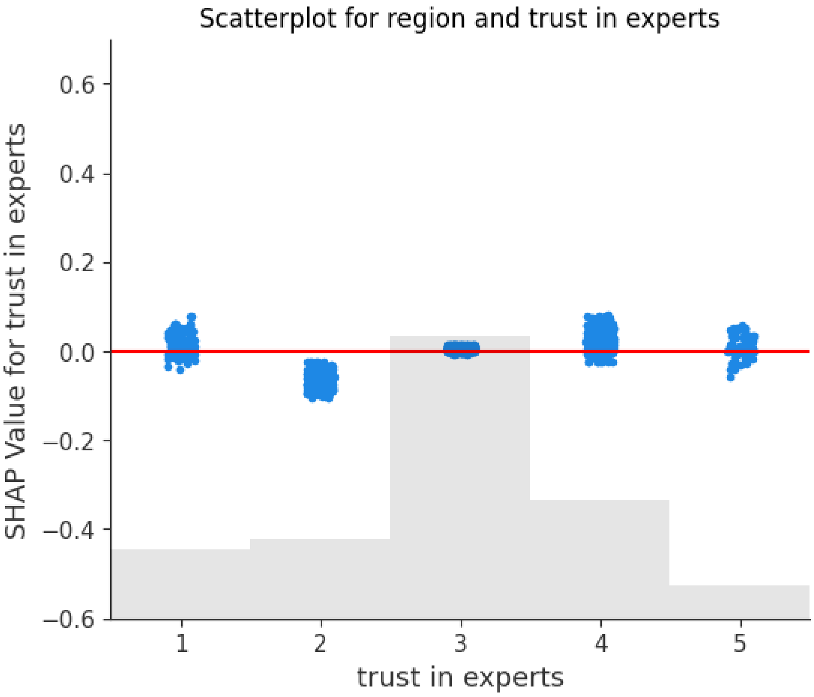


Figure A.20: SHAP Scatterplot for advice parameter and nature choice option

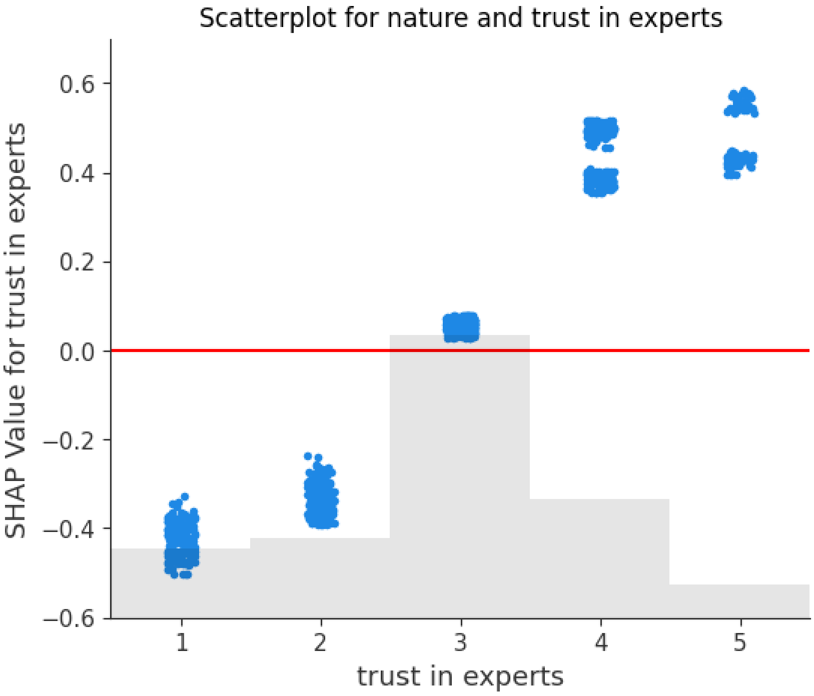


Figure A.21: SHAP Scatterplot for advice parameter and nuisance choice option

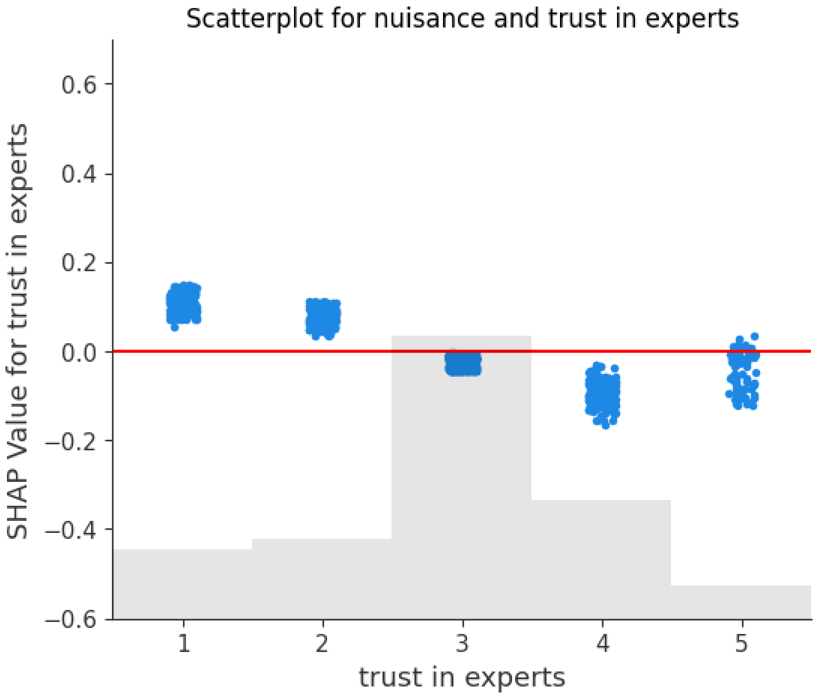


Figure A.22: SHAP Scatterplot for advice parameter and compensation choice option

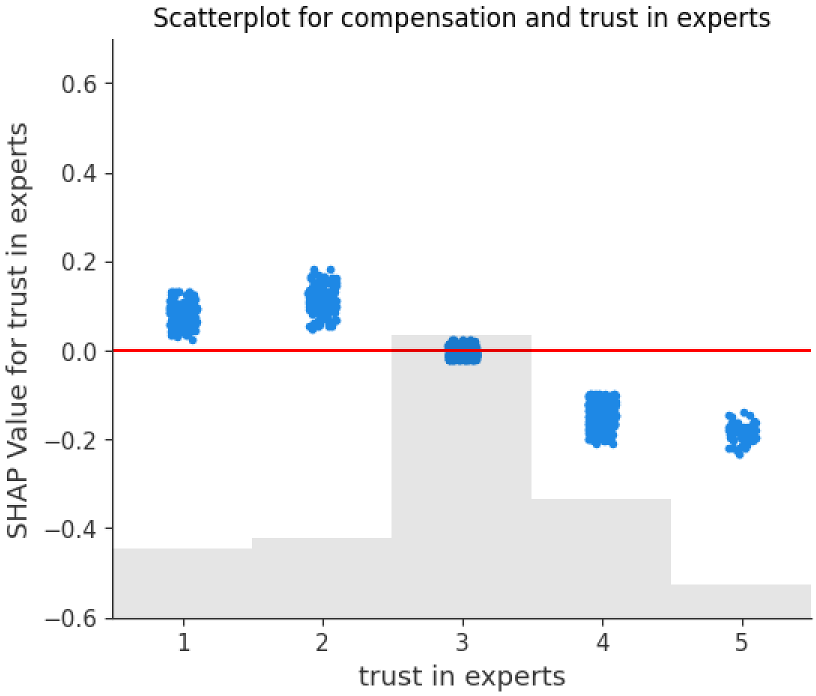


Figure A.23: SHAP Scatterplot for financial parameter and landscape choice option

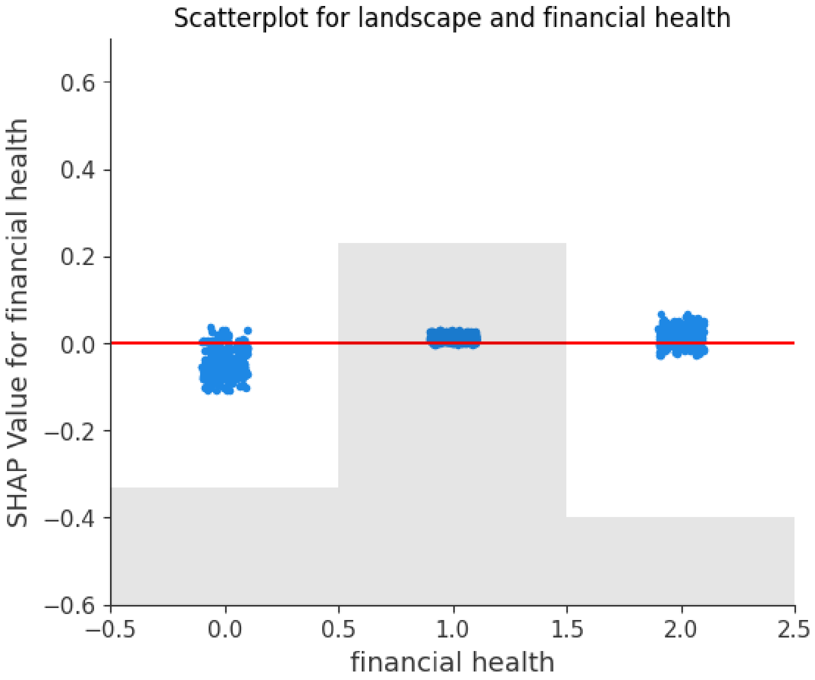


Figure A.24: SHAP Scatterplot for financial parameter and influence choice option

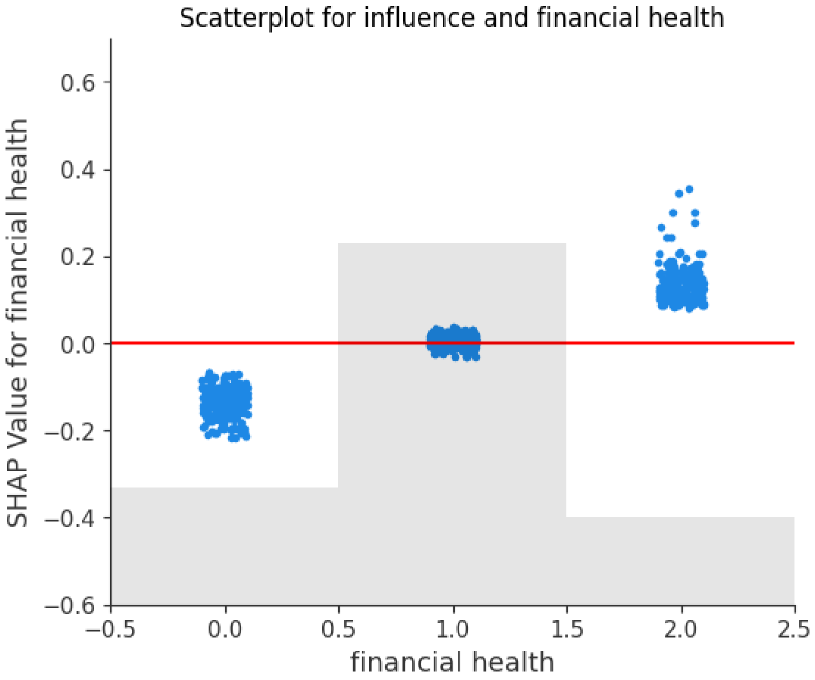


Figure A.25: SHAP Scatterplot for financial parameter and failure choice option

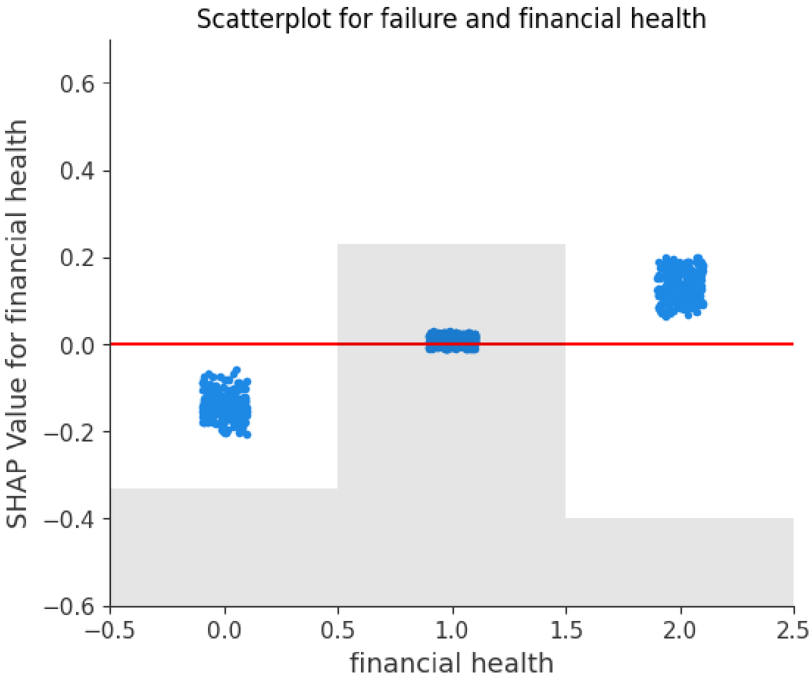


Figure A.26: SHAP Scatterplot for financial parameter and costs choice option

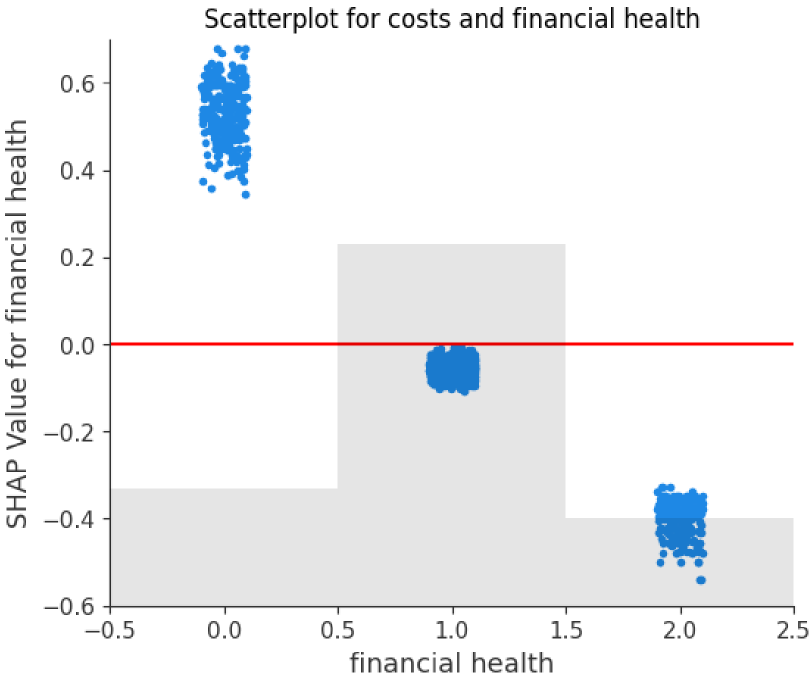


Figure A.27: SHAP Scatterplot for financial parameter and region choice option

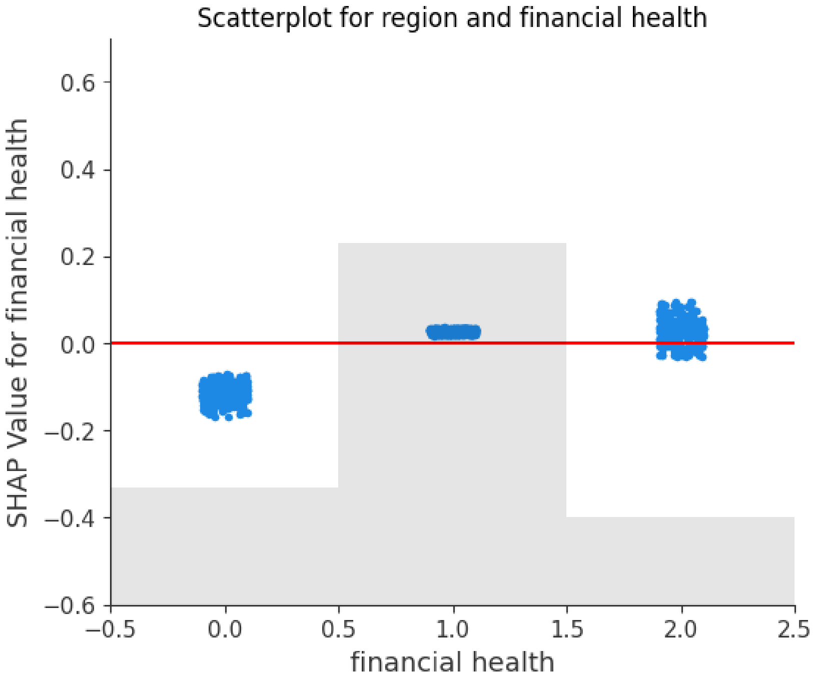


Figure A.28: SHAP Scatterplot for financial parameter and nature choice option

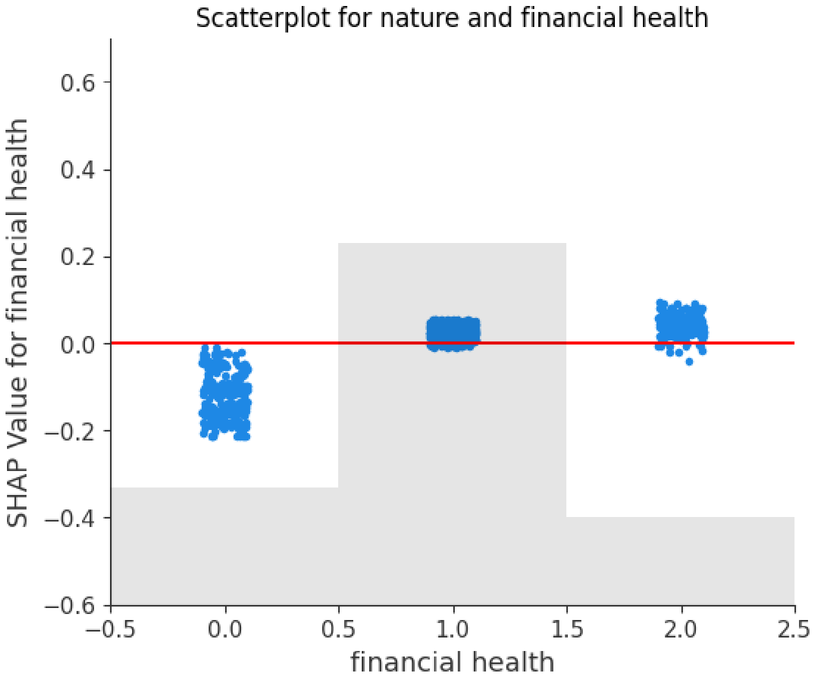


Figure A.29: SHAP Scatterplot for financial parameter and nuisance choice option

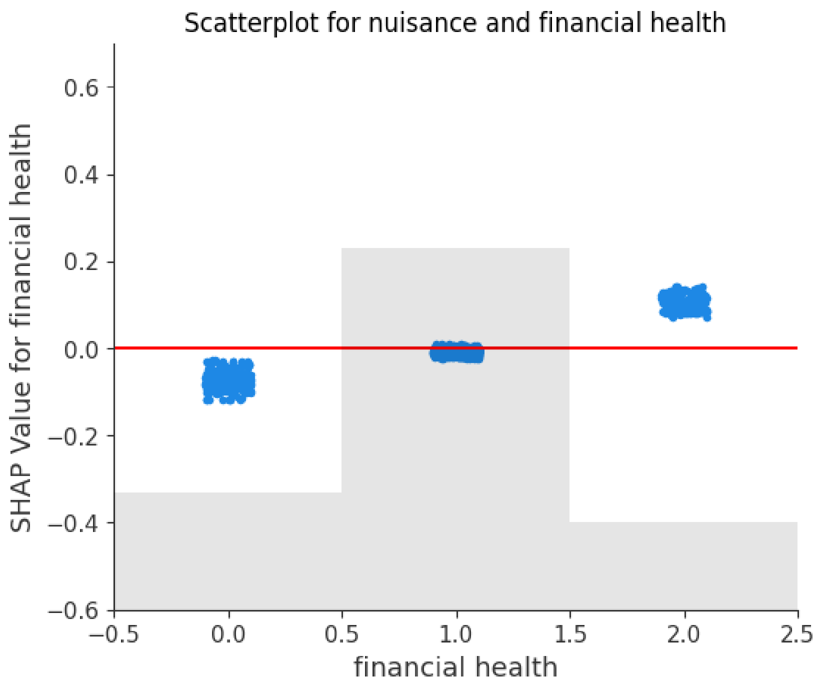


Figure A.30: SHAP Scatterplot for financial parameter and compensation choice option

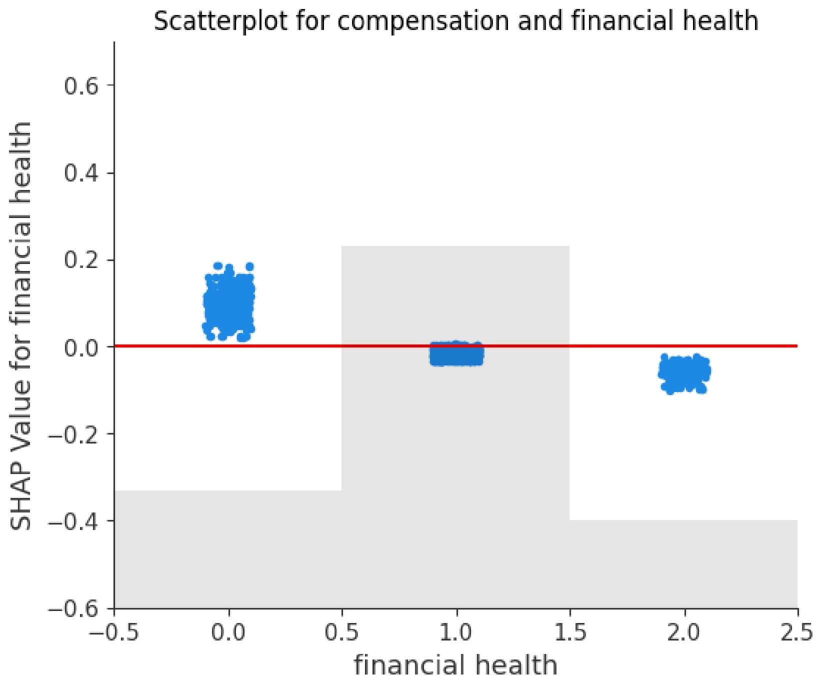


Figure A.31: SHAP Scatterplot for municipality parameter and landscape choice option

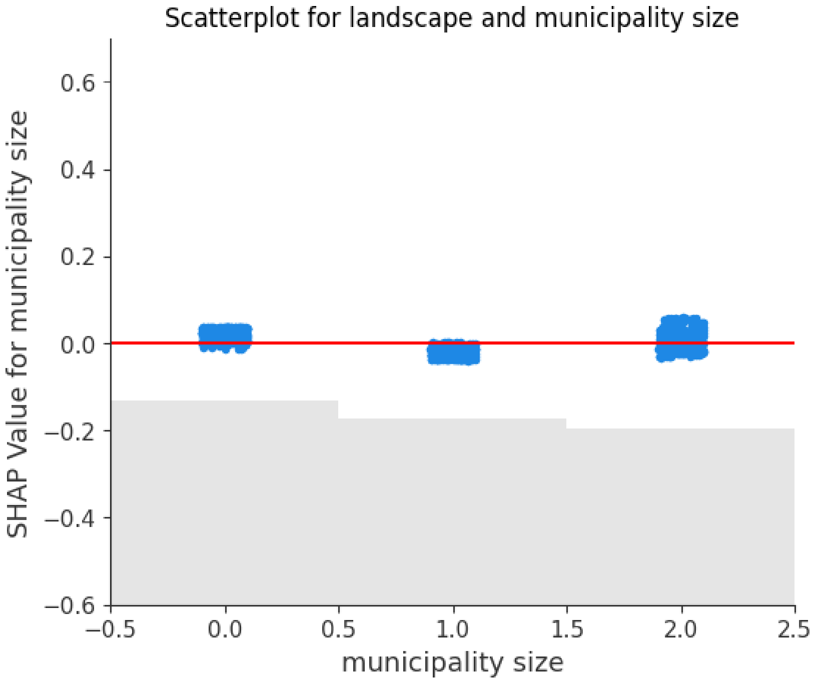


Figure A.32: SHAP Scatterplot for municipality parameter and influence choice option

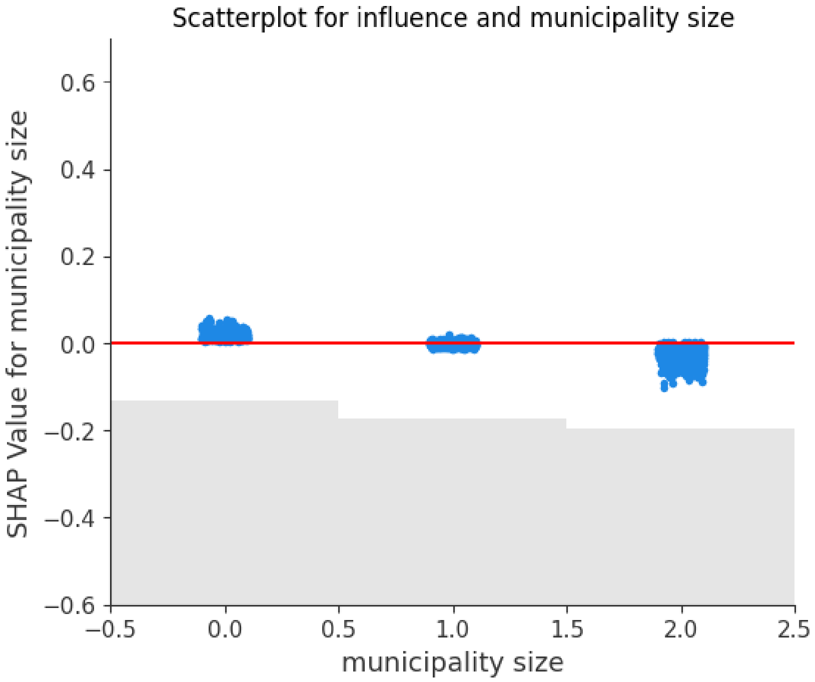


Figure A.33: SHAP Scatterplot for municipality parameter and failure choice option

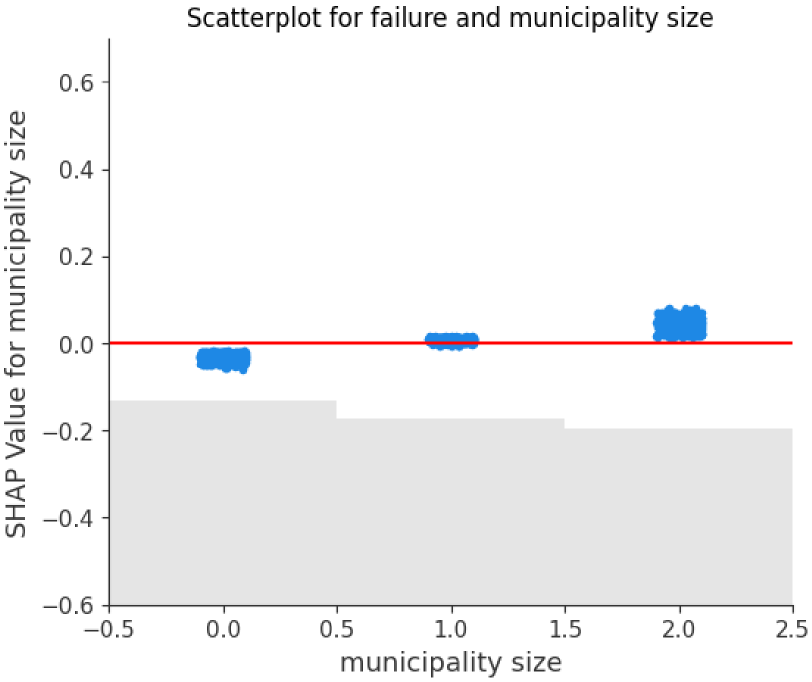


Figure A.34: SHAP Scatterplot for municipality parameter and costs choice option

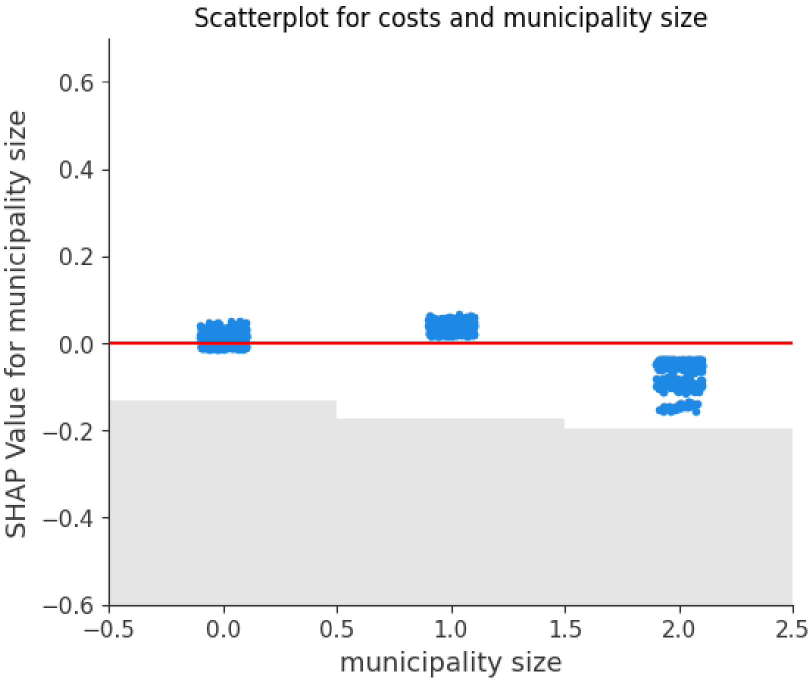


Figure A.35: SHAP Scatterplot for municipality parameter and region choice option

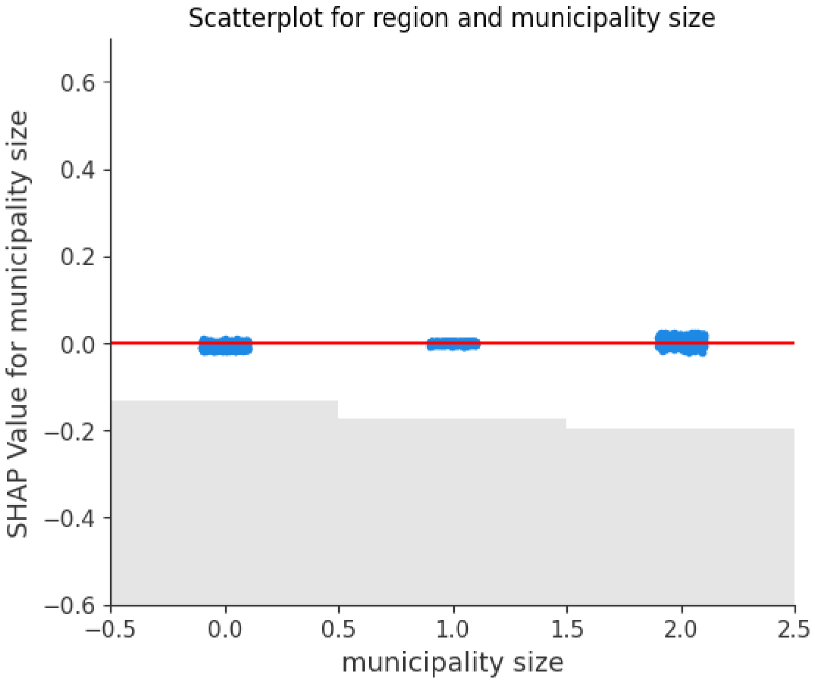


Figure A.36: SHAP Scatterplot for municipality parameter and nature choice option

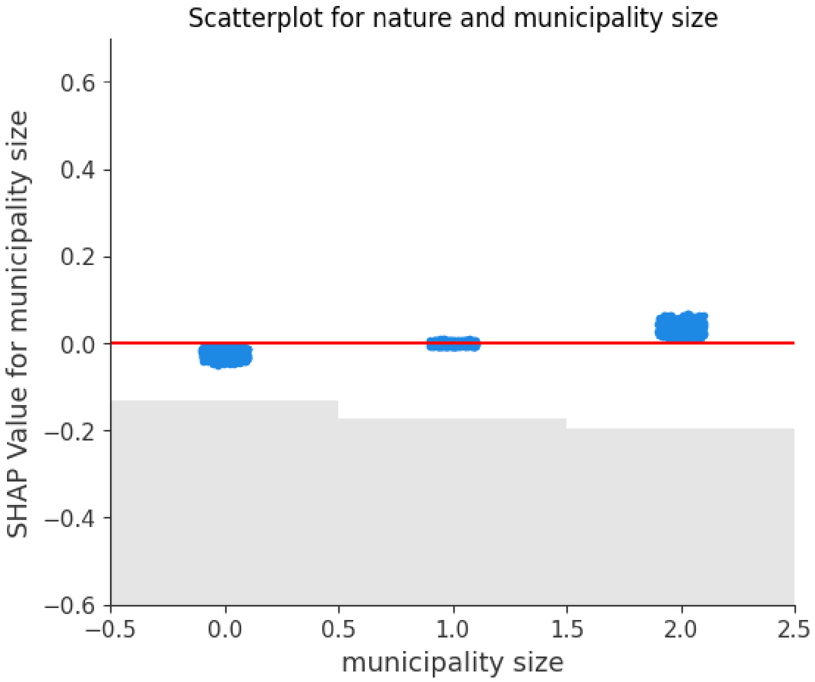


Figure A.37: SHAP Scatterplot for municipality parameter and nuisance choice option

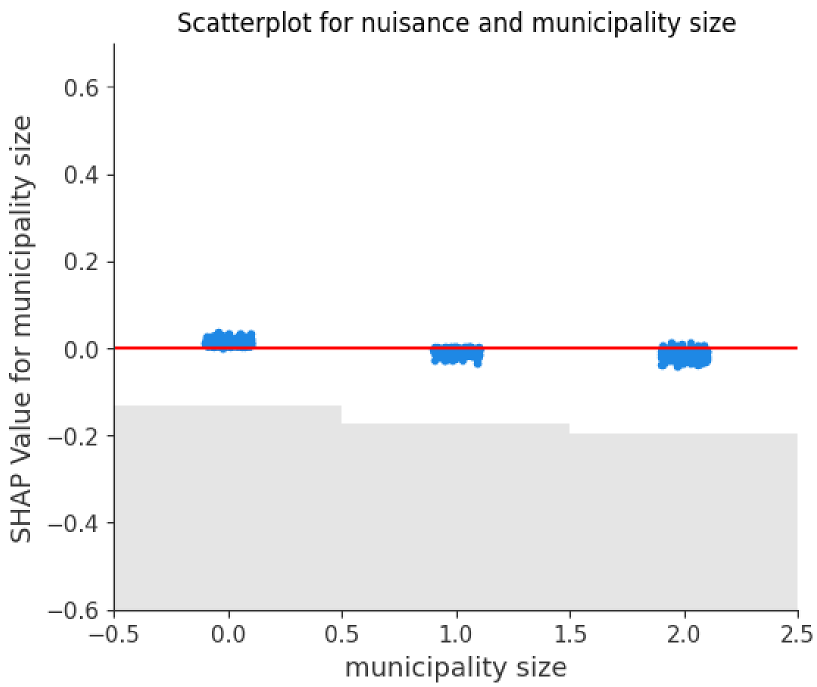


Figure A.38: SHAP Scatterplot for municipality parameter and compensation choice option

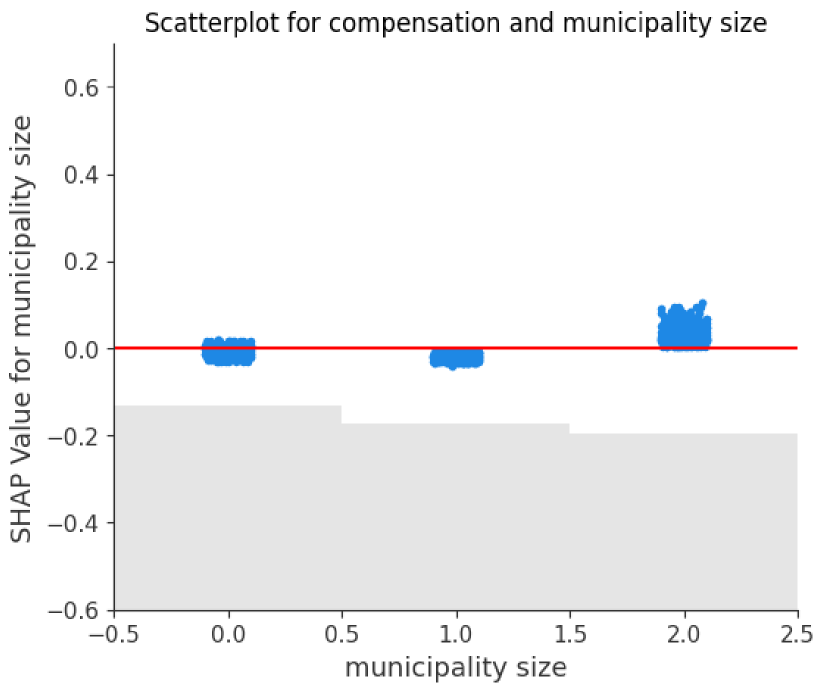


Figure A.39: SHAP Scatterplot for education parameter and landscape choice option

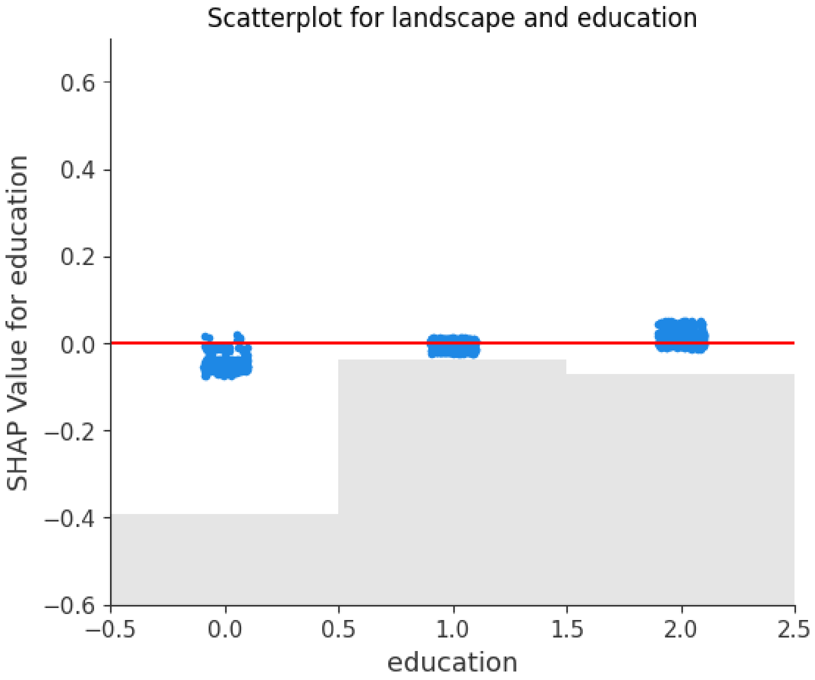


Figure A.40: SHAP Scatterplot for education parameter and influence choice option

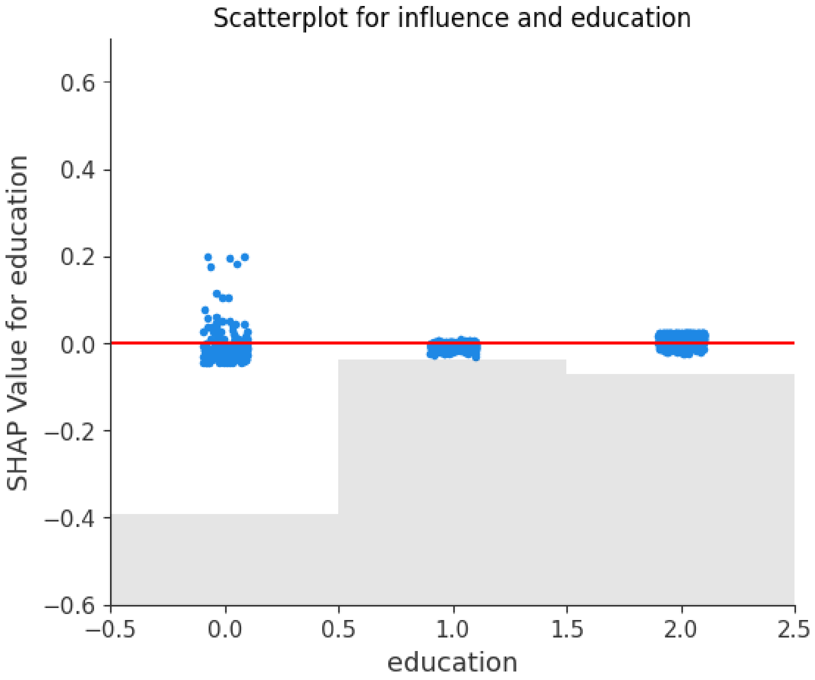


Figure A.41: SHAP Scatterplot for education parameter and failure choice option

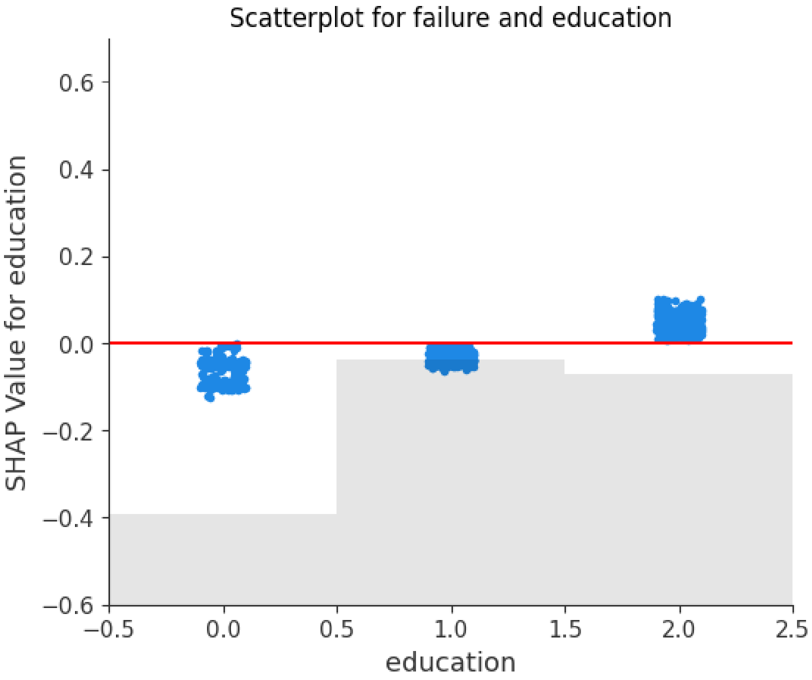


Figure A.42: SHAP Scatterplot for education parameter and costs choice option

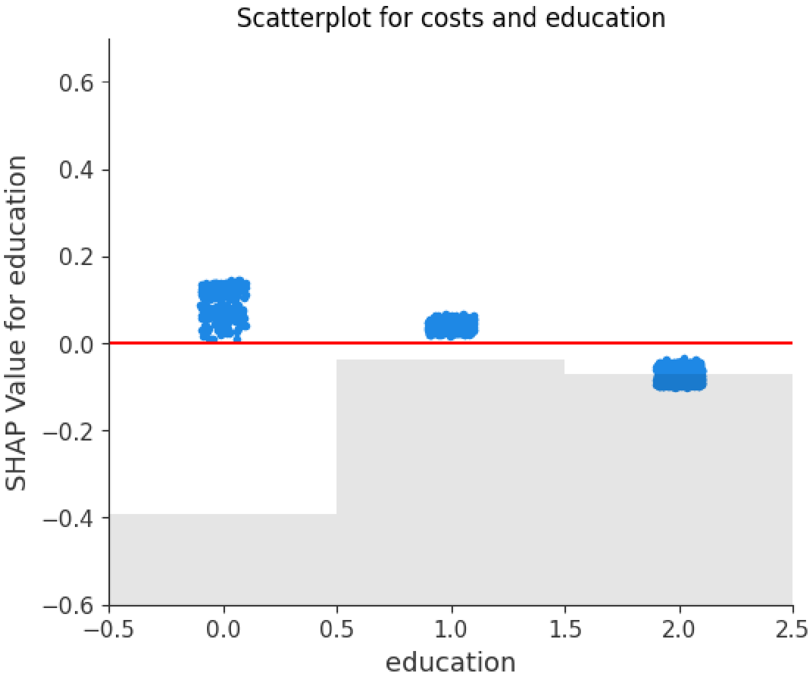


Figure A.43: SHAP Scatterplot for education parameter and region choice option

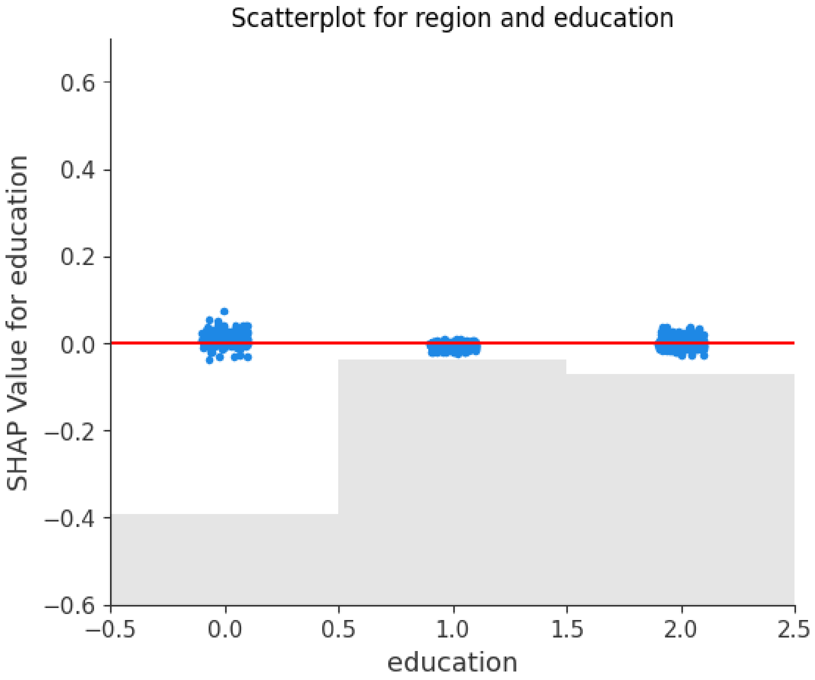


Figure A.44: SHAP Scatterplot for education parameter and nature choice option

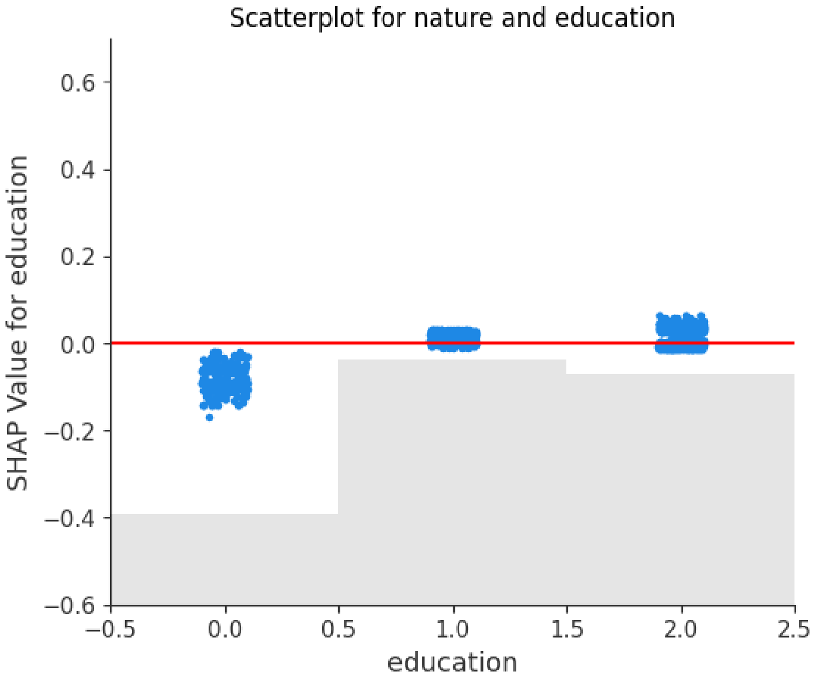


Figure A.45: SHAP Scatterplot for education parameter and nuisance choice option

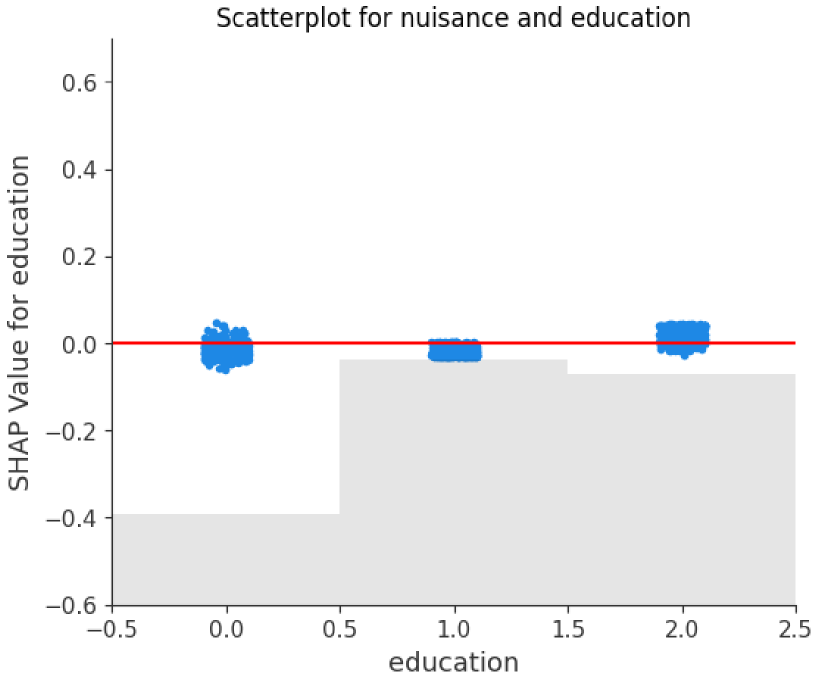


Figure A.46: SHAP Scatterplot for education parameter and compensation choice option

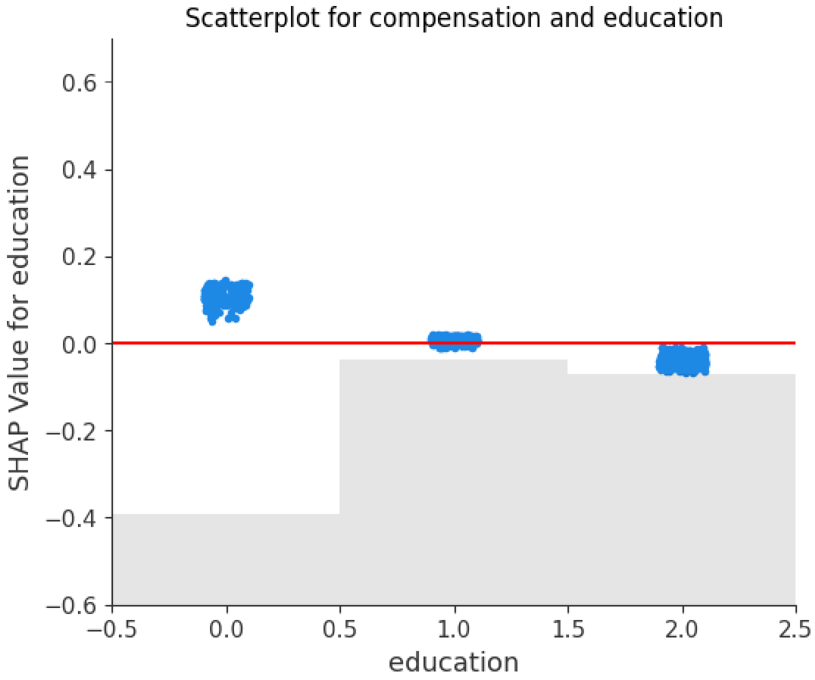


Figure A.47: SHAP Scatterplot for age parameter and landscape choice option

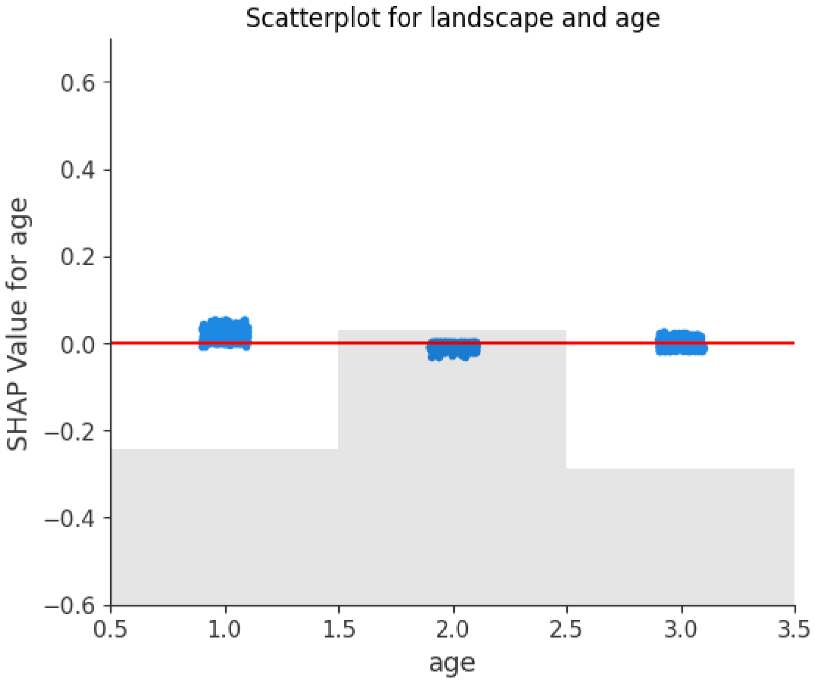


Figure A.48: SHAP Scatterplot for age parameter and influence choice option

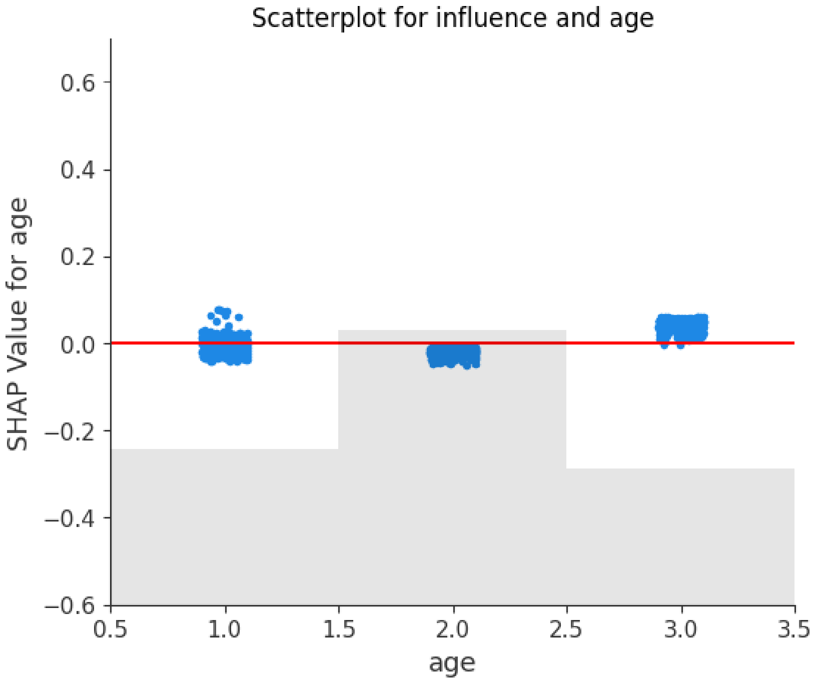


Figure A.49: SHAP Scatterplot for age parameter and failure choice option

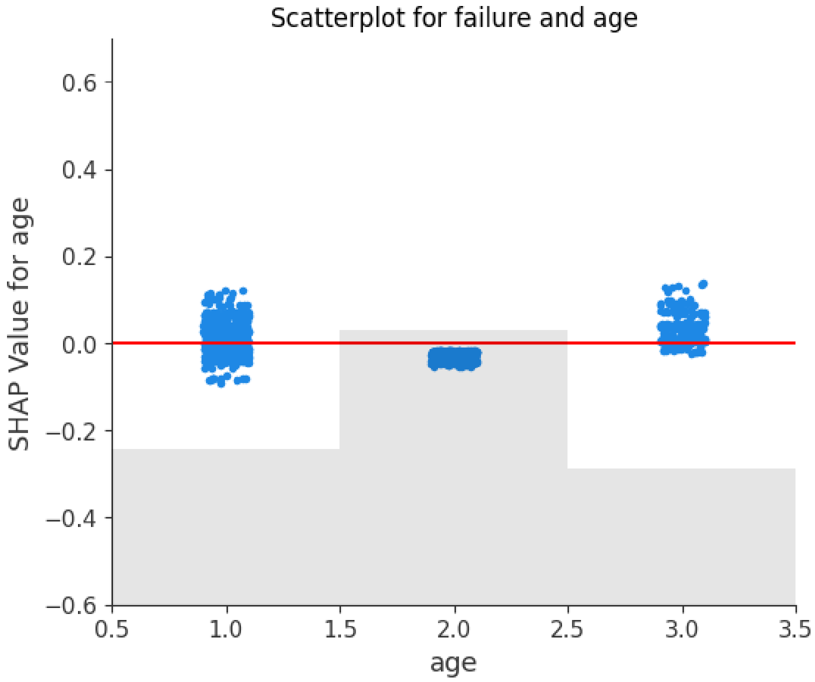


Figure A.50: SHAP Scatterplot for age parameter and costs choice option

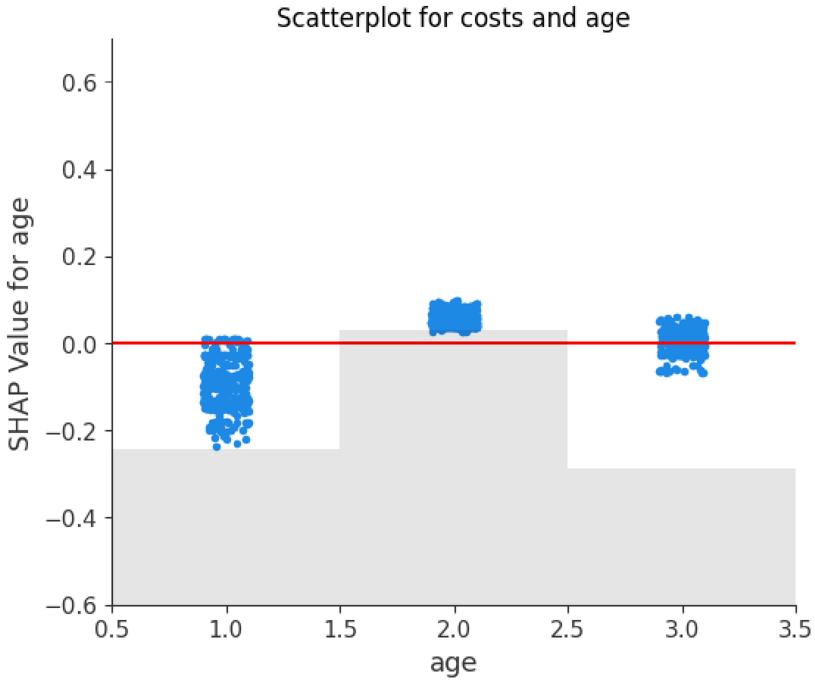


Figure A.51: SHAP Scatterplot for age parameter and region choice option

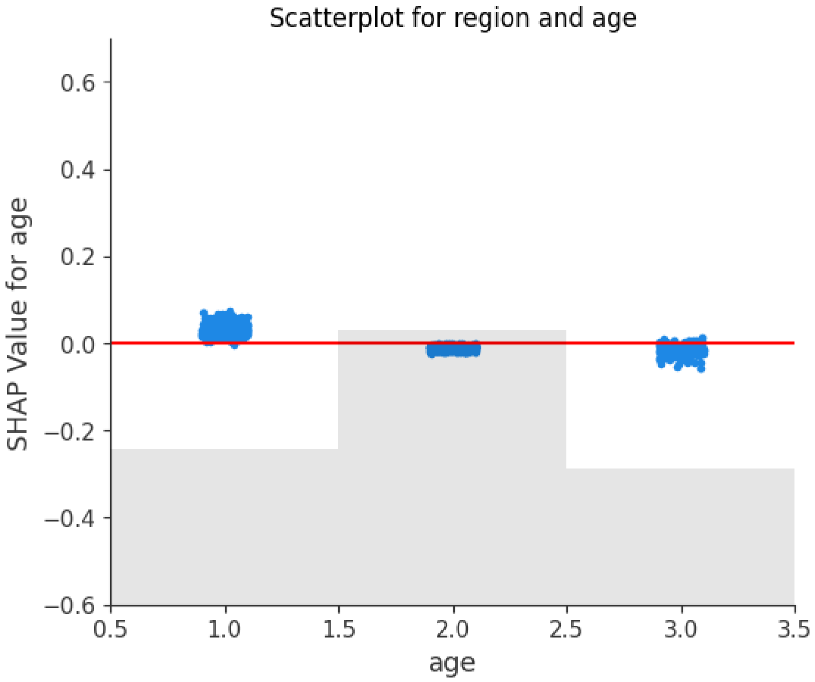


Figure A.52: SHAP Scatterplot for age parameter and nature choice option

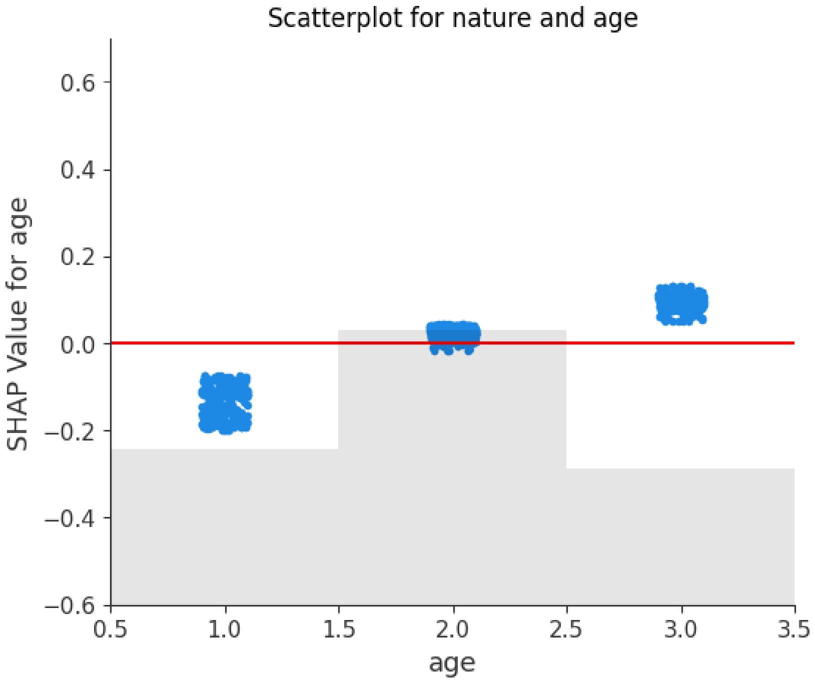


Figure A.53: SHAP Scatterplot for age parameter and nuisance choice option

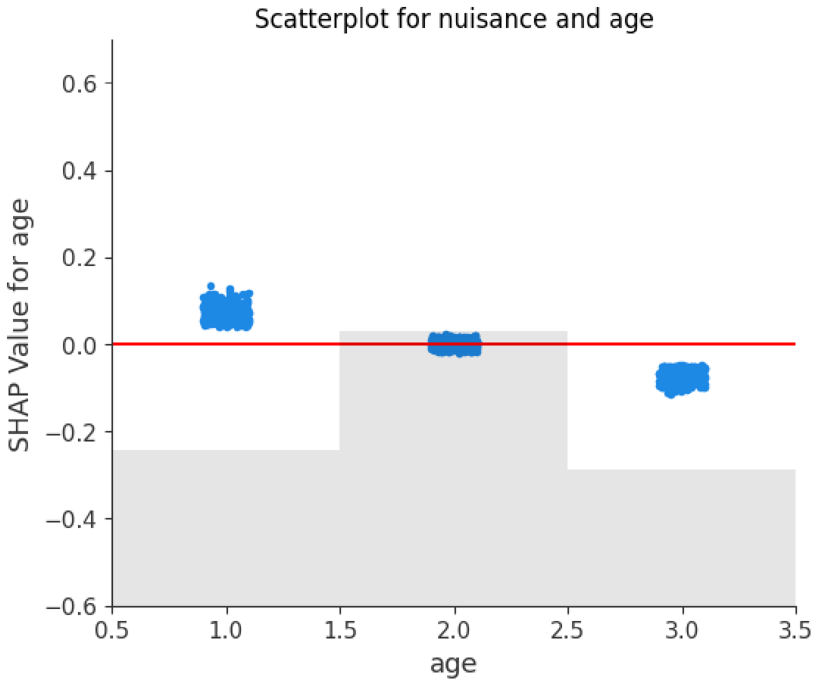


Figure A.54: SHAP Scatterplot for age parameter and compensation choice option

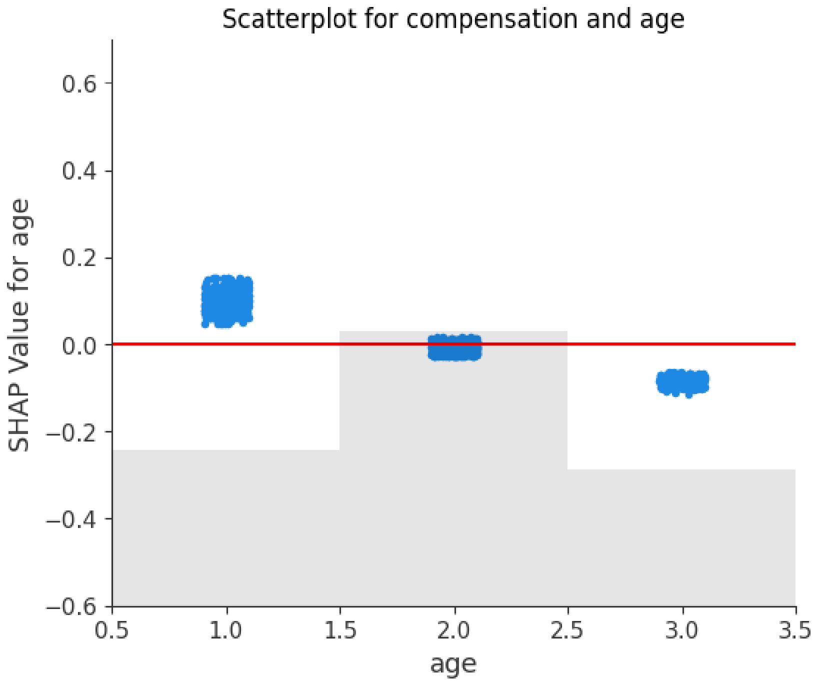


Figure A.55: SHAP Scatterplot for gender parameter and landscape choice option

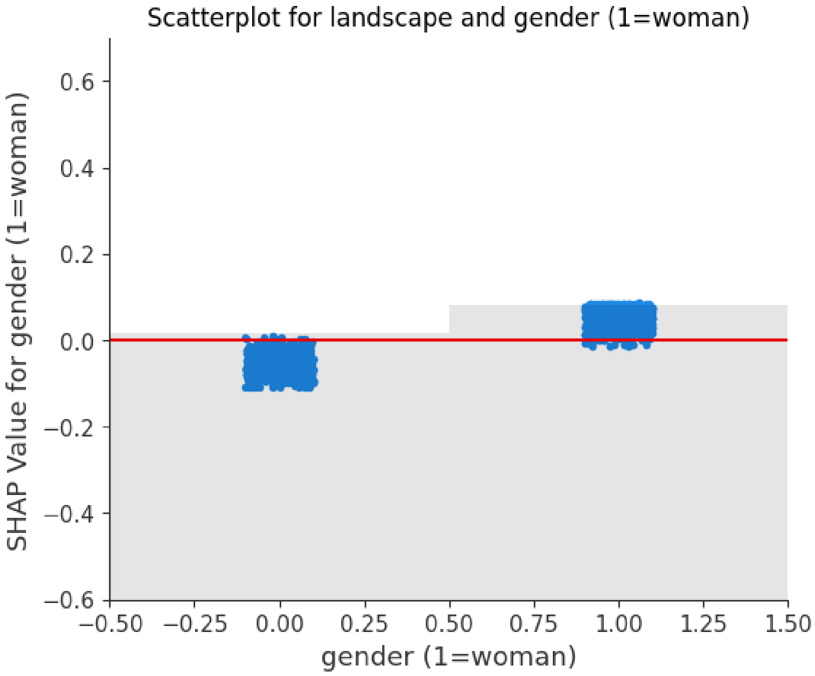


Figure A.56: SHAP Scatterplot for gender parameter and influence choice option

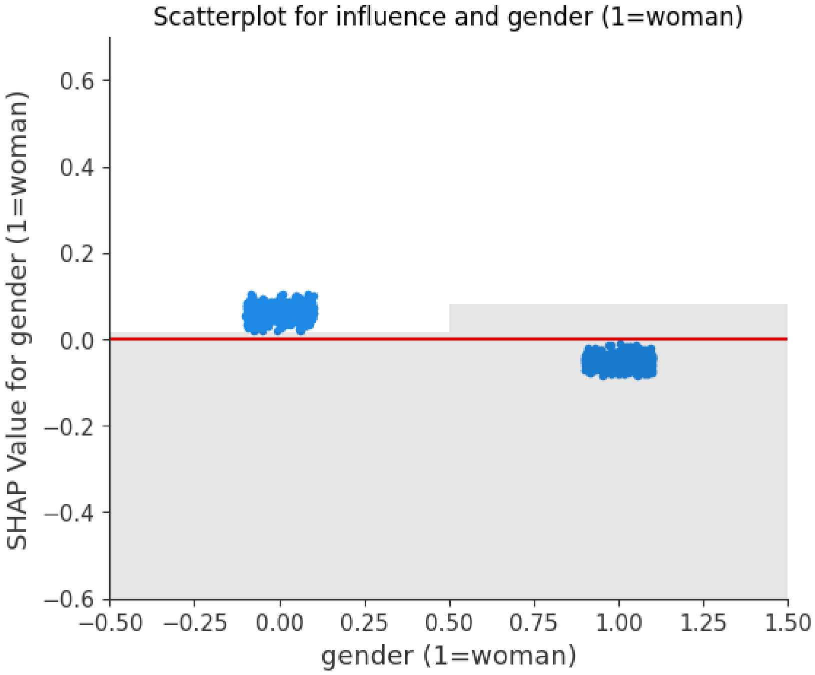


Figure A.57: SHAP Scatterplot for gender parameter and failure choice option

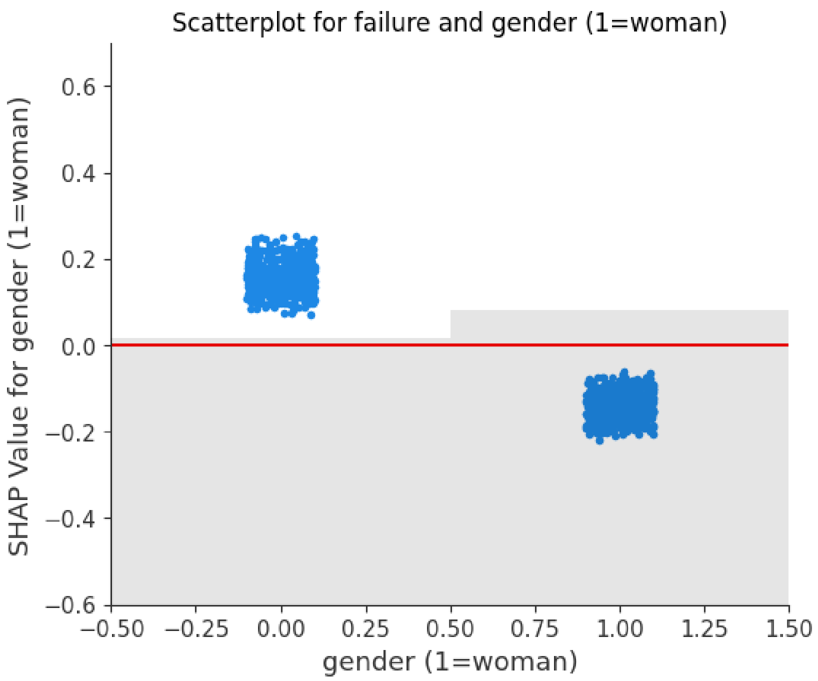


Figure A.58: SHAP Scatterplot for gender parameter and costs choice option

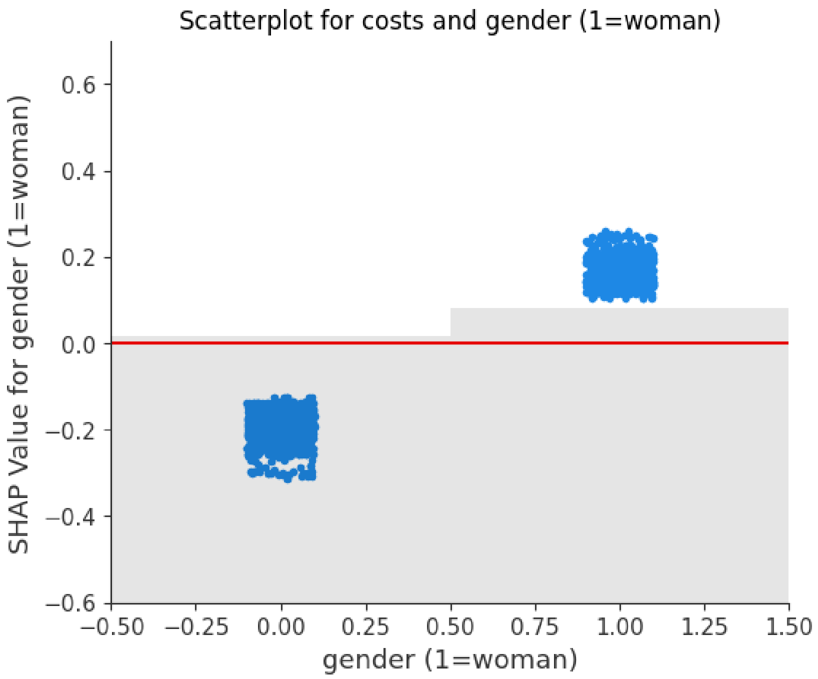


Figure A.59: SHAP Scatterplot for gender parameter and region choice option

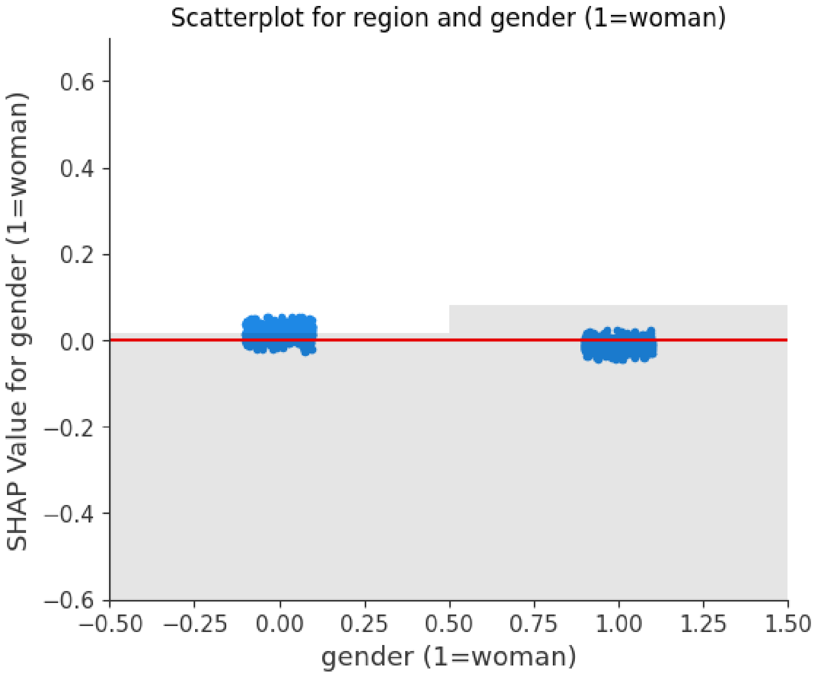


Figure A.60: SHAP Scatterplot for gender parameter and nature choice option

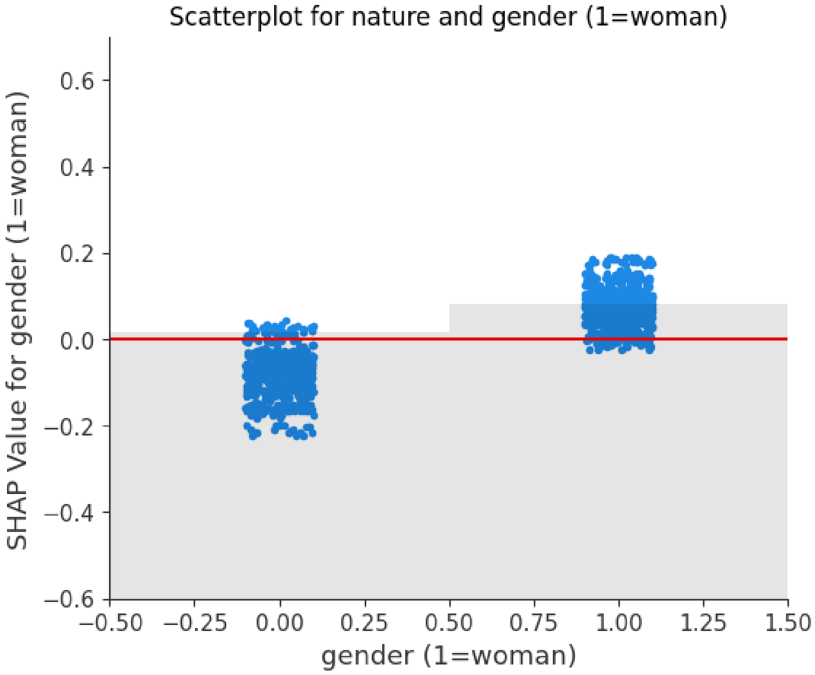


Figure A.61: SHAP Scatterplot for gender parameter and nuisance choice option

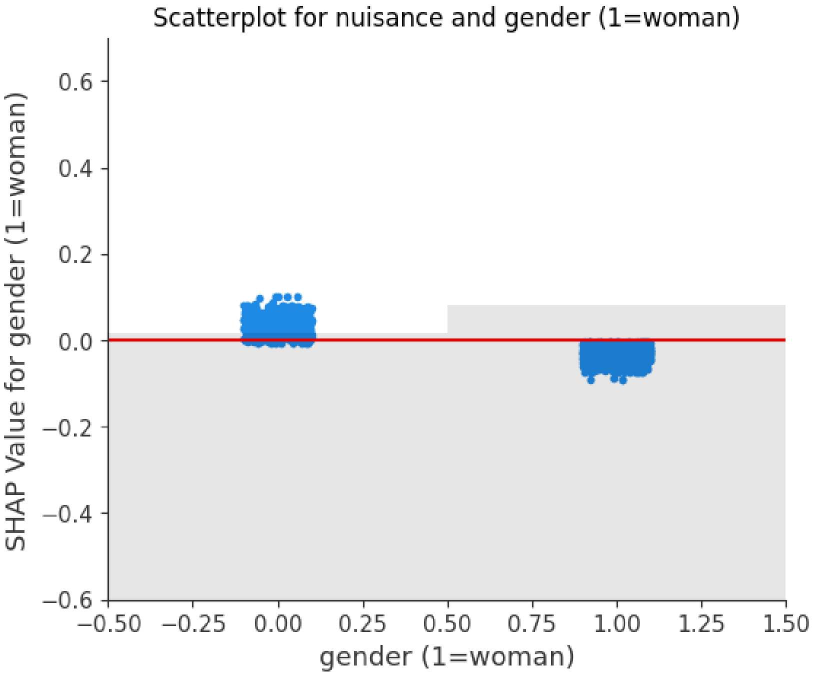
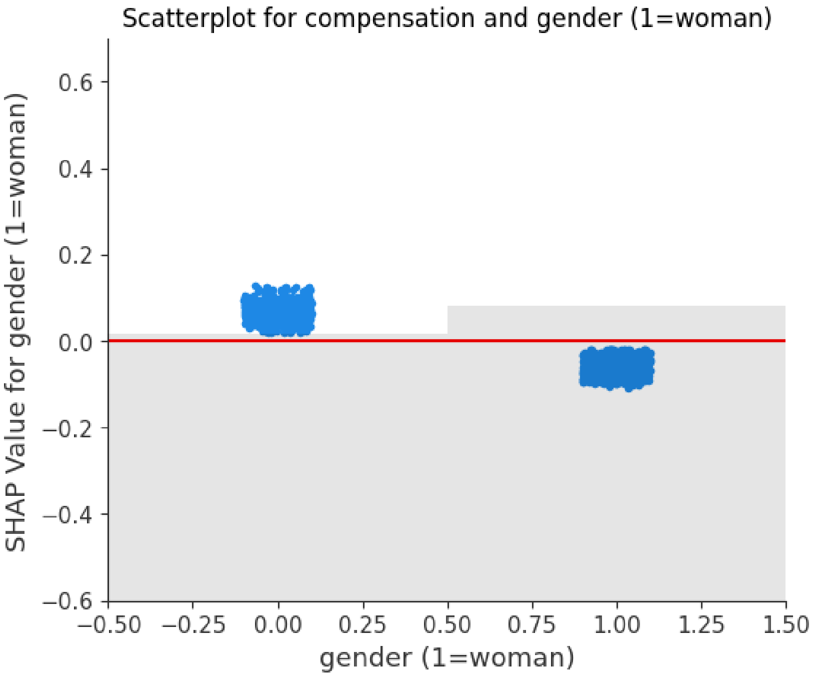


Figure A.62: SHAP Scatterplot for gender parameter and compensation choice option



B

Appendix B: Translation Table

Dutch	English
Het bestaande landschap moet zoveel mogelijk behouden blijven	The existing landscape should be conserved as much as possible
Inwoners mogen zo vroeg mogelijk invloed uitoefenen	Citizens are allowed to exert influence as soon as possible
De kans op stroomstoringen moet zo klein mogelijk zijn	The chance of grid failures should be as low as possible
De energiekosten voor Nederlanders moeten zo laag mogelijk zijn	Energy prices for Dutch citizens should be as low as possible
Regio's die meer schone energie opwekken betalen minder voor energie	Regions that produce more clean energy, pay less for energy
De bestaande natuur moet zoveel mogelijk behouden blijven	The existing nature should be conserved as much as possible
Inwoners moeten zo min mogelijk last hebben	Citizens should have the least amount of nuisance possible
Inwoners die last hebben moeten extra geld krijgen	Citizens that experience nuisance should be compensated with extra money
In dit onderzoek hebben we het advies gevraagd aan inwoners van Nederland. We kunnen ook advies vragen aan experts. Welk advies vind je het belangrijkste?	In this PVE we asked advise from Dutch citizens. We can also ask advise from experts. Which advise is most important to you?
Wat is je leeftijd?	What is your age?
Wat is de hoogste opleiding die je hebt afgemaakt?	What is the highest level of education you completed?
Welke zin past het beste bij je?	Which sentence describes your situation best?
Wat past het beste bij je?	Which statement fits you best?
De overheid moet het advies van inwoners overnemen	The government should fully follow citizen advice
Het advies van de inwoners is belangrijker dan het advies van de experts	The advice of citizens is more important than expert advise
Het advies van de inwoners is even belangrijk als het advies van de experts	Both advices are equally important
Het advies van de experts is belangrijker dan het advies van de inwoners	The advice of experts is more important than citizen advice
De overheid moet het advies van experts overnemen	The government should fully follow expert advice
Jonger dan 35 jaar	Below 35 years old
Tussen 35 en 65 jaar	Between 35 and 65 years old
65 jaar of ouder	Over 65 years old
Basisschool, Vmbo, mbo niveau 1 , Klas 1, 2 of 3 havo/vwo	Primary school, VMBO, MBO-1, Year 1/2/3 of HAVO/VWO
Klas 4, 5 of 6 havo/vwo of mbo-niveau 2, 3 of 4 (de basisberoepsopleiding, de vakopleiding, of de middenkader- en specialistenopleidingen)	Year 4/5/6 of HAVO/VWO, MBO-2, MBO-3 or MBO-4
Hbo of WO bachelor, Hbo of WO master	HBO or University Bachelor, HBO or University Master
Minder dan 50000 inwoners	Less than 50,000 inhabitants
50000 tot 150000 inwoners	Between 50,000 and 150,000 inhabitants
150000 inwoners of meer	More than 150,000 inhabitants
Ik heb iedere maand te weinig geld	Every month, I have a shortage of money
Ik heb iedere maand genoeg geld	Every month, I have enough money
Ik heb iedere maand meer dan genoeg geld	Every month, I have more than enough money
Ik ben een man	I am a man
Ik ben een vrouw	I am a woman

Table B.1: Translation table of initially Dutch statements

C

Appendix C: NP RES Data Preparation

In this section the data of NP RES is further explored. The rationale for the cleaning process is given, along with a description of every step in the process. The cleaning process is important as a model is only as good as the data it is trained on. If the data is unstructured and full of errors, the Machine Learning models have trouble making a good prediction. It will result in an unreliable model. First, I explore why we clean the data and what the goal of the cleaning process will be. After that a step-by-step description of the cleaning process is given, and to conclude a reflection on the cleaning process is provided in this section.

To get Machine Learning models to run, the data has to formatted correctly. In short this means that clear inputs and outputs need to be defined, between which the algorithm can find relations. The algorithm uses the inputs as given parameters that describe a participant. It uses these parameters to predict the output. It can evaluate the accuracy by taking the actual output and comparing that to the model prediction. For the NP RES data the outputs used are the choice task points given. The inputs used are demographic variables of the participants. The full list of demographic input parameters used is:

- Age group: Categorical data which gives the age of a participant in one of three age groups: younger than 35 years old, between 35 and 65 years old, and older than 65 years old.
- Gender: Categorical data which gives the gender of a participant in one of three categories: Male, Female, Other.
- Opinion on whether experts or citizens opinion should be more valued. Categorical data on a 5-point scale ranging from full trust on expert opinion to full trust on citizen opinion.
- Income: income scale based on self-evaluation in three categories: I have enough money, I don't have enough money, or I have an abundance of money.
- Municipal Size: Categorical data about municipal size.
- Education: Categorical data with three categories: low, middle or high-educated.

The NP RES data requires cleaning before it can be fed in the Machine Learning algorithms. The cleaning process is of influence on the final results of the analysis. Therefore an explanation is given of which steps are taken. This creates transparency about the assumptions taken in data cleaning. These assumptions are important limiting factors for the interpretation of results.

The dataset contains entries from a pilot. After the pilot, some of the wordings of choice task options were changed. Therefore all entries from the pilot are removed, to not create noise in the data. Using the started at data column, the entries from before the 6th of December 2022, 12:00 are removed. This was given as cut-off date by analysts of Populytics. All entries which were outside of the data collection period are removed, by using the consultation active criterion. All respondents who didn't give consent are removed via the consent given criterion.

It is desirable to provide a complete set of data to the machine learning algorithm. Therefore all respondents with a NaN in the choice task are removed. Furthermore there are respondents with a zero completion time. These are removed as well. Finally, it is possible to exploit some edge cases in the software to give more or less than 25 points in the choice task. Since this is not the intention of the choice task, all respondents with a summed value that is not equal to 25 are removed.

The process of cleaning results is a dataset that is cleaned from 2843 respondents, to 1534 respondents. The dataset had 58 data columns, and is now down to 16 data columns.

The data column names are not very descriptive and intuitive. Therefore these are renamed. Many data columns have answers in the shape of a text string. These can not be used by the machine learning algorithms. Therefore they need to be encoded into categories. One critical note is that in reality it is probably possible to give the string categories to the machine learning algorithms, thereby reducing the amount of manual processing steps. As this is not done here, it is a recommendation for future projects.

Categories with too few respondents in it are difficult to handle for the machine learning algorithms. If data is too scarce, the algorithms overfit on the data. This is mainly the case for NaN categories. In the NaN categories are mainly respondents who did not want to answer a question, or have chosen ”other” in the question about gender. Since we can not meaningfully draw conclusions about these categories due to the lack of respondents in those categories, they are removed after the renaming process. A dataset of 1361 respondents is exported for the machine learning algorithms.

The cleaning process involves mainly cutting out data that is irrelevant or noisy. Approximately half of the respondents is removed in the process. That means that while an analyst thinks there are many respondents, in reality only half of those are useful for a machine learning algorithm.

In future research one might investigate a less rigid cleaning process. The cleaning now was very harsh, removing all NaN values in both inputs and outputs. Future effort can be placed on selecting models and setting them up so they can deal with NaN values without disrupting the results.

Bibliography

Aggarwal, K., Mijwil, M., Garg, S., Al-Mistarehi, A.-H., Alomari, S., Gök, M., Zein Alaabdin, A., & Abdul Rahman, S. (2022). Has the future started? the current growth of artificial intelligence, machine learning, and deep learning. *Iraqi Journal for Computer Science and Mathematics*, *3*, 115–123. https://doi.org/10.52866/ijcsm.2022.01.01.013

Arnstein, S. R. (2019). A ladder of citizen participation. *American Planning Association. Journal of the American Planning Association*, *85*(1), 24–34. https://doi.org/10.1080/01944363.2018.1559388

Awan, U., Shamim, S., Khan, Z., Zia, N. U., Shariq, S. M., & Khan, M. N. (2021). Big data analytics capability and decision-making: The role of data-driven insight on circular economy performance. *Technological Forecasting and Social Change*, *168*, 120766. https://doi.org/https://doi.org/10.1016/j.techfore.2021.120766

Ayodele, T. (2010). Types of machine learning algorithms. https://doi.org/10.5772/9385

Bahamonde-Birke, F., & Mouter, N. (2019). About positive and negative synergies of social projects: Treating correlation in participatory value evaluation.

Bouwmeester, M. (2021). *Effects of goal-dependent implementation choices on the achievement of goals in participatory value evaluation processes* (Thesis).

Boxall, P. C., Adamowicz, W. L., Swait, J., Williams, M., & Louviere, J. (1996). A comparison of stated preference methods for environmental valuation. *Ecological Economics*, *18*(3), 243–253. https://doi.org/https://doi.org/10.1016/0921-8009(96)00039-0

Buskirk, T. D., Kirchner, A., Eck, A., & Signorino, C. S. (2018). An introduction to machine learning methods for survey researchers. *Survey Practice*, *11*(1). https://doi.org/10.29115/SP-2018-0004

Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, *300*, 70–79. https://doi.org/https://doi.org/10.1016/j.neucom.2017.11.077

Caputo, V., & Lusk, J. L. (2022). The basket-based choice experiment: A method for food demand policy analysis. *Food Policy*, *109*, 102252. https://doi.org/https://doi.org/10.1016/j.foodpol.2022.102252

Chen, T., & Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2939672.2939785

Chowdhury, M. S. (2006). Human behavior in the context of training: An overview of the role of learning theories as applied to training and development. *Journal of Knowledge Management Practice*, *7*(2), 1–11. https://www.scopus.com/inward/record.uri?eid=2-s2.0-69949098370&partnerID=40&md5=40fe27de84df968ed54b6bdfccad2758

Creighton, J. L. (2005). *The public participation handbook : Making better decisions through citizen involvement* (1st ed). Jossey-Bass.

Crowe, S., Cresswell, K., Robertson, A., Huby, G., Avery, A., & Sheikh, A. (2011). The case study approach. *BMC Med Res Methodol*, *11*, 100. https://doi.org/10.1186/1471-2288-11-100

Dekker, T., Koster, P., & Mouter, N. (2019). The economics of participatory value evaluation. *Tinbergen Institute Discussion Paper*, *2019-008/VIII*.

Devine, D., Gaskell, J., Jennings, W., & Stoker, G. (2021). Trust and the coronavirus pandemic: What are the consequences of and for trust? an early review of the literature. *Political Studies Review*, *19*(2), 274–285. https://doi.org/10.1177/1478929920948684

de Vries, M., Mouter, N., Jenninga, S., Tuit, C., Spruit, S., Fillerup, L., & Munyasya, A. (2022). *Participatieve waarde evaluatie over flevoland res 2.0* (Report). Populytics. https://populytics.nl/wp-content/uploads/2023/03/20221118-Eindrapport-PWE-RES-Flevoland-1.pdf

de Vries, M., Spruit, S. L., Mouter, N., Jenninga, S., & Tuit, C. (2022). *Een participatieve waarde evaluatie over windenergie in vijfheerenlanden* (Report). https://populytics.nl/cases/welke-keuzes-zouden-de-inwoners-van-vijfheerenlanden-maken-over-windmolens/

Edelenbos, J. (1999). Design and management of participatory public policy making. *Public Management: An International Journal of Research and Theory*, *1*(4), 569–576. https://doi.org/10.1080/14719039900000027

Geijsen, T., Mouter, N., Beumer, M., Jenninga, S., Tuit, C., Spruit, S., Poppe, T., & Korthals, D. (2023). *Schone energie in de toekomst: Waarmee moet de overheid rekening houden?* (Report).

Grunert, K. G., Juhl, H. J., Esbjerg, L., Jensen, B. B., Bech-Larsen, T., Brunsø, K., & Madsen, C. Ø. (2009). Comparing methods for measuring consumer willingness to pay for a basic and an improved ready made soup product. *Food Quality and Preference*, *20*(8), 607–619. https://doi.org/https://doi.org/10.1016/j.foodqual.2009.07.006

Hanley, N., Mourato, S., & Wright, R. E. (2001). Choice modelling approaches: A superior alternative for environmental valuatioin? *Journal of Economic Surveys*, *15*(3), 435–462. https://doi.org/10.1111/1467-6419.00145

Hernandez, J. I., van Cranenburgh, S., Chorus, C., & Mouter, N. (2023). Data-driven assisted model specification for complex choice experiments data: Association rules learning and random forests for participatory value evaluation experiments. *Journal of Choice Modelling*, *46*, 100397. https://doi.org/https://doi.org/10.1016/j.jocm.2022.100397

Hess, S., & Daly, A. (2014). *Handbook of choice modelling*. Edward Elgar.

Hössinger, R., Peer, S., & Juschten, M. (2023). Give citizens a task: An innovative tool to compose policy bundles that reach the climate goal. *Transportation Research Part A: Policy and Practice*, *173*, 103694. https://doi.org/https://doi.org/10.1016/j.tra.2023.103694

Irvin, R. A., & Stansbury, J. (2004). Citizen participation in decision making: Is it worth the effort? *Public Administration Review*, *64*(1), 55–65. http://www.jstor.org/stable/3542626

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Juschten, M., & Omann, I. (2023). Evaluating the relevance, credibility and legitimacy of a novel participatory online tool. *Environmental Science & Policy*, *146*, 90–100. https://doi.org/https://doi.org/10.1016/j.envsci.2023.05.001

Kahneman, D., & Knetsch, J. L. (1992). Valuing public goods: The purchase of moral satisfaction. *Journal of Environmental Economics and Management*, *22*(1), 57–70. https://doi.org/10.1016/0095-0696(92)90019-S

Kalton, G., & Flores Cervantes, I. (2003). Weighting methods. *http://lst-iiep.iiep-unesco.org/cgi-bin/wwwi32.exe/[in=ep*
19.

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, *3*, 100074. https://doi.org/https://doi.org/10.1016/j.caeai.2022.100074

Kim, S., & Lee, J. (2012). E-participation, transparency, and trust in local government. *Public Administration Review*, *72*(6), 819–828. http://www.jstor.org/stable/41688008

Köhler, W. (1967). Gestalt psychology. *Psychologische Forschung*, *31*(1), XVIII–XXX. https://doi.org/10.1007/BF00422382

Kühl, N., Goutier, M., Baier, L., Wolff, C., & Martin, D. (2022). Human vs. supervised machine learning: Who learns patterns faster? *Cognitive Systems Research*, *76*, 78–92. https://doi.org/https://doi.org/10.1016/j.cogsys.2022.09.002

Layard, R., & Glaister, S. (1994). *Cost-benefit analysis* (2nd ed.). Cambridge University Press. https://doi.org/DOI:10.1017/CBO9780511521942

Lee, J. (2020). Statistics, descriptive. In A. Kobayashi (Ed.), *International encyclopedia of human geography (second edition)* (pp. 13–20). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-08-102295-5.10428-7

Lezhnina, O., & Kismihók, G. (2022). Latent class cluster analysis: Selecting the number of clusters. *MethodsX*, *9*, 101747. https://doi.org/https://doi.org/10.1016/j.mex.2022.101747

Lund, B., & Wang, T. (2023). Chatting about chatgpt: How may ai and gpt impact academia and libraries? *Library Hi Tech News*. https://doi.org/10.2139/ssrn.4333415

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions.

MacNell, N., Feinstein, L., Wilkerson, J., Salo, P. M., Molsberry, S. A., Fessler, M. B., Thorne, P. S., Motsinger-Reif, A. A., & Zeldin, D. C. (2023). Implementing machine learning methods with complex survey data: Lessons learned on the impacts of accounting sampling weights in gradient boosting. *PLoS One*, *18*(1), e0280387. https://doi.org/10.1371/journal.pone.0280387

Mangion, M.-L., & Frendo, G. (2022). *Measuring political trust: Recognising the drivers of trust in public institutions*.

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI magazine*, *27*(4), 12–12.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*, 105–142.

MIT Sloan School of Management. (2019). Machine learning explained. https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

Mouter, N., de Vries, M., Munyasya, A., Tuit, C., Hoefsloot, N., Cieraad, F., Lier, F., & Tromp, H. (2023). *Uitkomsten van de lelylijnraadpleging* (Report). Populytics. https://populytics.nl/wp-content/uploads/2023/05/230515-Rapport-Lelylijn-PWE-raadpleging-Populytics.pdf

Mouter, N., Geijsen, T., Tuit, C., de Vries, M., Spruit, S. L., & Fillerup, L. (2022). *Een participatieve waarde evaluatie over de inrichting van de maatschappelijke raad schiphol en het omgevingshuis* (Report). Populytics. https://populytics.nl/wp-content/uploads/2022/07/Rapport-PWE-MRS-en-Omgevingshuis-Populytics.pdf

Mouter, N., Hernandez, J. I., & Itten, A. V. (2021). Public participation in crisis policymaking. how 30,000 dutch citizens advised their government on relaxing covid-19 lockdown measures. *PLoS ONE*, *16*(5). https://doi.org/10.1371/journal.pone.0250614

Mouter, N., Koster, P., & Dekker, T. (2021a). Contrasting the recommendations of participatory value evaluation and cost-benefit analysis in the context of urban mobility investments. *Transportation Research Part A: Policy and Practice*, *144*, 54–73. https://doi.org/10.1016/j.tra.2020.12.008

Mouter, N., van Beek, L., de Ruijter, A., Hernandez, J. I., Schouten, S., van Noord, L., & Spruit, S. L. (2021). *Brede steun voor ambitieus klimaatbeleid als aan vier voorwaarden is voldaan* (Report). TU Delft & Populytics. https://populytics.nl/wp-content/uploads/2022/06/Eindrapport-Klimaatraadpleging-Populytics.pdf

Mouter, N., Jara, K. T., Hernandez, J. I., Kroesen, M., de Vries, M., Geijsen, T., Kroese, F., Uiters, E., & de Bruin, M. (2022). Stepping into the shoes of the policy maker: Results of a participatory value evaluation for the dutch long term covid-19 strategy. *Social Science & Medicine*, *314*, 115430. https://doi.org/https://doi.org/10.1016/j.socscimed.2022.115430

Mouter, N., Koster, P., & Dekker, T. (2019). An introduction to participatory value evaluation. *Tinbergen Institute Discussion Paper*, *2019-024/V*.

Mouter, N., Koster, P., & Dekker, T. (2021b). Participatory value evaluation for the evaluation of flood protection schemes. *Water Resources and Economics*, *36*, 100188. https://doi.org/https://doi.org/10.1016/j.wre.2021.100188

Mouter, N., Shortall, R. M., Spruit, S. L., & Itten, A. V. (2021). Including young people, cutting time and producing useful outcomes: Participatory value evaluation as a new practice of public participation in the dutch energy transition. *Energy Research & Social Science*, *75*, 101965. https://doi.org/https://doi.org/10.1016/j.erss.2021.101965

Mulderij, L. S., Hernández, J. I., Mouter, N., Verkooijen, K. T., & Wagemakers, A. (2021). Citizen preferences regarding the public funding of projects promoting a healthy body weight among people with a low income. *Social Science & Medicine*, *280*, 114015. https://doi.org/https://doi.org/10.1016/j.socscimed.2021.114015

Paparini, S., Green, J., Papoutsi, C., Murdoch, J., Petticrew, M., Greenhalgh, T., Hanckel, B., & Shaw, S. (2020). Case study research for better evaluations of complex interventions: Rationale and challenges. *BMC Medicine*, *18*(1), 301. https://doi.org/10.1186/s12916-020-01777-6

Patricia Anne, B. (2008). A review of the literature on case study research. *1*(1).

Pope, C., Ziebland, S., & Mays, N. (2000). Qualitative research in health care. analysing qualitative data. *Bmj*, *320*(7227), 114–6. https://doi.org/10.1136/bmj.320.7227.114

Populytics. (2022). 6881 nederlanders denken mee over medisch keuren van rijbewijshouders. https://populytics.nl/wp-content/uploads/2022/06/Rapport-raadpleging-Optimalisatie-Stelsel-Medische-Rijgeschiktheid-Populytics.pdf

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). ”why should i trust you?” explaining the predictions of any classifier. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778

Rotteveel, A., Lambooij, M., Over, E., Hernandez, J. I., Suijkerbuijk, A., de Blaeij, A., de Wit, G., & Mouter, N. (2022). If you were a policymaker, which treatment would you disinvest? a participatory value evaluation on public preferences for active disinvestment of health care interventions in the netherlands. *Health Economics, Policy and Law*, *17*(4), 428–443. https://doi.org/10.1017/S174413312200010X

Rozas, L. W., & Klein, W. C. (2010). The value and purpose of the traditional qualitative literature review. *Journal of Evidence-Based Social Work*, *7*(5), 387–399. https://doi.org/10.1080/15433710903344116

Stevens, T. H., Belkner, R., Dennis, D., Kittredge, D., & Willis, C. (2000). Comparison of contingent valuation and conjoint analysis in ecosystem management. *Ecological Economics*, *32*(1), 63–74. https://doi.org/https://doi.org/10.1016/S0921-8009(99)00071-3

Subasi, A. (2020). Chapter 1 - introduction. In A. Subasi (Ed.), *Practical machine learning for data analysis using python* (pp. 1–26). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-821379-7.00001-1

Tuit, C. (2022). *The face validity of the participatory value evaluation method* (Thesis).

van Cranenburgh, S., Wang, S., Vij, A., Pereira, F., & Walker, J. (2022). Choice modelling in the age of machine learning - discussion paper. *Journal of Choice Modelling*, *42*, 100340. https://doi.org/https://doi.org/10.1016/j.jocm.2021.100340

Wallach, D., & Goffinet, B. (1989). Mean squared error of prediction as a criterion for evaluating and comparing system models. *Ecological Modelling*, *44*(3), 299–306. https://doi.org/https://doi.org/10.1016/0304-3800(89)90035-5