# Data-driven Methods to Study Individual Choice Behaviour

## with Applications to Discrete Choice Experiments and Participatory Value Evaluation Experiments

Hernández, J.I.

**DOI**
10.4233/uuid:c18b3f07-1bbc-4ef1-9f6d-cc116d8b4cb4

**Publication date**
2023

**Document Version**
Final published version

**Citation (APA)**
Hernández, J. I. (2023). *Data-driven Methods to Study Individual Choice Behaviour: with Applications to Discrete Choice Experiments and Participatory Value Evaluation Experiments*. [Dissertation (TU Delft), Delft University of Technology]. TRAIL Research School. https://doi.org/10.4233/uuid:c18b3f07-1bbc-4ef1-9f6d-cc116d8b4cb4

# Data-driven Methods to Study Individual Choice Behaviour

## with Applications to Discrete Choice Experiments and Participatory Value Evaluation Experiments

José Ignacio HERNÁNDEZ HERNÁNDEZ

The research leading to this dissertation has received funding from the Dutch Research Council (*Nederlandse Organisatie voor Wetenschappelijk Onderzoekm NWO*), NWO Responsible Innovation grant – 313-99-324: “Designing innovative and socially responsible heating systems”, as well as the “Development of Participative Value Evaluation within Socio-economic Cost Benefit analysis (SCBA) (CIVILIAN)” project, funded by the Dutch National Institute for Public Health and the Environment (*Rijksinstituut voor Volksgezondheid en Milieu, RIVM*).

# Data-driven Methods to Study Individual Choice Behaviour

## with Applications to Discrete Choice Experiments and Participatory Value Evaluation Experiments

**Dissertation**

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus, prof. dr. ir. T.H.J.J. van den Hagen
chair of the Board for Doctorates
to be defended publicly on
Monday 16 October 2023 at 12:30 o'clock

by

**José Ignacio HERNÁNDEZ HERNÁNDEZ**
Magister en Economía de Recursos Naturales y del Medio Ambiente
Universidad de Concepción, Chile
born in Concepción, Chile

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

| | |
|---|---|
| Rector Magnificus | Chairperson |
| Dr. mr. N. Mouter | Delft University of Technology, promotor |
| Dr. ir. S. van Cranenburgh | Delft University of Technology, promotor |

Independent members:

| | |
|---|---|
| Prof. dr. O. Cats | Delft University of Technology |
| Prof. dr. S. Hess | University of Leeds, United Kingdom |
| Prof. dr. N. J. A. van Exel | Erasmus University Rotterdam |
| Dr. T. Hillel | University College London, United Kingdom |

Non-independent member:

| | |
|---|---|
| Prof. dr. ir. C. G. Chorus | Delft University of Technology |

**TRAIL Thesis Series no. T2023/14, The Netherlands Research School TRAIL**

TRAIL
P.O. BOX 5017
2600 GA Delft
The Netherlands
E-mail: info@rsTRAIL.nl

ISBN: 978-90-5584-334-3

Printed in The Netherlands

*To all those who are full of curiosity.*

José Ignacio HERNÁNDEZ HERNÁNDEZ

# Preface

This thesis summarises four years of research work, five novel studies, thousands of lines of codes and a bunch of ideas that will eventually see the light. Certainly, this work could not be the same without the insightful supervision of Niek Mouter and Sander van Cranenburgh as my promotors, and Caspar Chorus as my promotor during the first three years of my PhD. Niek, I will always be grateful for your trust in me and my skills, for your enthusiasm to make me think out of the box, and for reminding me, time by time, when it is wise to stop. Sander, thank you for encouraging me to push further during my research, for making me never succumb to conformism, and for constantly flooding me with new research ideas. Caspar, during our three years working together, I always found a wise word in you, either by finding meaning in my immature research ideas at the beginning of my PhD or by making me think how all the dots connect into a sounding storyline, which after some time it became this thesis. To you three, thank you for your trust and support, which allowed me to write this thesis.

Many thanks to my defence committee members, who spent time during the summer to assess the contribution of this PhD thesis. Thanks to all those anonymous referees who took a moment to read my work and provide me with comments to improve my work. Thanks to all those people who provided me with feedback during conferences, seminars and colloquia, both online and in real life. And, of course, many thanks to all my PhD fellows who gave me feedback on this work. We often forget how much undervalued the (unpaid) reviewing work is. Properly recognising their effort in a few lines of text is not enough compared to the added value of their feedback that improved my research work.

To the friends and comrades I've met while abroad, thank you for all your academic and emotional support during this adventure. Special thanks to my friends from "The FF club" in Toulouse: Oscar and Sébastien. I will never be thankful enough for all your support during my time in France, from forcing me to speak in French to our countless hours in the Resto' U or having a coffee next to the Garonne. To my friends from The Netherlands: Baiba, Steven, Karen T., Ylenia, Joao, Mahendra, Felipe, Francisco, Cona, Karen O., Rocio, Veronica, Amanda, Gabriel; to my current and former officemates: Ahmad, Mahsa, Hong, Patrick, Mehrnaz, Kristel, Jeroen, Pelin, Lion and Floris; to my cooking buddies from the Cooking Club; to the Chilean community in Delft; To my many friends I've met in Europe (Alice, Alvaro, Vale, Camille); and in general, to all those people who made my staying in Europe a bit more comfortable.

To my close ones from Chile. To Alvaro, Sebastian, Carlos and Pablo. Although our contact is not as frequent as before, I always carry you in my thoughts since that moment in 2011 when we had the crazy idea of putting our minds on the works to contribute to a better future. To Felipe, my master thesis advisor, mentor and guide. I would never be thankful enough for all your lessons and support. To my friends from the university, although we parted in different ways, I am glad we are still in contact. To Camila, my confidant and friend for ten years. To Manuel and Alex, my friends and former flatmates. And to all those friends that are still there, either physically or in my thoughts. And to Nicole, for her invaluable support and advice during the last weeks of polishing this work.

A special place in my acknowledgements is for the Toulouse School of Economics in France, my first stop abroad before landing in Delft. Oftentimes, we forget that lessons come more frequently from our failures. In TSE, I learned one of the most valuable lessons in my life: it is okay not to succeed all the time and you always deserve a second chance. This thesis -and the second chance I gave myself- couldn't be possible without that year in TSE.

Finally, I would like to thank my family. We are far from being an archetype of family, but I am always thankful for your unconditional love during all these years.

# Contents

# Chapter 1

# Introduction

## 1.1. Research background

Since its origins in the 1970s (McFadden, 1974), choice modelling has become an important field of study in diverse areas, such as transportation (Ben-Akiva et al., 1985), health economics (Lancsar & Burge, 2014), environmental economics (Haab & McConnell, 2002; Carson & Czajkowski, 2014) and marketing (McFadden, 1986; Chandukala et al., 2008). Over the years, researchers in the choice modelling field have developed several methods to collect and model individual choices, e.g., stated choice (SC) experiments or econometric models of choice behaviour. These methods allow researchers and policymakers to, for instance, understand individuals' preferences in different contexts (e.g., as consumers, citizens, patients, etc.), derive economic values (e.g., willingness to pay, willingness to accept), determine the acceptability for new policy interventions and predict behaviour under new scenarios (e.g., a new train line is proposed).

The methods developed by choice modellers can be divided in two categories:

- <u>Methods to collect choice data:</u> Choice modellers rely on choice data. Such data may come either from observed choices (known as revealed preferences data) or from SC experiments. For years, discrete choice experiments (DCEs) have been the predominant SC experiment on empirical studies in diverse fields (see, for instance Hoyos, 2010; Carson & Czajkowski, 2014; de Bekker-Grob et al., 2012). However, the experimental setting of DCEs (i.e., choosing a single alternative among mutually-exclusive options) may not reflect how decision-makers make choices in real-life. In light of this, scholars have proposed new data collection methods that extend the DCE to consider more realistic forms of decision-making, such as bundling, choices for goods and quantities, or budget allocations

(Wiley & Timmermans, 2009; Caputo & Lusk, 2022; Carson et al., 2022; Mouter et al., 2021b).

- Models of choice behaviour: Choice modellers rely on models to analyse and extract interpretable outcomes from choice data. For years, the predominant modelling paradigm has been discrete choice models based on the Random Utility Maximisation (RUM) theory (McFadden, 1974). The RUM theory is often praised for its mathematical elegance, interpretability and connection with economics theory. However, RUM models face limitations on at least two aspects: 1) the assumption of a specific type of decision-makers behaviour (i.e., utility maximisation), 2) strict model specifications that, if misaligned with the true choice behaviour, may lead to biased outcomes and misguided advice. Furthermore, the limitations of RUM models can be more severe when these models are used to analyse data from data collection methods beyond discrete choices, as the number of possible interactions between chosen alternatives, quantities, attributes or budget allocations that decision-makers face in these experiments is considerably higher than in a DCE and, in consequence, the probability of overlooking them while specifying a choice model is higher. In response, scholars have proposed alternative modelling approaches, namely new choice models based on alternative behavioural theories (e.g., Chorus et al., 2008; Chorus, 2010; Chorus & van Cranenburgh, forthcoming) and, more recently, models that adopt a data-driven paradigm (e.g., machine learning) to learn directly from the data (van Cranenburgh et al., 2022).

This thesis contributes to the choice modelling field in the categories mentioned above, namely 1) methods to collect choice data, and 2) models of choice behaviour. In the sequel, I first provide a literature overview about the development of methods to collect choice data, from conventional DCEs to Participatory Value Evaluation (PVE) experiments, a preference elicitation framework that gained considerable attention in the last years for policymaking in the Netherlands. Then, I provide an overview concerning the methodological progress of modelling choice behaviour, from choice models to data-driven modelling approaches.

## 1.2. Literature overview

### New data collection methods

Choice modellers rely on choice data to obtain insights and provide advice. Choice data can be categorised into two paradigms. The first paradigm is revealed preference

(RP) data, which consists of observed choices made by decision-makers in real-life markets (e.g., barcode scanners or public transport card data). The second paradigm is using stated choice (SC) experiments, which is the focus of this thesis.

SC experiments are data collection methods where decision-makers make choices in hypothetical scenarios as if they were in a real-life situation. An SC experiment consists of one or two hypothetical scenarios named choice situations. Each choice situation contains a number of alternatives that are characterised by a set of attributes and levels. The alternatives, attributes and levels presented in the choice situations of a SC experiment are designed and controlled by the analyst. Then, decision-makers are asked, based on the information provided in the choice situation, to perform a specific choice.

There are various types of SC experiments. The most popular SC experiment is the DCE. A DCE is a type of SC experiment where decision-makers must do a single choice among mutually-exclusive alternatives (i.e., discrete choice) that vary on their attributes (Carson et al., 1994).

To illustrate how DCEs work, consider the following example:

| Choose your preferred alternative | | | |
|---|---|---|---|
| **Alternative** | A | B | None of them |
| **Trip time** | 15 mins. | 20 mins. | |
| **Chance of traffic jam** | 30% | 20% | |
| **Trip cost** | 10 euro | 8 euro | |
| **YOUR CHOICE** | **X** | | |

This example consists of two alternatives, plus an "opt-out" option. Each alternative is characterised by the trip time, the chance of being in a traffic jam and trip cost. A decision-maker must select one of the alternatives presented in the experiment, or he/she can "opt-out".

DCEs are highly customisable data collection tools. Notably, in a DCE, the analyst controls the number of alternatives of each choice set, the number of attributes and attribute levels of each alternative, whether the alternatives are labelled and whether there is an "opt-out" (i.e., no-choice) alternative (see Hensher et al., 2005, for a detailed manual to design DCEs). Furthermore, the analyst can explicitly specify restrictions such that certain combinations of alternatives or attributes do not appear in the same choice situation, in order to avoid dominant (dominated) alternatives or to capture preferences for specific choice situations that are of the researchers' interest.

Given this versatility, DCEs are highly popular for collecting stated choices in diverse fields (see, for instance Hoyos, 2010; Carson & Czajkowski, 2014; de Bekker-Grob et al., 2012).

However, traditional DCEs have a key limitation. In a DCE, respondents are forced to select one alternative among mutually-exclusive options, which has three implications. Firstly, it is implicit that the alternatives of a DCE are perfect substitutes of each other. Secondly, the consumed quantity of the chosen alternative of a DCE is taken as given (i.e., the individual fully consumes the selected alternative). Thirdly, budget restrictions in a DCE are ignored or taken as exogenous to the choice situation. However, in real life, decision-makers often choose more than one good at the same time, they choose how much to consume, they choose how much to spend a scarce budget on their consumed goods, and the involved goods can be imperfect substitutes for each other or they can be complements. Some examples of these situations are, namely, vacation trips, supermarket purchases, or being a policymaker who must allocate a limited budget to public projects. In these contexts, DCEs cannot completely capture decision-makers preferences. Furthermore, arguably, DCEs suffer from *hypothetical bias* (Haghani et al., 2021), as the choice situation presented is unrealistic and, as a consequence, the choices made by decision-makers in such experiments may not reflect their “true” preferences in real life.

In light of this, scholars have proposed new SC experiments that extend DCEs by incorporating more realistic decision rules. Some examples of such decision rules include bundling or portfolio choices, i.e., choosing more than one alternative at once (Ben-Akiva & Gershenfeld, 1998; Wiley & Timmermans, 2009; van Cranenburgh et al., 2014; Caputo & Lusk, 2022), volumetric choices or choices over alternatives and quantities (Carson et al., 2022) and choices over budget expenditure (Neill & Lahne, 2022; Costa-Font * & Rovira, 2005).

A recently developed SC experiment that has received increasing interest from policymakers in the Netherlands is Participatory Value Evaluation (PVE). To date, PVE experiments have been applied in diverse fields, such as infrastructure projects (Mouter et al., 2021a,b), energy (Mouter et al., 2021c) and healthcare (Rotteveel et al., 2022; Mulderij et al., 2021). PVE is a preference elicitation framework based on an SC experiment, where individuals are asked to select their preferred combination of alternatives without surpassing a set of resource constraints.

To illustrate how PVE experiments work, consider the following example:

Choose your preferred alternatives.
Maximum budget: 100M

| **Alternative** | A | B | C | D |
|---|---|---|---|---|
| **Cost** | 20M | 30M | 40M | 50M |
| **YOUR CHOICE** | **X** | | **X** | |
| **Consumed Budget** | 60M | | | |
| **Remaining Budget** | 40M | | | |

This example consists of four alternatives (A, B, C and D), and a total amount of budget (resources) of 100 million euros that cannot be overspent. Each alternative is characterised by a cost of resources of 20, 30, 40 and 50 million euros. A decision-maker who answers this choice situation can either spends an amount lower or equal to 100 million euro (e.g., alternatives A, B and D) or select no alternative and leave the full budget unspent.

PVE experiments are highly versatile, as the analyst has control over the alternatives, attributes, levels and budget presented to decision-makers. Furthermore, PVE experiments allow decision-makers to express a broader range of preferences than in traditional DCEs, namely for different combinations of alternatives, for combinations of attributes of each alternative, and for allocations of scarce resources. In contrast, DCEs allow decision-makers to express their preferences only for different mutually-exclusive alternatives and their attributes.

## Models for choice data

To obtain behaviourally-meaningful and/or policy-relevant outcomes from choice data, choice modellers rely on models. The most popular approach for analysing choice data is through discrete choice models based on Random Utility Maximisation (RUM) theory. RUM is a theory to describe individual choice behaviour based on the notion that decision-makers seek to maximise the utility derived from mutually-exclusive alternatives. Formally, under the RUM theory, the analyst assumes that decision-makers -denoted by $n$- face $J$ alternatives. The utility of an alternative $j$ is denoted by $U_{nj}$, which depends on a set of observed characteristics (e.g., attributes) and a stochastic error term, as described in equation (1.1):

$$U_{nj} = V_{nj} + \varepsilon_{nj}, \tag{1.1}$$

where $V_{nj}$ is the observed part of the utility function and $\varepsilon_{nj}$ is the stochastic term.

The development of a RUM model follows a theory-driven approach. This means that the analyst assumes that the data-generating process (DGP), i.e., the form of the utility function, is known, and data is used to find the parameters that describe the model. Specifically, the analyst assumes a functional form for $V_{nj}$ and a distribution for $\varepsilon_{nj}$. In turn, the distribution of $\varepsilon_{nj}$ determines the form of the choice probabilities of the RUM model. For instance, when $\varepsilon_{nj}$ has an i.i.d. Extreme Value (Gumbel) distribution, the choice probabilities take the form of the multinomial logit (MNL) model (McFadden, 1974), as described in equation (1.2):

$$P_{ni} = Prob(U_{ni} \geq U_{nk}, \forall k \neq i) = \frac{\exp(V_{ni})}{\sum_j \exp(V_{nj})}, \tag{1.2}$$

where $P_{ni}$ is the probability of choosing alternative $i$ by decision-maker $n$. Other specific distributions of $\varepsilon_{nj}$ result in different discrete choice models, with the notable examples of the nested logit (NL) model (McFadden, 1978; Daly, 1987) or the mixed logit (MXL) model (McFadden & Train, 2000). Welfare measures can be directly obtained from linear-additive RUM models (Small & Rosen, 1981), which makes these type of models particularly attractive in policy applications such as for policy appraisal (i.e., for a cost-benefit analysis).

Whereas RUM models are suited for discrete choices, there exists a specific type of theory-driven utility maximisation choice model for discrete/continuous choices (i.e., over alternatives and quantities) named as Kuhn-Tucker (KT) models. KT models were proposed initially by Hanemann (1984) to study contexts where individuals must choose among alternatives and their quantities. KT models assume that decision-makers maximise a random utility function that depends on their selected goods and their quantities, subject to a budget constraint. Formally, KT models seek to solve a constrained optimisation problem such as the one provided in equation (1.3):

$$\begin{aligned}
\max_{\{x_{n1}, x_{n2}, \ldots, x_{nJ}\}} \quad & U_n = U_n(x_{n1}, x_{n2}, \ldots, x_{nJ}; \beta, \varepsilon_{n1}, \varepsilon_{n2}, \ldots, \varepsilon_{nJ}) \\
\text{s.t.} \quad & \sum_j p_{nj} \cdot x_j \leq B_n \\
& x_{nj} \geq 0, \forall j = \{1, \ldots, J\}
\end{aligned} \tag{1.3}$$

The choice probabilities of this model are derived by finding the Karush-Kuhn-Tucker conditions of the optimisation problem under distributional assumptions concerning the stochastic term(s). A notable KT model that gained considerable popularity in recent years is the Multiple Discrete/Continuous Extreme Value (MDCEV) model (Bhat, 2008). The MDCEV is known for its computational tractability and subsequent developments to accommodate, namely, nested choices (Pinjari & Bhat, 2010) and multiple budget constraints (Castro et al., 2012), as well as the possibility of deriving welfare measures from this model (Lloyd-Smith, 2018)

A limitation of RUM and KT models, however, is their reliance on strict behavioural assumptions concerning, namely: 1) whether decision-makers maximise utility when making a choice, 2) the attributes and variables that affect such utility function, 3) the shape of the utility function, and 4) the distribution of the stochastic part of the utility. In response, scholars have proposed alternative choice models that account for choice behaviour that departs from the utility maximisation theory. A detailed overview of such models is provided by Chorus & van Cranenburgh (forthcoming). However, the "true" choice behaviour is ultimately unknown from the analyst perspective, whereas choice models must assume a specific type(s) of choice behaviour *a priori*. As a consequence, the analyst can still overlook relevant behavioural assumptions related to relevant variables, interactions or non-linear effects that, if not accounted for, can lead to biased estimates.

In recent years, scholars have explored the potential of data-driven methods (e.g., machine learning, data mining, nonparametric models) for choice modelling van Cranenburgh et al. (2022). Data-driven methods are a set of methodological tools and models that aim to identify patterns or approximate functions directly from the data. Data-driven methods differ from choice models in their modelling paradigms. Choice models, on the one hand, are theory-driven, and they work upon a structured theory of choice behaviour (e.g., utility maximisation or regret minimisation). This structure is assumed as known by the analyst, and data is used to find the parameters that describe the assumed theory. In data-driven methods, on the other hand, the analyst acknowledges that the data-generating process (DGP), i.e., the individual choice behaviour, is ultimately unknown from his/her perspective, and data is used to obtain an approximation of the DGP. Some data-driven methods already explored for choice modelling include nonparametric estimators (e.g., Fosgerau, 2006; Rouwendal et al., 2010), data-mining (e.g., Keuleers et al., 2001; Guo et al., 2020), machine learning models (e.g., Hillel et al., 2019; Wang et al., 2021) and artificial neural networks (ANNs) (e.g., van Cranenburgh & Alwosheel, 2019; van Cranenburgh & Kouwenhoven, 2021; Alwosheel et al., 2021)[1].

[1] While ANNs are machine learning models, their considerable development during the last years make them deserve a dedicated topic.

## 1.3. Problem statement

Literature shows that choice modellers, on the one hand, have developed new data collection tools to account for more realistic forms of decision-making. On the other hand, there is a broader recognition that decision-makers behaviour is ultimately unknown from the analyst perspective and, in consequence, data-driven methods can help to uncover such behaviour. Yet, to the author's knowledge, at least three relevant methodological and practical challenges are still unresolved in literature:

- The first -methodological- challenge concerns the increased complexity of modelling non-DCE SC experiments and, notably, PVE experiments. In a PVE experiment, decision-makers can choose among several combinations of alternatives, attributes and allocations of budget. This presents a challenge to specify the theory-driven choice models that have been developed for PVE experiments (i.e., Dekker et al., 2019; Bahamonde-Birke & Mouter, 2019) since potential interactions between chosen alternatives or attributes become relevant. However, such interactions must be manually specified by the analyst in theory-driven choice models, which can become unfeasible to do as the number of possible specifications of the utility function considerably increases with the number of alternatives and attributes involved in the PVE experiment[2]. Data-driven methods can alleviate this burden, as they can identify interactions or model the utility function directly from the data. However, this opportunity has not been explored before for SC experiments outside DCEs in general, nor for PVE experiments in particular. By exploring this, researchers and policymakers can be benefited by easing the specification process of theory-driven models through data-driven methods and by the additional insights that can be obtained from the new models derived from this.

- The second -methodological- challenge concerns the development of data-driven methods for discrete choice data, and notably supervised machine learning, that balance flexibility, interpretability and consistency with economic assumptions. Supervised machine learning, i.e., machine learning models aimed to predict a response variable, are often praised by their predictive power, given their flexibility to learn patterns from the data. However, machine learning models face challenges when it comes to interpretability (i.e., the ability to extract behaviourally-relevant insights), as their parameters have no behavioural meaning like, for instance, the taste parameters of a linear multinomial logit model. In response, in

[2] This problem also applies for discrete choice models. However, in the case of PVE experiments, this problem is even more so given their increased complexity compared to DCEs

recent years, scholars have put a greater focus on increasing the interpretability of machine learning models to study choice behaviour. In the notable case of ANNs, I highlight (e.g., Sifringer et al., 2020; Alwosheel, 2020; Alwosheel et al., 2021; Wong & Farooq, 2021) and Han et al. (2022). However, a limitation of these previous works is that they often either sacrifice flexibility to learn patterns from the data for the sake of keeping consistency with economic assumptions (e.g., RUM theory) or vice versa. As a consequence, more widespread use of these models is hindered, as the analyst ends up either with a flexible model that cannot provide economically-consistent outcomes for policymaking (e.g., willingness to pay) or an economically consistent but inflexible model that performs no better than a multinomial logit model. A new data-driven machine learning model that balances flexibility and consistency with economic assumptions can alleviate this, providing more room for these models to be used in real-life policy applications.

- The third -practical- challenge is the lack of software tools to estimate and compare data-driven methods to study individual choice behaviour. Most routines aimed to estimate data-driven methods for choice modelling consist of user-written pieces of code in different programming languages. This makes a comparison of different data-driven methods a burdensome task, hindering more widespread use of these methods in choice modelling. In contrast, there is widespread availability of unified software packages to estimate theory-driven choice models (e.g., Bierlaire, 2003; Hess & Palma, 2019; Gutiérrez-Vargas et al., 2021). A new software tool that unifies data-driven methods and allow to easily estimate and compare them can encourage researchers and policymakers to further use these methods in real-life policy applications.

## 1.4. Research goals and scope

### Main research goal and scope

Based on the problem statement, this thesis aims for the following **main research goal**:

*To investigate the extent that data-driven methods can be used for analysing individual choice behaviour from SC experiments, either to complement theory-driven choice models, or alternatives to theory-driven choice models; and to provide methodological and substantive contributions for such purposes.*

Given the methodological and practical challenges presented in the problem statement (Section 1.3), the main research goal is scoped on PVE experiments and DCEs. Specifically, the research goal is scoped such that it addresses the lack of research on: 1) the extent that data-driven methods can be used to model SC experiments outside DCEs in general and PVE experiments in particular, 2) the need for having data-driven methods for discrete choice data that can balance flexibility (to learn from the data) and consistency with economic assumptions, and 3) the lack of software tools that make these data-driven methods easily available to the research community.

### Sub-goals

In light of the problem statement, main research goal and scope presented above, the following sub-goals are presented:

- **RG1:** To examine the extent that data-driven methods can be used as a complement to theory-driven choice models for PVE experiments, and to develop methodological tools for this purpose.
- **RG2:** To examine the extent that data-driven methods can be used as alternatives to theory-driven choice models for PVE experiments, and to develop methodological tools for this purpose.
- **RG3:** To develop a new discrete choice model based on data-driven methods that balances flexibility to learn the utility function from the data, with consistency with economic assumptions.
- **RG4:** To develop a new software tool to estimate and compare the outcomes of different data-driven methods simply and conveniently.

## 1.5. Research studies and thesis outline

### Research studies

This thesis consists of five studies that address the main research goal and subgoals. Each study presented in this thesis corresponds to a scientific article submitted to different scientific journals, in a different stage of the peer-review process, including published articles. Below, a brief description of each study is provided. In addition, the original title of each scientific article, its peer-review stage and its target journal are provided. When applicable, the research question addressed by the study is also provided.

### Study 1: A large-scale deployment of a Participatory Value Evaluation experiment

- This study is published as: *Public participation in crisis policymaking. How 30,000 Dutch citizens advised their government on relaxing COVID-19 lockdown measures* in PLoS ONE.

- This study introduces PVE experiments to the reader and illustrates how they work in a real-life application. In addition, this study shows the standard approach to model PVE experiments using a theory-driven choice model based on Kuhn-Tucker models, what behaviourally-relevant outcomes are obtained and interpreted, and which challenges emerge from analysing PVE experiments with these models.

In this study, a large-scale PVE experiment was conducted for the very first time, in which 30,000 Dutch citizens advised the government about which COVID-19 restrictions should be relaxed during the first wave of the SARS-CoV-2 pandemic in April-May 2020. Respondents of this PVE experiment were able to select among eight COVID-19 measures that the government was considered to relax, ranging from lifting restrictions on specific sectors (e.g., hospitality, visits at nursing homes) to full relaxations on specific population groups. Respondents faced trade-offs between casualties derived from COVID-19, reductions of psychological stress, improvements in the economy and pressure on the healthcare system.

To the date this study was developed, the number of empirical applications of PVE experiments was limited. Furthermore, the studies conducted to that moment relied on samples composed of few respondents, in comparison to the complexity of the experiment at that moment. To illustrate this point, a previous PVE experiment presented 16 different alternative policies and collected a sample size of approximately 2,500 respondents.

### Study 2: Data-driven methods to assist choice models for Participatory Value Evaluation experiments

- This study is published as: *Data-driven assisted model specification for complex choice experiments data: Association rules learning and random forests for Participatory Value Evaluation experiments* in the Journal of Choice Modelling.

- This study addresses RG1: *To examine the extent that data-driven methods can be used as a complement to theory-driven choice models for PVE experiments, and to develop methodological tools for this purpose.*

PVE experiments are a type of SC experiment that poses a realistic choice situation. However, modelling PVE experiments has proven difficult, as accounting for all combinations of attributes and potential interactions between alternatives is computationally intractable. As a consequence, analysts often rely on simple model specifications that may overlook relevant interactions between alternatives, attributes and/or budget allocations. This, in turn, can lead to biased estimates, misleading interpretation of parameters and, as a consequence, misguided policy advice. In this study, three procedures based on data-driven methods, namely association rules (AR) learning and random forests (RF), are proposed to assist the specification of choice models for PVE experiments. A methodological-iterative (MI) procedure combined with AR learning and RF is used to identify relevant interactions and covariates directly from the PVE choice data and use such information to iteratively correct the model specification of a portfolio choice model for PVE experiments. Additionally, the predictions of an RF model are used as a contrast with the predictions of a portfolio choice model to identify the validity of the behavioural assumptions of the latter. The resulting assisted choice models led to model fit and behavioural interpretation improvements, compared with manually-specified portfolio choice models. The results of this study show the potential of data-driven methods to complement the currently developed choice models for PVE experiments.

### Study 3: Explainable artificial intelligence to study Participatory Value Evaluation experiments

- This study is currently under review as: *Explaining citizens' policy support for reimposing COVID-19 measures in the Netherlands with machine learning techniques.*
- This study addresses RG2: *To examine the extent that data-driven methods can be used as alternatives to theory-driven choice models for PVE experiments, and to develop methodological tools for this purpose.*

In the last years, there is an increasing interest in machine learning analysis of choice experiments data. While machine learning models are proven a promising method for predicting choice data, they have the key limitation of not being easily interpretable. This study proposes the use of SHAP, an explainable artificial intelligence (XAI) technique, to explain the preferences of individuals in a PVE experiment for reimposing COVID-19 measures in the Netherlands. SHAP is an XAI technique based on Shapley Values, a concept of game theory to determine the contribution of the covariates for specific predictions of a model. Therefore, SHAP can be used to explain the predictions of an otherwise “opaque” machine learning model. In addition, the results of SHAP are compared with the insights obtained from two conventional data analysis tools of PVE experiments, namely choice models and latent class cluster analysis (LCCA). The results of this study show that SHAP reaches a higher level of detail that the previous methods, as it provides explanations to predictions at the individual level, as a difference from the alternative methods (i.e., choice models and LCCA) that provide insights for ’average citizens’ or averages across specific population groups. Furthermore, SHAP allows to identification of non-linear effects and clusters of preferences that are otherwise hidden from the analyst. This study investigates the potential of data-driven methods as an extension of conventional data analysis approaches.

**Study 4: An economically-consistent discrete choice model based on artificial neural networks**

- This study is currently under review as: *An economically-consistent discrete choice model with flexible utility specification based on artificial neural networks.*

- This study addresses RG3: *To develop a new discrete choice model based on data-driven methods that balances flexibility to learn the utility function from the data, with consistency with economic assumptions.*

ANNs are gaining increasing popularity for modelling discrete choice data, as they can learn complex interactions and reach higher prediction performance than a conventional discrete choice model. However, extracting behaviourally- and economically-relevant information from ANNs has proven challenging since the parameters of these data-driven methods lack behavioural interpretations. Furthermore, while previous efforts attempted to extract such information from ANNs by incorporating prior expert knowledge, such approaches either restrict the utility functions of the model in such a way that the resulting interpretable measures, namely marginal utilities or willingness to pay, capture a limited range of preferences from individuals, or they break core assumptions to make the ANN consistent with RUM theory. This study proposes a new network structure called Alternative-Specific and Shared weights Neural Network (ASS-NN), a novel ANN model that balances flexibility of the utility functional form, consistency with RUM theory and fungibility of money, also known as “one euro is one euro”, a concept from economic theory that guarantees that the value of money is equal across alternatives for the same amount of money.

### Study 5: A new software package to estimate nonparametric models to compute the value of travel time distribution from binary choice experiments

- This study is published as: *NP4VTT: a new software for estimating the value of travel time with nonparametric models* in the Journal of Choice Modelling.
- This study addresses RG4: *To develop a new software tool to estimate and compare the outcomes of different data-driven methods simply and conveniently.*

The value of travel time (VTT) is a key concept in transport policy appraisal. To elicit the VTT, a conventional approach is through two-alternative-two-attribute choice experiments. A key advantage of two-alternative-two-attribute choice experiments is the possibility of estimating the distribution of the VTT with nonparametric estimators. However, the availability of software routines to apply such methods is rather scarce. Furthermore, estimating the VTT distribution with nonparametric methods usually involves user-written pieces of code, with makes it difficult to compare different methods and jeopardises the more widespread use of these methods. This study presents NP4VTT, a new software package written in Python to estimate and compare the VTT distribution with five different nonparametric methods, namely a local constant model, local logit, Rouwendal method, an ANN-based method and a shallow ANN model that is equivalent to a logistic regression model.

## Thesis Outline

This thesis is divided into two parts detailed in figure 1.1. The first part addresses data-driven methods for PVE experiments and consists of three chapters: Chapter 2 (Study 1) introduces a large-scale PVE experiment and shows how these experiments are analysed using theory-driven choice models; Chapters 3 and 4 (Studies 2 and 3, respectively) provide new data-driven methods to analyse PVE experiments.

The second part of this thesis addresses data-driven methods and software for DCEs and consists of two studies: Chapter 5 (Study 4) provides a new discrete choice model based on data-driven methods for analysing discrete choice data; Chapter 6 (Study 5) provides a new software package with to estimate and compare the outcomes of different data-driven methods. Finally, Chapter 7 summarises the conclusions of each chapter, provides overall conclusions of this thesis and discusses implications and further research directions.

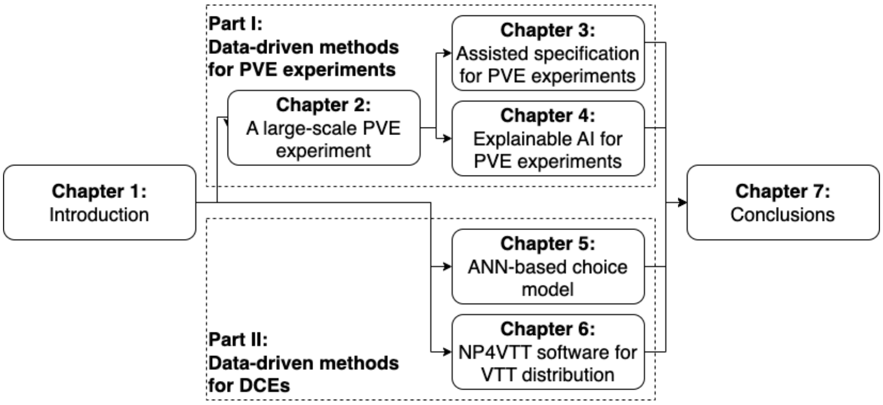


*Figure 1.1: Organisation of this thesis*

# Bibliography

Alwosheel, A., S. van Cranenburgh, C. G. Chorus (2021) Why did you predict that? Towards explainable artificial neural networks for travel demand analysis, *Transportation Research Part C: Emerging Technologies*, 128, p. 103143.

Alwosheel, A. S. A. (2020) Trustworthy and Explainable Artificial Neural Networks for Choice Behaviour Analysis.

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

Ben-Akiva, M., S. Gershenfeld (1998) Multi-featured products and services: Analysing pricing and bundling strategies, *Journal of Forecasting*, 17(3-4), pp. 175–196.

Ben-Akiva, M., S. Lerman, S. Lerman (1985) *Discrete Choice Analysis: Theory and Application to Travel Demand*, MIT Press.

Bhat, C. R. (2008) The multiple discrete-continuous extreme value (MDCEV) model: Role of utility function parameters, identification considerations, and model extensions, *Transportation Research Part B: Methodological*, 42(3), pp. 274–303.

Bierlaire, M. (2003) BIOGEME: A free package for the estimation of discrete choice models, in: *Swiss Transport Research Conference*, CONF.

Caputo, V., J. L. Lusk (2022) The Basket-Based choice Experiment: A method for food demand policy analysis, *Food Policy*, 109, p. 102252.

Carson, R. T., M. Czajkowski (2014) The discrete choice experiment approach to environmental contingent valuation, in: *Handbook of Choice Modelling*, Edward Elgar Publishing, pp. 202–235.

Carson, R. T., T. C. Eagle, T. Islam, J. J. Louviere (2022) Volumetric choice experiments (VCEs), *Journal of Choice Modelling*, p. 100343.

Carson, R. T., J. J. Louviere, D. A. Anderson, P. Arabie, D. S. Bunch, D. A. Hensher, R. M. Johnson, W. F. Kuhfeld, D. Steinberg, J. Swait (1994) Experimental analysis of choice, *Marketing letters*, 5, pp. 351–367.

Castro, M., C. R. Bhat, R. M. Pendyala, S. R. Jara-Díaz (2012) Accommodating multiple constraints in the multiple discrete–continuous extreme value (MDCEV) choice model, *Transportation Research Part B: Methodological*, 46(6), pp. 729–743.

Chandukala, S. R., J. Kim, T. Otter, P. E. Rossi, G. M. Allenby (2008) Choice models in marketing: Economic assumptions, challenges and trends, *Foundations and Trends® in Marketing*, 2(2), pp. 97–184.

Chorus, C. G. (2010) A new model of random regret minimization, *European Journal of Transport and Infrastructure Research*, 10(2).

Chorus, C. G., T. A. Arentze, H. J. Timmermans (2008) A random regret-minimization model of travel choice, *Transportation Research Part B: Methodological*, 42(1), pp. 1–18.

Chorus, C. G., S. van Cranenburgh (forthcoming) Capturing alternative decision rules in travel choice models: A critical discussion, in: *Handbook of Choice Modelling*, Edward Elgar Publishing.

Costa-Font *, J., J. Rovira (2005) Eliciting preferences for collectively financed health programmes: The ‘willingness to assign’ approach, *Applied Economics*, 37(14), pp. 1571–1583.

Daly, A. (1987) Estimating “tree” logit models, *Transportation Research Part B: Methodological*, 21(4), pp. 251–267.

de Bekker-Grob, E. W., M. Ryan, K. Gerard (2012) Discrete choice experiments in health economics: A review of the literature, *Health Economics*, 21(2), pp. 145–172.

Dekker, T., P. Koster, N. Mouter (2019) The economics of participatory value evaluation.

Fosgerau, M. (2006) Investigating the distribution of the value of travel time savings, *Transportation Research Part B: Methodological*, 40(8), pp. 688–707.

Guo, X., D. Z. Wang, J. Wu, H. Sun, L. Zhou (2020) Mining commuting behavior of urban rail transit network by using association rules, *Physica A: Statistical Mechanics and its Applications*, 559, p. 125094.

Gutiérrez-Vargas, Á. A., M. Meulders, M. Vandebroek (2021) Randregret: A command for fitting random regret minimization models using Stata, *The Stata Journal*, 21(3), pp. 626–658.

Haab, T. C., K. E. McConnell (2002) *Valuing Environmental and Natural Resources: The Econometrics of Non-Market Valuation*, Edward Elgar Publishing.

Haghani, M., M. C. Bliemer, J. M. Rose, H. Oppewal, E. Lancsar (2021) Hypothetical bias in stated choice experiments: Part II. Conceptualisation of external validity, sources and explanations of bias and effectiveness of mitigation methods, *Journal of choice modelling*, 41, p. 100322.

Han, Y., F. C. Pereira, M. Ben-Akiva, C. Zegras (2022) A Neural-embedded Choice Model: TasteNet-MNL Modeling Taste Heterogeneity with Flexibility and Interpretability.

Hanemann, W. M. (1984) Discrete/continuous models of consumer demand, *Econometrica: Journal of the Econometric Society*, pp. 541–561.

Hensher, D. A., J. M. Rose, J. M. Rose, W. H. Greene (2005) *Applied Choice Analysis: A Primer*, Cambridge University Press.

Hess, S., D. Palma (2019) Apollo: A flexible, powerful and customisable freeware package for choice model estimation and application, *Journal of Choice Modelling*, 32, p. 100170.

Hillel, T., M. Bierlaire, M. Elshafie, Y. Jin (2019) Weak teachers: Assisted specification of discrete choice models using ensemble learning, in: *hEART 2019: 8th Symposium of the European Association for Research in Transportation. Budapest, Hungary.*

Hoyos, D. (2010) The state of the art of environmental valuation with discrete choice experiments, *Ecological economics*, 69(8), pp. 1595–1603.

Keuleers, B., G. Wets, T. Arentze, H. Timmermans (2001) Association rules in identification of spatial-temporal patterns in multiday activity diary data, *Transportation Research Record*, 1752(1), pp. 32–37.

Lancsar, E., P. Burge (2014) Choice modelling research in health economics, in: *Handbook of Choice Modelling*, Edward Elgar Publishing, pp. 675–687.

Lloyd-Smith, P. (2018) A new approach to calculating welfare measures in Kuhn-Tucker demand models, *Journal of choice modelling*, 26, pp. 19–27.

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

McFadden, D. (1978) Modelling the choice of residential location, *Transportation Research Record*, (673).

McFadden, D. (1986) The choice theory approach to market research, *Marketing science*, 5(4), pp. 275–297.

McFadden, D., K. Train (2000) Mixed MNL models for discrete response, *Journal of Applied Econometrics*, 15(5), pp. 447–470.

Mouter, N., P. Koster, T. Dekker (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, pp. 54–73.

Mouter, N., P. Koster, T. Dekker (2021b) Participatory value evaluation for the evaluation of flood protection schemes, *Water Resources and Economics*, 36, p. 100188.

Mouter, N., R. M. Shortall, S. L. Spruit, A. V. Itten (2021c) 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, p. 101965.

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

Neill, C. L., J. Lahne (2022) Matching reality: A basket and expenditure based choice experiment with sensory preferences, *Journal of Choice Modelling*, p. 100369.

Pinjari, A. R., C. Bhat (2010) A multiple discrete–continuous nested extreme value (MDCNEV) model: Formulation and application to non-worker activity time-use and timing behavior on weekdays, *Transportation Research Part B: Methodological*, 44(4), pp. 562–583.

Rotteveel, A. H., M. S. Lambooij, E. a. B. Over, J. I. Hernández, A. W. M. Suijkerbuijk, A. T. de Blaeij, G. A. de Wit, N. Mouter (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), pp. 428–443.

Rouwendal, J., A. de Blaeij, P. Rietveld, E. Verhoef (2010) The information content of a stated choice experiment: A new method and its application to the value of a statistical life, *Transportation Research Part B: Methodological*, 44(1), pp. 136–151.

Sifringer, B., V. Lurkin, A. Alahi (2020) Enhancing discrete choice models with representation learning, *Transportation Research Part B: Methodological*, 140, pp. 236–261.

Small, K. A., H. S. Rosen (1981) Applied Welfare Economics with Discrete Choice Models, *Econometrica*, 49(1), pp. 105–130.

van Cranenburgh, S., A. Alwosheel (2019) An artificial neural network based approach to investigate travellers' decision rules, *Transportation Research Part C: Emerging Technologies*, 98, pp. 152–166.

van Cranenburgh, S., CG. Chorus, B. van Wee (2014) Vacation behaviour under high travel cost conditions–A stated preference of revealed preference approach, *Tourism Management*, 43, pp. 105–118.

van Cranenburgh, S., M. Kouwenhoven (2021) An artificial neural network based method to uncover the value-of-travel-time distribution, *Transportation*, 48(5), pp. 2545–2583.

van Cranenburgh, S., S. Wang, A. Vij, F. Pereira, J. Walker (2022) Choice modelling in the age of machine learning - Discussion paper, *Journal of Choice Modelling*, 42, p. 100340.

Wang, S., B. Mo, S. Hess, J. Zhao (2021) Comparing hundreds of machine learning classifiers and discrete choice models in predicting travel behavior: An empirical benchmark, *arXiv preprint arXiv:2102.01130*.

Wiley, J. B., H. J. Timmermans (2009) Modelling Portfolio Choice in Transportation Research, *Transport Reviews*, 29(5), pp. 569–586.

Wong, M., B. Farooq (2021) ResLogit: A residual neural network logit model for data-driven choice modelling, *Transportation Research Part C: Emerging Technologies*, 126, p. 103050.

# Chapter 2

# A large-scale deployment of a Participatory Value Evaluation experiment

- 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), e0250614.
- All authors contributed equally to this work.

Following the outbreak of COVID-19, governments took unprecedented measures to curb the spread of the virus. Public participation in decisions regarding (the relaxation of) these measures has been notably absent, despite being recommended in the literature. Here, as one of the exceptions, we report the results of 30,000 citizens advising the government on eight different possibilities for relaxing lockdown measures in the Netherlands. By making use of the novel method Participatory Value Evaluation (PVE), participants were asked to recommend which out of the eight options they prefer to be relaxed. Participants received information regarding the societal impacts of each relaxation option, such as the impact of the option on the healthcare system. The results of the PVE informed policymakers about people's preferences regarding (the impacts of) the relaxation options. For instance, we established that participants assign an equal value to a reduction of 100 deaths among citizens younger than 70 years and a reduction of 168 deaths among citizens older than 70 years. We show how these preferences can be used to rank options in terms of desirability. Citizens advised to relax lockdown measures, but not to the point at which the healthcare system becomes heavily overloaded. We found wide support for prioritising the re-opening of contact professions. Conversely, participants disfavoured options to relax restrictions for specific groups of citizens as they found it important that decisions lead to “unity” and not to “division”. 80% of the participants state that PVE is a good method to let citizens participate in government decision-making on relaxing lockdown measures. Participants felt that they could express a nuanced opinion, communicate arguments, and appreciated the opportunity to evaluate relaxation options in comparison to each other while being informed about the consequences of each option. This increased their awareness of the dilemmas the government faces.

## 2.1. Introduction

The Corona crisis is a vivid example of a critical juncture in the history of nations (Acemoglu & Robinson, 2012). Following the outbreak of COVID-19, governments around the world took unprecedented measures to curb the spread of the virus, to protect high-risk groups and to prevent the overloading of health care systems. These government measures resulted in a range of unprecedented economic and social impacts (Weible et al., 2020). Imposing such restrictions is a significant challenge for political leaders, who are pressured to decide under time constraints, often with limited knowledge of the future course of the crisis and the impacts of their decisions. While this is common to many types of disasters, pandemics are a rising tide, with prolonged uncertainty and accumulating cases. The potential mortality, morbidity, and life disruptions are difficult to predict, but waiting to act until the facts are certain is unacceptable to many political leaders (Vaughan & Tinker, 2009). From the beginning of the crisis up to the time of writing, one can observe a myriad of national and local responses to COVID-19, which differ in the composition of the policy mix but also in the timing and intensity of policy adoption (OECD, 2020).

During periods of crisis and high uncertainty, the demand for scientific and technical expertise increases as governments and the public search for certainty in understanding problems and choosing responses (Weible et al., 2020; Lavazza & Farina, 2020). In many countries, this creates a need for what is perceived as evidence-based policymaking, which signals to the public that decisions are being made based on reasoned and informed judgments that serve the public good, rather than special interests (Cairney, 2016). Scientific and technical experts have become part of decision-making processes, as their names and images join political leaders as the face of how governments respond (Weible et al., 2020; Lavazza & Farina, 2020). For instance, the Dutch prime minister Mark Rutte has said that he navigated this crisis guided by the knowledge of health experts from the Dutch Outbreak Management Team (OMT), members of which regularly participated in official press conferences. In Germany, the Chancellor received advice from two health experts: namely Christian Drosten, head of virology at Berlin's Charité hospital and Lothar Wieler, the head of the government-funded Robert Koch-Institute (Dostal, 2020).

As scientific and technical experts become more prominent in defining problems and solutions during a crisis, the question of who is accountable for policymaking becomes more difficult to answer (Weible et al., 2020). Moreover, the increased centrality of health experts in policy networks raises questions about the extent to which other types of expertise and interests (e.g. social and economic) are sufficiently heard and the extent to which the advice of health experts produces decisions that align with society's preferences. In Germany, all virus-related policies made at the early stage of the pandemic were negotiated in an ad hoc way, largely bypassing the parliamentary system (Dostal, 2020). The core executives at the national and regional levels succeeded in rapidly concentrating decision-making power at the top of the pyramid. As Dostal (2020) concludes, the most important point of critique towards the German approach was the decision to limit the utilisation of expertise to a very small number of hand-picked experts. Avoiding 'counter-expertise' produced a form of tunnel vision among decision-makers, and many ostensibly 'neutral' expert recommendations involved value judgements and moral questions. Unsurprisingly, considerable differences in people's attitudes towards COVID-19 policies are not only visible between countries but also within, especially across regions and age groups (Sabat et al., 2020).

When government decisions misalign with citizens' preferences, society can correct political decisions by 'voting with their feet'. For instance, the government of Serbia backtracked on its plans to enforce a second lockdown after major protests, and the Dutch government decided to close schools following protests, even though health experts from the Outbreak Management Team advised against school closure. However, democracy theorists would argue that such protests may not necessarily represent the preferences of society at large, since any given policy generates its own opposition, ready to be exploited by elites and mass communication, making it difficult to detect the signal amid the noise (Dryzek et al., 2019). While without protest and opposition there would be little reason for democratic innovations (Warren & Mansbridge, 2013), government-driven public participation in COVID-19 policymaking has been notably absent (Weible et al., 2020; OECD, 2020; Partnership, 2020; Pearse, 2020). This is all the more remarkable after acknowledging that public participation is repeatedly recommended in health disaster response literature (OECD, 2020; Bernier, 2014; Schoch-Spana et al., 2006).

In a broad sense, the literature offers three rationales for involving citizens in crisis policymaking: the substantive, the normative and the instrumental rationale. The substantive rationale suggests that involving citizens will improve the quality of government decisions. Citizen participation allows a better evaluation of people's preferences towards the impacts of government policies, which can provide input for governments to align their decisions with citizens' preferences (Schoch-Spana et al., 2006; Farrell, 2009; Lewis-Kraus, 2020). Through a participatory process, the public may bring in new ideas, arguments, values and conditions that were not on the radar of (experts who inform the) decision-makers (Fung & Wright, 2003). For instance, the celebrated concept of drive-through testing was a citizen's idea (Lee et al., 2020). The normative rationale asserts that involving citizens in policymaking is 'the right thing to do' in a democracy, as citizens should have a say in (governmental) decisions that will deeply affect their lives and society (Delgado et al., 2011). According to Lavazza and Farina (Lavazza & Farina, 2020), health emergency policies that have strong ethical implications, deeply affecting people in very sensitive domains, should be participatory in character. Government-initiated participation in COVID-19 policies allows citizens to raise their voices in a more constructive and peaceful way than the protests in Serbia, Chile, Italy or the United States (Kingsley, 2020; Povoledo et al., 2020). Finally, public participation exercises can be said to be motivated by an instrumental rationale when they aim to achieve a particular predefined end, such as increasing citizens' acceptance of COVID-19 policies or restoring public trust. Greater public support for measures during a crisis can increase citizens' compliance, which in turn is likely to increase the effectiveness of non-pharmaceutical measures (SteelFisher et al., 2012; Moon, 2020).

In the Netherlands, an attempt was made to involve 30,000 Dutch citizens in policy decisions regarding relaxing lockdown measures for the period of 20 May to 20 July, 2020 through a Participatory Value Evaluation (PVE). PVE is a preference elicitation method which can ameliorate the potential misalignment between government decisions and public preferences by measuring the latter in a large and diverse group of citizens. The essence of a PVE is that citizens can give advice on government decisions in an easy-to-access manner (Mouter et al., 2021a); they are effectively put in the shoes of a policymaker. For example, in an online environment, they see: 1) which policy options the government is considering; 2) the concrete impacts of the options among which the government can choose and; 3) the constraint(s) that the government faces. Subsequently, citizens are asked to provide a recommendation to the government in terms of the policy options the government should choose, subject to the constraint(s). Individuals' preferences over (the impacts of) policy options can be determined by feeding these choices into behaviourally-informed choice models (Dekker et al., 2019). The obtained preferences can be used to rank government policies in terms of their desirability.

The essence of a PVE can be illustrated with the following example. Suppose that a government considers four policy options (A, B, C and D). Each policy results in costs (let us assume 5, 10, 15 and 20 million euros) and a range of impacts (X, Y, Z). Suppose that the government faces a public budget constraint of 20 million euros. In this case, participants in the PVE will be asked how they would suggest the government allocate the 20 million euros over the policy options while being informed about the impacts of each of the policy options.

In this paper, we report the results of the PVE regarding the relaxation of lockdown measures in the Netherlands between 20 May to 20 July 2020. The primary goal of this paper is to show what type of insights a PVE can bring to policymakers and other stakeholders who have to decide on corona policies. A secondary objective of this paper is to improve understanding of the strengths and weaknesses of PVE in terms of involving citizens into crisis policymaking. To achieve this, we compare PVE with other methods and discuss the merits, in terms of the three rationales for public participation, of PVE in involving citizens in crisis policymaking. This comparison might provide policymakers with arguments as to why PVE is an appealing and feasible participatory method in times of a pandemic. That said, we do not aim to provide a conclusive answer to the question of whether PVE is better or worse than other participatory methods.

The remainder of this paper is organized as follows: section 2.2 discusses the three rationales for public involvement in crisis policymaking. Section 2.3 reasons why PVE is an attractive method for involving citizens in crisis policymaking by comparing the method with other participatory approaches. Section 2.4 discusses our methodology. Section 2.5 presents our results and section 2.6 provides a conclusion and discussion.

## 2.2. The rationale for active public involvement in crisis policymaking

Since the outbreak of COVID-19 in the first quarter of 2020, most governments have been operating in "emergency mode". Scholars, pundits and journalists began warning at the beginning of the pandemic about risks like authoritarian power grabs, speeding up surveillance and other 'temporary' measures that will eventually outlast the pandemic Harari (2020); Mudde (2020); Roth (2020). Despite the fact that some political actors were indeed ready to exploit crises to change policies or institutions (Liu & Boin, 2020; Capano et al., 2020), effective and agile, coordinated, consultative and collaborative approaches among government and non-government actors have taken the spotlight (Moon, 2020). However, public participation in COVID-19 policymaking – using citizen advice in value-laden health policy decisions – has been notably absent (Weible et al., 2020; OECD, 2020; Partnership, 2020; Pearse, 2020). Even routine forms of obtaining public input requiring minimal effort from public officials were hardly deployed. There have been a few instances of citizen involvement in COVID-19 policymaking in South Korea, Scotland, Belgium or Estonia, which we will discuss later in this section. However, even these examples only relate to the gathering of citizens' ideas or evaluating attitudes towards new government measures. In the following passages, we present a range of prominent theoretical rationales for involving citizens in policymaking in general and crisis policymaking in particular. We classify the arguments according to Fiorino's (Fiorino, 1990) distinction between substantive, normative and instrumental justifications.

### 2.2.1. Substantive rationale

Due to the high urgency associated with decision-making during a pandemic, governments might easily overlook important details. For instance, some of the current policy plans might incorrectly assume that the public's response will be guided by an almost exclusive focus on risk beliefs about the danger of the pandemic and the likelihood of being infected. Risks are evaluated within the context of people's lives and priorities, and because of this, some risks may be judged as acceptable (Driscoll et al., 2020). For example, low-income groups might have a stronger need to ignore self-quanrantine orders or travel restrictions in order to earn money to survive, since their relative earning losses are higher than for other income groups (Driscoll et al., 2020). As studies have shown, the general public weighs pandemic policy decisions differently than professionals (who might have a tendency to view the world through a narrower lens) (Bernier, 2014). Hence, understanding how the risks and benefits of an intended policy are seen by the public will require input from groups outside the government and the health sector (Schoch-Spana et al., 2006). Through a participatory process, the public may bring in new ideas, arguments, values and conditions that were not on the radar of (the experts who inform) decision-makers (Fung & Wright, 2003). In Scotland, such an exercise by its government led to over 4,000 ideas and 18,000 comments from citizens about the lockdown (Webster, 2020). Citizens' imaginations are not necessarily constrained by legalistic, bureaucratic or scientific views of disaster management, but have the potential to be a source of collective wisdom and capability to solve problems (Schoch-Spana et al., 2006). In South Korea, the government adopted some citizen-led strategies to fight COVID-19. For example, a student in that country developed a mobile application that citizens could use to access information on confirmed patients. Furthermore, as mentioned in the introduction, the concept of drive-through testing was also a citizen's idea (Lee et al., 2020). A potential caveat is that citizens' input often needs to be produced in a short timeframe to have an impact on policy decisions. In crisis management, this window of opportunity can be rather small. Hence, once public officials have made up their minds, it can be too late for incorporating the publics' input.

### 2.2.2. Normative rationale

When citizen participation is driven by a normative rationale, it is seen as 'the right thing to do'. Citizens should have a say in governmental decisions when policies will affect their lives in significant ways Delgado et al. (2011). Deliberative scholars argue that the far-reaching involvement of citizens in the design of public policies is especially important at the time of world-changing events like a pandemic. This is because elected officials have to take ethical decisions - ones that produce clear winners and losers which are beyond the mandate they received during elections held prior to the pandemic (Pearse, 2020; Esaiasson et al., 2017; Gutmann & Thompson, 2004; Mansbridge, 1997). More importantly, the chances for greater victimization during a disaster or epidemic are unevenly distributed in society, as are the opportunities for enhanced safety. Economic means, social class, ethnicity and race, gender, and social connectedness are factors that often determine the extent of harm suffered (Schoch-Spana et al., 2007). For example, Hispanic Americans and African Americans have succumbed to COVID-19 in disproportionately higher numbers than the population as a whole (Xafis, 2020). Isolated individuals with few social ties are also more vulnerable to disasters (Dynes, 2006). Including groups that might be un(der)represented in policymaking is therefore not only the ‘right thing’ to do, but such efforts also feed positively into the substantial rational of public participation; in many responses to COVID-19, policy effectiveness was reduced by ‘blindspots’ in otherwise well-performing systems due to failure to adequately care for vulnerable groups (Capano et al., 2020). Moreover, the way we perceive the impact of government measures on the lives (and deaths) of others, will likely affect the way in which we sacrifice our personal freedoms for the benefit of the extended community. As studies and the protests in Serbia, Chile, Italy or the United States have shown, the general public weighs pandemic policy decisions differently than do professionals (who might tend to view the world from a narrower perspective) (Bernier, 2014).

### 2.2.3. Instrumental rationale

Public participation exercises can be said to be motivated by an instrumental rationale when they aim to achieve a particular predefined end, (e.g. increasing citizens’ compliance and trust). Greater public support for imposed lockdown measures can increase citizens’ compliance, which in turn is likely to increase the effectiveness of non-pharmaceutical measures (SteelFisher et al., 2012; Moon, 2020). Yet support for and compliance with a policy measure are difficult to model before that measure has been implemented (Blendon et al., 2008), since a myriad of individual, group, and subgroup responses to disease outbreaks affect attitudes and behaviour (e.g., perceived gender roles, generational differences, religious beliefs, partisanship, varying health literacy and education levels) (Vaughan & Tinker, 2009; Allcott et al., 2020). Because of the high degree of uncertainty surrounding a new type of virus, people typically do not demonstrate the ability to fully process messages from the government. They must make quick judgments, based on emotion and a general feeling towards the government, in taking action (Chuang et al., 2015). This points to a circular relationship between how citizens evaluate their expectations towards their government and their evoked measures. In their survey before and after the lockdown in Western Europe, Bol et al. (Bol et al., 2021) note that the expectation of policies was not enough to spur policy support; rather it is retrospective policy evaluation. It is worth emphasizing that, in some cases, the intrinsic sense of responsibility citizens feel might have a stronger explanatory power in terms of successfully suppressing COVID-19 outbreaks than do government measures. Unlike Taiwan or South Korea, Hong Kong's success in fighting COVID-19 cannot be attributed to an executive that acted early, forcefully and with good governance backed by the people (Tufekci, 2020). In an environment of low public trust and a lack of political legitimacy – which would together normally result in policy failure – Hong Kong's citizens decided to organize their own COVID-19 response (Hartley & Jarvis, 2020).

Overall, involving the public in crisis policymaking is not something that government regularly do. Many policymakers remain sceptical about the contributions the public can make (Hendriks & Lees-Marshment, 2019; Koskimaa & Rapeli, 2020). It is often argued that citizens are too uninformed or uninterested in politics to formulate coherent and efficient policies (Pearse, 2020). Even in normal times, many public officials have come to view the public as something that should be kept at arm's length rather than as a potential resource helping to produce better decisions on health policies (Bernier, 2014). However, if policies align with citizens' preferences, then the likelihood of effective support from citizens will be greater (OECD, 2020; Pearse, 2020). Hence, citizen ownership of exit strategies will be essential to ensure that solidarity prevails over discrimination (Gilbert et al., 2020). And as the pandemic continues unabated, polls are showing waning public satisfaction with governments' handling of the resulting crises (Pearse, 2020).

## 2.3. Positioning PVE against other participatory approaches

PVE can be conceived of as a participatory approach to effectively involve a large and diverse group of citizens in public policymaking (Mouter et al., 2021a). At the same time, PVE is also a preference elicitation technique which can be used for the economic evaluation of government policy options (Mouter et al., 2021a; Dekker et al., 2019). Hence, PVE extends the substantive rationale for citizen participation by providing policymakers with insights into the economic costs and benefits of crisis policies. This section compares PVE with other participatory approaches to improve understanding of its strengths and weaknesses in terms of involving citizens in crisis policymaking. Note that we compare PVE with archetypes of other participatory approaches described in the literature and that we are aware of the fact that specific versions of an approach might exist with a different set of strengths and weaknesses. Moreover, we focus here on public participation in crisis policymaking, not in the overall management of a public health crisis, which can also include other forms of participation. The literature provides a range of criteria for defining whether a method or a process can be conceived as a 'participatory approach', and sometimes these criteria can be quite restrictive (Rowe & Frewer, 2005). In the present paper, we classify a method as a participatory approach when it is explicitly used as a public consultation preceding a governmental policy decision.

### 2.3.1. Mini-publics

The literature offers a range of participatory methods to involve citizens in the design and evaluation of public policies which centre around deliberative mini-publics; examples include citizen assemblies and consensus conferences. In essence, a mini-public is a demographically representative sample of the population, small enough to genuinely deliberate, and representative enough to be genuinely democratic (Goodin & Dryzek, 2006). A mini-public generally consists of around 15 to 100 randomly selected citizens (there are examples with 500) who, enabled by an independent facilitator, collectively provide advice on a policy issue (Ryan & , 2014). Citizen assemblies are one example of a mini-public that has been successful in dealing with divisive and highly politicised issues such as same-sex marriage, abortion and decarbonisation measures. The purpose of a citizen assembly is to employ a cross-section of the public to study the options available to the government on certain questions and to propose answers to these questions through dialogue and the use of various methods of inquiry such as directly questioning experts (Pal, 2012).

The basic reasoning behind deliberative approaches is that a diverse and inclusive group of citizens, if given adequate information, resources and time to deliberate on a given topic, can produce an informed judgement. The Deliberative Democracy Consortium defines deliberation as "an approach to decision-making in which citizens consider relevant facts from multiple points of view, converse with one another to think critically about options before them and enlarge their perspectives, opinions, and understandings" (Deliberative democracy consortium., 2020). Participants must consider a question from multiple viewpoints, exchange perspectives, opinions, and understandings and think critically about all possible options. The emphasis is to engage participants from the affected population, without excluding social groups or marginalised views (Ryan & , 2014).

The main downside of deploying deliberative approaches for involving citizens during a pandemic is that such processes generally take a lot of time. The biggest logistical task remains the selection process, which must deliver a representative sample of a given population, as well as a range of experts from different disciplines, with different perspectives on the matter in question (Pearse, 2020). Moreover, participants must take time to educate themselves and exchange viewpoints. This is tricky because policy questions during a pandemic are highly volatile, and governments have to respond quickly to new developments. For instance, the Irish Citizen Assembly on Abortion took more than a year to produce final recommendations and the French citizen convention on climate issues lasted for six months. And even though the actual face-to-face deliberations of the Public Engagement Project on Control Measures for Pandemic Influenza in the United States (Centers for Disease Control, 2020) lasted one month, the project's duration from planning to final report lasted eight months. Another issue with deliberations is that they are more effective offline, with participants able to engage in face-to-face interactions. This is relatively difficult in times of social distancing measures that were especially stringent at the peak of the pandemic. Furthermore, deliberation is usually carried out in small groups to ensure high-quality discussions, since this is unlikely to be possible with large groups (Goodin & Dryzek, 2006). This restricts the extent to which the public may bring in new ideas, arguments, values and conditions that were not on the radar of experts and decision-makers. Indeed, in a public health crisis, the aim should be to gather and circulate as many views as possible, to ensure that policymakers are as familiar as they can be with the social landscape that any resultant policy will need to be built upon (Pearse, 2020; Dryzek & Niemeyer, 2008).

Furthermore, as Goodin (2000) argues, mini-publics should be deployed only if the views they reach are representative or at least an accurate reflection of those that would have been reached by a larger group had similar processes been feasible at that scale. It can therefore be argued that a group of 100 citizens might be too small to be able to provide a representative picture of the population's preferences regarding a pandemic which is responsible for unprecedented and multi-dimensional impacts.

Finally, due to the participation of small groups, the number of citizens who will have increased their awareness through participation is also relatively limited. The way citizens perceive the impact of government measures on the lives (and deaths) of others will be mostly limited to the participants. During the deliberation on the US pandemic influenza policy in 2007, the exercise may itself have served as a trust-building exercise for the 260 citizens and the 50 government officials and stakeholders who participated. However, it was concluded that greater use of this method may be needed to assure both groups of the soundness of plans during an influenza pandemic (Deliberative democracy consortium., 2020).

### 2.3.2. Referendum

An alternative approach for involving citizens in the evaluation of public policies is the referendum. The referendum reaches a larger and more diverse group of citizens because of its low ‘barrier to entry for participating’. The only effort that citizens have to expend is in casting their vote, Moreover, organizing referenda can be an opportunity to restore the legitimacy of public decision-making (Frey & Stutzer, 2000). The lockdown measures imposed by governments were not discussed during previous election campaigns. Thus, citizens were not given the opportunity to take them into account when transferring authority to their elected representatives, something for which a referendum can correct. However, the referendum has several disadvantages in its application to crisis policymaking. Firstly, organising a ballot during a pandemic demands a great deal of time and effort in preparation. Secondly, citizens are only asked to vote ’for’ or ’against’ a proposal in a referendum, which prevents the public from expressing the kind of nuanced opinions which can enhance policy proposals or modify them to vulnerable groups. This is even more problematic if it neglects to address the subsequent policy implications of the choices on offer (for example, if the UK votes to leave the EU, how should it go about doing so?). Multi-dimensional policy issues such as those that arise during a pandemic generally do not lend themselves to a simple ’yes’ or ’no’ response. As Offe (2017) puts it, holding referenda on substantial yet unknown long-term results will only encourage the accountability-free expression of poorly considered mass preferences and de-emphasize requirements of consistency, compromise-building, and the reflection on consequences. Moreover, a referendum does not allow citizens to transmit new ideas, arguments, values or conditions to decision-makers. Finally, if the outcome of a referendum is considered to be binding, this would limit a government in responding quickly to new scientific insights or to new developments during a highly volatile pandemic. Therefore, depending on the qualification requirements and on the kinds of policy proposals that are open for the ballot, referenda are mostly used to guide long-term strategic government decisions, rather than short-term measures and regulations (Lupia & Matsusaka, 2004).

### 2.3.3. Opinion poll/survey

Governments also consult citizens through opinion polls, in which they ask them about the extent to which they support a certain policy or to rate several policy options. Such methods can be deployed rapidly and often make use of large randomised and representative panels, or are open for anyone to participate, such as 'the big Corona study' (Study, 2020) of the Universities of Antwerp, Hasselt and KU Leuven. However, similarly to the referendum, the questions that are asked in these opinion polls are frequently too generic to be of much policy relevance. Questions such as "do you support the lockdown" or "where should wearing face masks be obligatory" may provide policymakers with a quick understanding of public opinion regarding these topics. However, polls do not provide a deeper insight into the extent to which people value one potential policy over another and how their preferences for a certain policy option are influenced by its (societal) effects (Chorus et al., 2020). Nor do such questions provide an opportunity for participants to experience the dilemma of the policymaker during a pandemic. Hence, the ability of public polling to inform policymakers is generally limited, especially when the impacts of policy trade-offs on citizens' lives are not made visible.

### 2.3.4. Participatory Budgeting

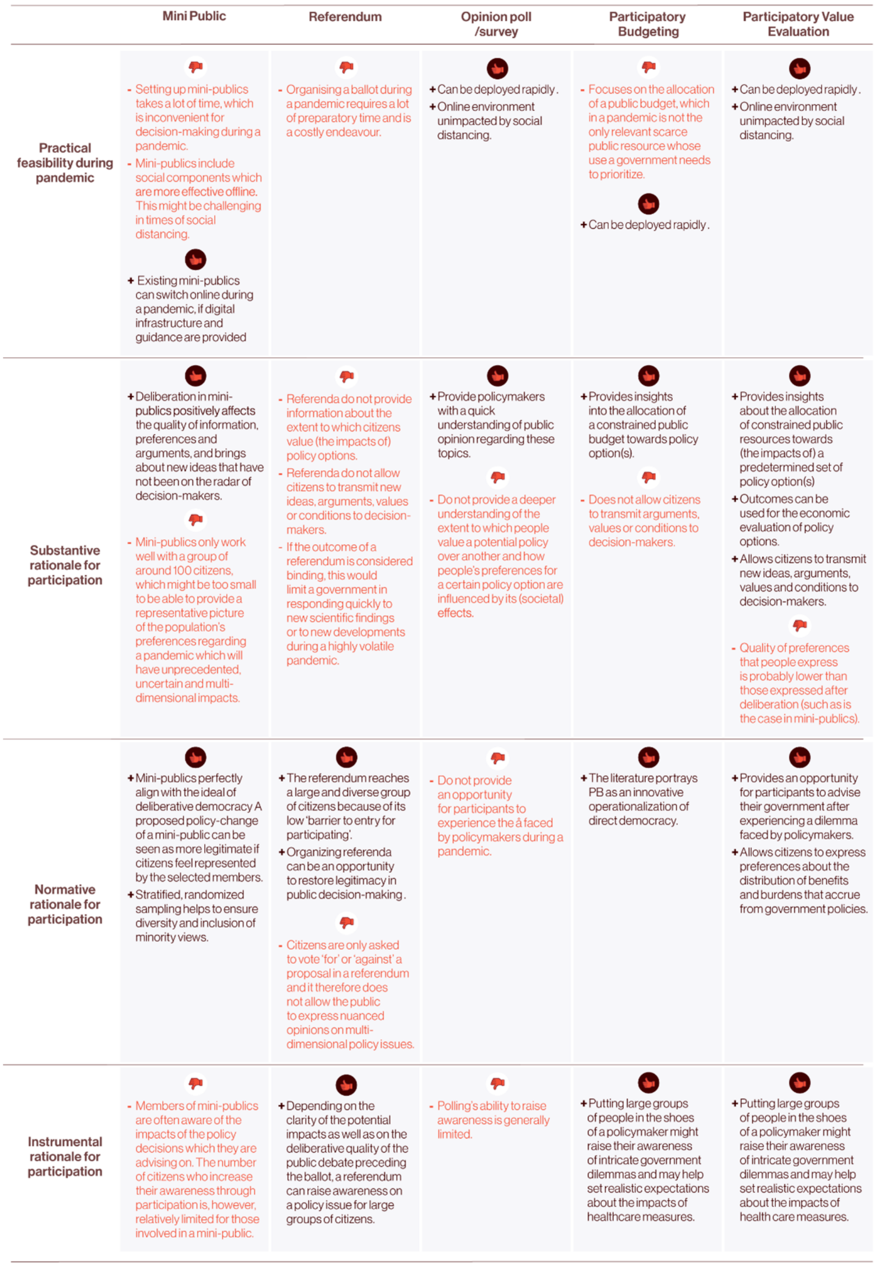
A relatively new member of the family of direct democracy institutions is participatory budgeting (PB) (Aragones & Sanchez-Pages, 2009; Cabannes, 2004). The essence of PB is that non-elected citizens are involved in the allocation of designated parts of the public budget (Sintomer et al., 2008); they do this by selecting a portfolio among the many portfolios that are possible within the budget. PB processes generally attract large and diverse groups of citizens because the barriers to entry are low. Putting large groups of people in the shoes of a policymaker might raise their awareness of intricate government dilemmas and may help set realistic expectations about the impacts of public health measures. It can be argued that PB constitutes a balancing point between the high barriers to entry and running time of mini-publics and the overly simplistic referendum/opinion poll. However, the subject of the exercise of a PB is pretty clear: to divide up a public budget. In contrast, during a pandemic, money is far from the only relevant scarce public resource over whose use a government needs to establish priorities.

## 2.3.5. Participatory Value Evaluation

Participatory Value Evaluation (PVE) closely resembles PB in the sense that citizens' optimal policy portfolios are elicited given a constraint faced by the government in allocating public resources. A fundamental difference between the two methods is that the design of a PVE can adopt other constraints than only public budget (e.g. sustainability targets, maximum pressure on the health care system). PVE has three practical advantages over PB in the sense that in theory these characteristics can also be incorporated in a PB. First, a PVE explicitly communicates to participants that they can advise against allocating public resources to the proposed policy options. That is, participants are asked whether they advise the government to allocate any resources at all, and if so, which policy options they would recommend. (Hanley et al., 2001) assert that such an experimental design, in which the baseline is clearly presented, will yield accurate estimates of the impacts of the implementation of policy options on citizens' welfare. A second practical advantage is that insights can be obtained from a PVE regarding the extent to which preferences for policy options are affected by impacts of policy options by using sensitivity analyses (we will provide examples in section 2.5.2). That is, analysts can identify how the desirability of policy options is affected by changes in impacts. Third, in a PVE, the written motivations that participants use to explain their choices provide policymakers with insights in people's arguments, concerns and values.

A difference between PVEs and mini-publics is that PVE experiments are based on individual preference formation. That is, respondents are provided with information on the policy alternatives they are meant to choose from, but they study this information individually, without the opportunity to ask questions or discuss. This approach has been criticised for implicitly or explicitly assuming that people have pre-formed preferences for quite abstract issues, such as COVID-19 lockdown measures, even when they do not have any relevant real-life experience (Czajkowski et al., 2015), or they are assumed to be able to form preferences in private based on informational material provided within the survey (Bartkowski & Lienhoop, 2017). Various scholars argue that discussions with others and the opportunity to ask questions are decisive for preference formation, as preference formation is an inherently social and dynamic process (Bartkowski & Lienhoop, 2017; Dietz et al., 2009).

Figure 2.1 provides a comparison between PVE and other participatory approaches on four dimensions. The goal of this comparison is to provide arguments as to why PVE could be an appealing and feasible participatory method in times of a pandemic. The purpose is not to provide a conclusive answer to the question of whether PVE is better or worse than other participatory methods.



| | Mini Public | Referendum | Opinion poll /survey | Participatory Budgeting | Participatory Value Evaluation |
|---|---|---|---|---|---|
| **Practical feasibility during pandemic** | 👎<br>- Setting up mini-publics takes a lot of time, which is inconvenient for decision-making during a pandemic.<br>- Mini-publics include social components which are more effective offline. This might be challenging in times of social distancing.<br>👍<br>+ Existing mini-publics can switch online during a pandemic, if digital infrastructure and guidance are provided | 👎<br>- Organising a ballot during a pandemic requires a lot of preparatory time and is a costly endeavour. | 👍<br>+ Can be deployed rapidly .<br>+ Online environment unimpacted by social distancing. | 👎<br>- Focuses on the allocation of a public budget, which in a pandemic is not the only relevant scarce public resource whose use a government needs to prioritize.<br>👍<br>+ Can be deployed rapidly . | 👍<br>+ Can be deployed rapidly .<br>+ Online environment unimpacted by social distancing. |
| **Substantive rationale for participation** | 👍<br>+ Deliberation in mini-publics positively affects the quality of information, preferences and arguments, and brings about new ideas that have not been on the radar of decision-makers.<br>👎<br>- Mini-publics only work well with a group of around 100 citizens, which might be too small to be able to provide a representative picture of the population's preferences regarding a pandemic which will have unprecedented, uncertain and multi-dimensional impacts. | 👎<br>- Referenda do not provide information about the extent to which citizens value (the impacts of) policy options.<br>- Referenda do not allow citizens to transmit new ideas, arguments, values or conditions to decision-makers.<br>- If the outcome of a referendum is considered binding, this would limit a government in responding quickly to new scientific findings or to new developments during a highly volatile pandemic. | 👍<br>+ Provide policymakers with a quick understanding of public opinion regarding these topics.<br>👎<br>- Do not provide a deeper understanding of the extent to which people value a potential policy over another and how people's preferences for a certain policy option are influenced by its (societal) effects. | 👍<br>+ Provides insights into the allocation of a constrained public budget towards policy option(s).<br>👎<br>- Does not allow citizens to transmit arguments, values or conditions to decision-makers. | 👍<br>+ Provides insights about the allocation of constrained public resources towards (the impacts of) a predetermined set of policy option(s)<br>+ Outcomes can be used for the economic evaluation of policy options.<br>+ Allows citizens to transmit new ideas, arguments, values and conditions to decision-makers.<br>👎<br>- Quality of preferences that people express is probably lower than those expressed after deliberation (such as is the case in mini-publics). |
| **Normative rationale for participation** | 👍<br>+ Mini-publics perfectly align with the ideal of deliberative democracy A proposed policy-change of a mini-public can be seen as more legitimate if citizens feel represented by the selected members.<br>+ Stratified, randomized sampling helps to ensure diversity and inclusion of minority views. | 👍<br>+ The referendum reaches a large and diverse group of citizens because of its low 'barrier to entry for participating'.<br>+ Organizing referenda can be an opportunity to restore legitimacy in public decision-making .<br>👎<br>- Citizens are only asked to vote 'for' or 'against' a proposal in a referendum and it therefore does not allow the public to express nuanced opinions on multi-dimensional policy issues. | 👎<br>- Do not provide an opportunity for participants to experience the å faced by policymakers during a pandemic. | 👍<br>+ The literature portrays PB as an innovative operationalization of direct democracy. | 👍<br>+ Provides an opportunity for participants to advise their government after experiencing a dilemma faced by policymakers.<br>+ Allows citizens to express preferences about the distribution of benefits and burdens that accrue from government policies. |
| **Instrumental rationale for participation** | 👎<br>- Members of mini-publics are often aware of the impacts of the policy decisions which they are advising on. The number of citizens who increase their awareness through participation is, however, relatively limited for those involved in a mini-public. | 👍<br>+ Depending on the clarity of the potential impacts as well as on the deliberative quality of the public debate preceding the ballot, a referendum can raise awareness on a policy issue for large groups of citizens. | 👎<br>- Polling's ability to raise awareness is generally limited. | 👍<br>+ Putting large groups of people in the shoes of a policymaker might raise their awareness of intricate government dilemmas and may help set realistic expectations about the impacts of healthcare measures. | 👍<br>+ Putting large groups of people in the shoes of a policymaker might raise their awareness of intricate government dilemmas and may help set realistic expectations about the impacts of health care measures. |

*Figure 2.1: Comparing PVE and other participatory approaches*

In conclusion, there are various reasons why PVE could be an appealing participatory approach for involving citizens in policy decisions during a pandemic. In terms of its practical feasibility, citizens can participate in a PVE online, which is appealing in times of social distancing. Moreover, a PVE can be deployed rapidly, which is important during a pandemic as governments have to respond quickly to new developments. The design of a PVE can also adopt other constraints than just the public budget, which is a key benefit compared to PB. In terms of improving the quality of decision-making (substantive rationale for participation), PVE provides information to policymakers about the extent to which the desirability of policy options is affected by the impacts of those options. It also allows citizens to transmit new ideas, arguments, values and conditions to decision-makers. From a normative point of view, a benefit of PVE is that it enables citizens to participate in multi-dimensional policy issues that do not lend themselves to a simple ‘yes’ or ‘no’ or the allocation of a constrained amount of public budget. From an instrumental point of view, letting citizens experience intricate government dilemmas improves their understanding of the social, health and economic impacts of proposed measures, which might also subsequently increase levels of acceptance and compliance.

## 2.4. Methodology

Before presenting the specifics of the PVE, section 2.4.1 compares PVE with contingent valuation (CV) and discrete choice experiments (DCE), which are two related preference elicitation techniques that can be used for the economic evaluation of government policy options. In this section, we also provide arguments as to why we selected PVE instead of these two other elicitation techniques for studying Dutch citizens’ preferences over the relaxation of lockdown measures. In section 2.4.2, we discuss the choices that we made in the design of the PVE. In section 2.4.3, we discuss the analysis techniques that were used in this study.

### 2.4.1. Comparing PVE with CV and DCE

CV is a valuation method based in surveys, designed to create a hypothetical market for public goods, and determine the amount of money that people would be willing to pay (willingness-to-pay, WTP) or accept as compensation (willingness-to-accept, WTA) for specific changes in the quantity or quality of such goods (Carson et al., 2003). CV is a popular method in the field of environmental economics for answering questions such as how to value changes in environmental quality (Carson, 2012; Halkos et al., 2020). In the CV survey, participants first receive a detailed description of a proposed government project as well as the consequences of the project. Then, they are asked whether they are willing to pay a predetermined amount of money, commonly presented as a one-time tax, to finance the implementation of the project. The CV survey is completed by a representative sample of the population, while varying the amount of money required to implement the project. In this way, it is possible to obtain an estimate of the mean WTP of the population through econometric techniques (Haab & McConnell, 2002). In turn, this mean WTP estimate represents a measure of the welfare change generated by implementing the government project (Carson & Hanemann, 2005).

While CV seems to be an effective method for determining the value of a whole project, its applicability as a preference elicitation technique is limited. Crucially, it is not possible to determine the extent to which different characteristics of the project (hereafter “attributes”) affect these preferences. Hence, CV is an attractive preference elicitation technique if the government wants to know society’s aggregate willingness to pay for one specific relaxation option, but from a CV it is not possible to infer how the aggregate willingness to pay for a particular relaxation option is affected by its impact on COVID-19 related deaths, physical injuries and mental injuries respectively.

An alternative for CV is to use a discrete choice experiment (DCE). The core idea behind DCEs is that individuals’ preferences for a government project are established by decomposing the project into separate attributes and different specifications of these attributes (referred to as ’attribute levels’) (Lancaster, 1966). The relative importance of these attributes can be empirically assessed by presenting respondents a series of choice tasks in which they are asked to choose a preferred alternative (in this case a specific relaxation option for lockdown measures) from a set of two or more alternatives with varying combinations of attribute levels (Hensher et al., 2005). By collecting the choices of a large group of respondents, statistical methods known as discrete choice models (Train, 2009) are used to estimate the preferences of individuals for policy options and attributes. These models have a solid foundation in random utility theory (McFadden, 1974), allowing researchers to compute welfare measures for changes in the quantity or quality of the attributes, and to determine the WTP of individuals for these changes (Haab & McConnell, 2002).

The literature distinguishes between labelled DCEs and unlabeled DCEs (Hensher et al., 2005). Unlabeled DCEs only focus on estimating people’s preferences for the concrete attributes of policy options and do not specify policies in terms of their nature, whereas labelled DCEs also specify the policy options which are evaluated by respondents in terms of their nature (e.g. re-opening the hospitality industry or relaxing restrictions for young citizens). The advantage of unlabeled DCEs is that it allows policymakers to use outcomes for the assessment of (combinations of policies), including those that are currently not on the table but might be considered in later phases of the crisis. A recent application of an unlabelled DCE to study the preferences for the relaxation of COVID-19 measures is provided by Chorus et al. (Chorus et al., 2020). An advantage of labelled DCEs is that it allows participating citizens to express their preferences towards a particular relaxation option regardless of the impacts that are included in the DCE.

Labelled DCE and PVE are closely related in the sense that both preference elicitation techniques allow individuals to express preferences towards specific policies as well as policy impacts. A first fundamental distinction is that participants in a DCE express preferences through selecting a single policy option, whilst participants in a PVE can select a bundle of policy options. Hence, a PVE better enables participants to evaluate policy options in relation to each other. Participants in a PVE can select one policy option or none of the options (just as in a DCE with an opt-out option), but – unlike in a DCE – they can also choose two or more options. A second fundamental distinction is that participants in a PVE express preferences not only towards specific government policies, but also towards the allocation of scarce public resources. Participants make a continuous choice regarding the extent to which they think that public resources should be allocated and discrete choices as to whether or not to include specific policy options in the bundle that they recommend to the policymaker. Participants in DCEs generally do not receive information concerning the scarcity of public resources and when such information is provided, participants are asked to recommend a single policy option from a set of policy options that all require the same investment of public resources (Mouter et al., 2019).

Whether or not a policymaker should choose PVE, (labelled or unlabelled) DCE or CV as a preference elicitation technique depends, in our view, on the policy question that should be answered. CV is an appealing technique when a policymaker wants to know whether a single relaxation option should be implemented; an unlabelled DCE is an appealing technique if the policymaker wants to know how individuals value the impacts of known and unknown relaxation options; labelled DCE is a promising elicitation technique when a policymaker wants to obtain information concerning people's preferences towards both the impacts of policy options as well as the options in and of themselves; finally, a PVE is appealing when policymakers want to know people's preferences regarding the extent to which scarce public resources should be allocated towards the (impacts of) a predefined set of options.

After the first wave of the pandemic had reasonably flattened, leaders in the Netherlands began contemplating about lifting lockdown policies. In the first week of April 2020, the research team heard from Dutch policymakers that they were expecting a major decision to be made in May. This decision concerned the ways in which the lockdown measures could be relaxed without overloading the healthcare system. Policymakers told the research team that they were considering various relaxation options which would have a range of societal impacts. We found PVE to be the most suitable preference elicitation method for this decision problem, as it concerned the allocation of scarce public resources (available capacity of the health care system) towards (the impacts of) a predetermined set of policy option(s).

### 2.4.2. Design of the PVE

We started on 9 April, 2020 with the design stage of the PVE via an online brainstorm with policymakers and researchers from the RIVM (the Dutch National Institute for Public Health and Environment), the Ministry of Health, Welfare and Sport and the Ministry of Finance about the relaxation options and impacts that they were considering. Based on this brainstorm, we compiled a shortlist of relaxation options and their impacts, which we discussed with various academics. In these meetings, we inquired as to whether we had overlooked important relaxation options and whether they could help us with providing information regarding the order of magnitude of the impacts of these strategies. For instance, we spoke with several epidemiologists to learn about the effect of relaxation options on the available capacity of the healthcare system as well as the number of deaths and people with permanent injuries caused by COVID-19. Moreover, as a result of these meetings, we included the option “All restrictions lifted in the Northern provinces”, as some academics we spoke with found this an attractive option and argued in the public debate for its inclusion (Klaassen, 2020). These researchers considered this a promising approach, since at the time that the PVE was conducted there were only a few infections in these provinces; this made it easier to keep infection levels low through testing and tracing. In addition, we decided to split the attribute ‘increase in the number of deaths caused by the relaxation option’ into ‘additional deaths of people of +70 years’ and ‘additional deaths of people younger than 70 years’ as various academics we consulted found it interesting to know whether Dutch citizens weigh the increase of mortality risk differently between these two age groups.

Based on the information and feedback we received from policy makers and academics, we selected eight relaxation options and sent a draft version of the PVE to the policymakers for feedback. In the meantime, the research team collected reports and media content to describe the eight relaxation options in the PVE and to provide estimates of the attribute levels. For instance, we used projections regarding the increase in the number of people with lasting physical injuries caused by postponed operations (Authority, 2020), data on the increase in domestic violence resulting from the corona crisis in the United Kingdom (Guardian, 2020), information on domestic violence in the Netherlands prior to the crisis (NLTimes, 2019) and estimates concerning bankruptcies, unemployment and income loss (NIBUD, 2020; Centers for Disease Control, 2020; Rabobank, 2020). We integrated this information and the feedback of policymakers into a new draft version of the PVE and this experiment was tested by a convenience sample of 80 respondents. We incorporated this feedback into the final version of the PVE.

In the PVE, participants were invited to advise the government on which lockdown measures should be relaxed between 20 May and 20 July 2020. They were asked if the government should relax lockdown measures during this period at all and, if so, which relaxation option(s) should be favoured. In an online environment, participants were presented with eight relaxation options which they could advise to the government (see Appendix 2.G for a detailed description of these options);

- Nursing and care homes allow visitors
- Re-open businesses (other than contact professions and hospitality industry)
- Re-open contact professions
- Young people may come together in small groups
- All restrictions lifted for people with immunity
- All restrictions lifted in Northern provinces
- Direct family members from other households can have social contact
- Re-open hospitality and entertainment industry

The order in which the options were presented was randomised across respondents. For each of these relaxation options, they received information regarding the option's projected impact on the pressure on the health care system (which was expressed as the percentage in which the pressure on the health care system would increase due to the relaxation option). Moreover, for each option participants received information regarding its impact on increase of deaths among people younger than 70 years and older than 70 years, increase in the number of people with permanent physical injury, decrease in the number of people with permanent mental injury and the decrease in the number of households with long-term loss of income. For example, participants were shown that the relaxation option "re-open contact professions" would reduce the number of households that lose at least 15% of their income, but increase the number of deaths among people under the age of 70. The constraint that participants faced in the PVE was the maximum capacity of the healthcare system in the sense that they were not able to recommend a bundle of relaxation options that in total resulted in a greater than 50% increase of the pressure on the healthcare system. Hence, they could only select a limited amount of relaxation options. Furthermore, participants were notified that the healthcare system could handle the pressure if it increased between 0% and 25%, that it would be overstretched if the pressure increased between 26% and 40%, and that it would be seriously overstretched if the pressure increased between 41% and 50%. After submitting their advice to the government, participants were asked to provide written motivations for their choices. Subsequently, they were asked which of the eight relaxation options should not be considered by the government and again they were asked to qualitatively underpin their choice. The main reason for including these open questions is that new arguments and ideas can emerge from the qualitative data and the government can learn about the arguments they can anticipate from those for and against specific relaxation options. Participants were also asked to answer various follow-up questions (e.g. gender, income, education and age) and they were also asked about the extent to which they themselves would experience impacts from each of the relaxation options they recommended to the government (see Appendix 2.G for more detail). The PVE is also explained in a video: https://www.youtube.com/watch?v=1D_g_HTnS50

In order to estimate how much value respondents derive from different impacts of the relaxation options, it is necessary to vary the levels of the impacts of the relaxation options across respondents. To give an example, some respondents were shown that the option ”Re-open contact professions” would lead to 200 additional deaths among people over the age of 70, while there were also respondents who saw that choosing this option would lead to 400, 600 or 1000 additional deaths in this age group. We illustrate the need for presenting different information with the following example: suppose we want to know how much money people are willing to pay for a cup of coffee and we ask 1,000 people if they would be willing to pay 50 cents for the cup of coffee. If all individuals answer ”yes” to this question, then we don't know if these people are also willing to pay 80 cents, or even $1.50 for the coffee. The analyst obtains much more information regarding people's preferences for a cup of coffee by dividing the 1,000 people into 10 groups, for example. The first group is then asked if they are willing to pay 50 cents, the second group is asked if they are willing to pay 75 cents etc. Similarly, we learn much more about people's preferences for preventing COVID-19 deaths in the context of relaxing lockdown measures by presenting respondents with different information about the impact of re-opening contact professions on deaths among people over the age of 70. Appendix 2.A details per relaxation option the possible levels of each impact. Since collecting data for all possible combinations of impacts is unfeasible in a real-life situation, we constructed 60 different profiles of relaxation options and impact levels, based on the values presented in Appendix 2.A.

To avoid an excessive correlation between impacts and between pressure levels, we followed an experimental design process of three stages. First, the number of possible impact levels were defined for each relaxation option. In the second stage, we constructed an initial design matrix of 60 rows and 48 columns, with rows representing each profile, and columns representing the impacts of each policy option. Each column is filled with random levels of the corresponding impact of each policy option, and then all columns are randomized. In the final stage, we iteratively make single changes in the values of random columns of the design matrix, and we store the resulting design on each iteration in which the correlation between impacts is reduced. This process is repeated during a certain amount of time, or after no further improvement is observed. For this design, we fixed the randomization time to ten minutes, and we observed no further improvement after three minutes approximately. Appendix 2.A provides a more detailed description of the iterative algorithm and the correlation improvement criterion.

In the PVE, we made a substantial effort to ensure consequentiality, by (truthfully) informing respondents that the outcomes of this study would be shared with the Netherlands Institute of Public Health and Environment and high-ranking policymakers at relevant ministries. Consequentiality means that respondents must feel that their choices might have real-life consequences; the literature indicates that this substantially improves the reliability of the outcomes of preference elicitation studies (Johnston et al., 2017; Carson & Groves, 2007).

We carried out the PVE with two different samples. First, a randomly selected sample from the online Kantar Public panel, which was drawn to be representative of the Dutch population (¿18 years) in terms of age and gender. Kantar Public approached members of their panel by e-mail to take part in our on-line survey and participants received a small monetary compensation. 3,358 respondents completed the experiment. The panel PVE was conducted to measure the preferences of ‘the average Dutch citizen’. A disadvantage of a ’panel PVE’ is that only Dutch citizens that are part of the Kantar Public sample can participate. For this reason, we decided to open the PVE to the general public. A disadvantage of this ’open PVE’ is that we, as researchers, have no control over which Dutch people participate and which do not. The results could be influenced by supporters or opponents of measures that mobilise many likeminded citizens. Hence, we carried out both a ’panel PVE’ and an ’open PVE’ because both have advantages and disadvantages. Participants received information on the study purpose, questionnaire content, data storage and who had access to their data before starting the questionnaire. Written informed consent was obtained at the start of the questionnaire. Our data collection effort was approved by the Ethics Board of the Delft University of Technology.

Data was collected in the period 29 April – 4 May. Because our experiment was widely covered by the media, the number of participants was far higher than expected. As a result, the server could no longer cope with the volume and the PVE was offline on 30 April between 10.00 and 15.00. Eventually, 26,293 citizens participated in this ‘open PVE’. Appendix 2.B presents the socio-demographic characteristics of the participants and provides a comparison with those of the population. Close correspondence was found between the gender distribution in the sample and the population. Highly educated respondents were over-represented in the sample. In the panel PVE middle-aged respondents were underrepresented, but in the ‘open PVE’ this age group was over-represented. In section 2.5.2, we will explore what this means in terms of the general applicability of our findings.

### 2.4.3. Analysis of the data

The econometric framework to analyse people’s choices in a PVE is a Kuhn-Tucker type choice model based in the work of Bhat (2008), developed by Dekker et al. (2019) for PVE (henceforth, the MDCEV-PVE model), and adapted for this study. This framework is rooted in the consumer’s theory of microeconomics and relies on three key assumptions. First, it is assumed that an individual chooses the bundle of policy options that maximises their utility (i.e. satisfaction), subject to satisfying the resource constraint (in this case the limited capacity of the health sector). The second assumption is that part of the utility for each relaxation option depends on the impacts that are explicitly presented to individuals. For example, an individual may prefer relaxation options that reduce economic losses. Using the MDCEV-PVE model, the researcher can estimate so-called “taste parameters” to know the importance that individuals give to each impact on their choice of policy options. Additionally, the preferences for policy options can depend on other factors not associated with the impacts. The researcher can estimate so-called policy-specific constants to determine the benefits and costs individuals obtain from specific relaxation options, irrespective of the impacts that are explicitly communicated in the PVE. These policy-specific constants can also be complemented by including individual-specific variables to analyse sociodemographic differences in the preferences for relaxation options. Third, it is assumed that an individual can derive utility not only from (the impacts of) each relaxation option, but also from the resources that are not allocated. In the context of this PVE, individuals might want to advise against allocating the full capacity of the health care system because they do not want to overstretch the system.

We proceed to briefly formalize the MDCEV-PVE model used in this paper. Let n be an individual who faces $J$ policy options and an amount of resources equal to $B$. When a policy $j$ is chosen, it consumes a portion of $B$ by an amount of $c_j$. Following Dekker et al. (2019) specification of the individual's utility function, the choice problem that individual $n$ faces is given by:

$$\begin{aligned} \max \qquad & U_n = y_0 \Psi_{n0} + \sum_j y_{nj} \Psi_{nj} \\ \text{s.t.} \qquad & \sum_j y_{nj} c_{nj} + y_0 = B, \end{aligned} \tag{2.1}$$

where $y_0$ is the amount of non-spent public resources, $y_{nj}$ is a variable that takes value 1 if the individual chooses policy option $j$ and zero otherwise, $\Psi_{n0}$ is the utility provided by the non-spent resources, whereas $\Psi_{nj}$ is the utility provided by the individual policy $j$. In the modelling, we assume that the utility for each policy option depends on the preferences for each known impact, as well as other factors apart from the impacts, encompassed in a policy-specific constant and sociodemographic characteristics. Therefore, we model the individual utility for policy options as $\Psi_{nj} = \exp\left(\delta_j + \sum_k \beta_k x_{njk} + \sum_m \theta_m z_{jm} + \varepsilon_{nj}\right)$, where $\delta_j$ is the specific constant for policy $j$, $\beta_k$ is the taste parameter for impact $k$, $x_{njk}$ is the level of impact $k$ for policy $j$, $\theta_{jm}$ is a parameter that captures the extent that the sociodemographic characteristic $m$ affects the preferences for policy $j$, and $\varepsilon_{nj}$ is an extreme-value type I stochastic term. The utility of non-spent resources is modelled in a similar form, by assuming $\Psi_{n0} = \exp(\delta_0 + \varepsilon_{n0})$. Dekker et al. (2019) provide an expression for the probability of choosing a bundle of policies under the MDCEV-PVE framework, allowing to estimate the model parameters using maximum likelihood.

The estimates of the MDCEV-PVE model can be used to determine the aggregate utility that a given bundle of policy options provides to society. Following Dekker et al. (2019) the aggregate utility of a given bundle of policies is given by:

$$EU = y_0 E\left[\Psi_{n0}\right] + \sum_{j=1}^{J} y_{nj} E\left[\Psi_{nj}\right], \tag{2.2}$$

where $E\left[\Psi_{nj}\right] = \Gamma(2) \cdot \exp\left(\hat{\delta}_j + \sum_k \hat{\beta}_k x_{jk}\right)$ and $E\left[\Psi_{n0}\right] = \Gamma(2) \cdot \exp\left(\hat{\delta}_0\right)$. It is assumed that all individuals in society face the same levels of policy impacts. Thus, only a single level for each policy impact $x_{jk}$ and $y_0$ are considered for the computation of the aggregate utility. In general, these values are assumed to be the average value of each impact level and cost, for each policy option, or either the minimum or maximum levels when a sensitivity analysis of the aggregate utility is performed.

The aggregate utility function can be used to determine the bundle of policy options that maximizes the aggregate utility of society, provided that a policymaker has limited resources. Dekker et al. (2019) suggest a procedure to determine the optimal bundle by enumerating the aggregate utility of all possible combinations of policy options that satisfy a given resource limit and sorting them in descending order. The bundle with the highest aggregate utility is called the "optimal portfolio" of policy options.

Finally, the participants produced more than 100,000 written motivations for the choices they made in the PVE. As the time between the start of our data collection and the publication of our results for Dutch policy makers was very limited (29 April – 6 May) we decided to analyse the written arguments of only a share of the respondents. We randomly selected 3,000 respondents and assigned the written arguments of these respondents to six annotators. To obtain an exhaustive list of arguments for and against each of the relaxation measures we asked the annotators to analyse these arguments until saturation was reached. One annotator experienced that saturation occurred after he had analysed the written motivations of 200 participants (no new arguments were added to the list of arguments), while another annotator had to review the responses of 500 participants to reach that point. The remaining annotators reached saturation between these two extremes. Eventually, the written arguments of 2,237 participants were analysed. In a second round of analysis, three annotators counted the number of times that 600 respondents mentioned the arguments that were identified in the first round. The aim of this was to provide policymakers with information about the number of respondents who cited a specific argument. For reasons of time we could only include 600 respondents in this second round.

## 2.5. Results

### 2.5.1. Descriptive results

The vast majority of participants supported a degree of relaxation of lockdown measures in the period 20 May – 20 July. We found little support for far-reaching relaxations that might cause the healthcare system to become heavily overloaded (higher than 41% increase in pressure on the health care system), but this varied across segments of the population. Figure 2.2 shows that men with high incomes and high education levels expressed a relatively strong preference for opening up (which would result in a relatively high pressure on the health care system). In contrast, older people on low incomes, who estimated that they themselves ran a high risk of becoming seriously ill from COVID-19, were relatively conservative in this regard. A further distinction is noticeable between the two survey groups. Participants in the panel PVE were significantly more cautious than participants in the open PVE in terms of their advice on relaxing lockdown measures. On average, participants in the 'panel PVE' recommended options resulting in a 28% increase in pressure on the healthcare system, while for those in the open PVE this was 32%. The percentage of participants advising against any relaxation whatsoever was much higher for the panel PVE than for the open PVE. This result suggests that citizens who participated in the open PVE were inclined to support a somewhat more extensive relaxation of lockdown measures than the average Dutch citizen (participants in the panel PVE).

Figure 2.3 shows that in both the open PVE and the panel PVE participants most often recommended the option: ”Re-open contact professions”. Figure 2.3 also shows that the strategy “Re-open hospitality and entertainment industry” was evaluated differently in the panel PVE and the open PVE. In the panel PVE 20% of the participants recommended this option and 45% discouraged this option, whilst in the open PVE the percentage of respondents who recommended this option was higher than the share of respondents opposing it. Moreover, Figure 2.3 shows participants divided about the desirability of the relaxation option 'nursing and care homes should allow visitors'.

One area of broad agreement was opposition to the relaxation of restrictions for specific groups of citizens. In both the panel PVE and the open PVE, the option “All restrictions lifted in Northern provinces” was least often advised, with “All restrictions lifted for people with immunity” not far behind. As seen in Figure 2.3, both options were rejected by more than 45% of the participants in the open PWE.

A normative objective in public participation is to secure distributional justice. The design of the PVE allowed citizens to consider the distributions of burdens and benefits of relaxing lockdown measures and enabled them to choose policy options from which they themselves would not benefit at all. To verify the extent to which participants choose relaxation policies that do (not) benefit themselves we asked them to indicate the impacts they predicted they would experience from each of the relaxation options they recommended. Table 2.1 shows that 71% of the respondents who recommended the relaxation option “Nursing and care homes allow visitors” would not personally experience any impacts from its implementation. 69% of the respondents would not expect to experience impacts from the relaxation option “Direct family members from other households can have social contact”. The written motivations (which we discuss more in detail in section 2.5.3) show that the interpretation of this result is ambiguous. On the one hand, there are respondents who choose this option for altruistic purposes. For instance, one respondent says: “I do not have any family, but I think that people

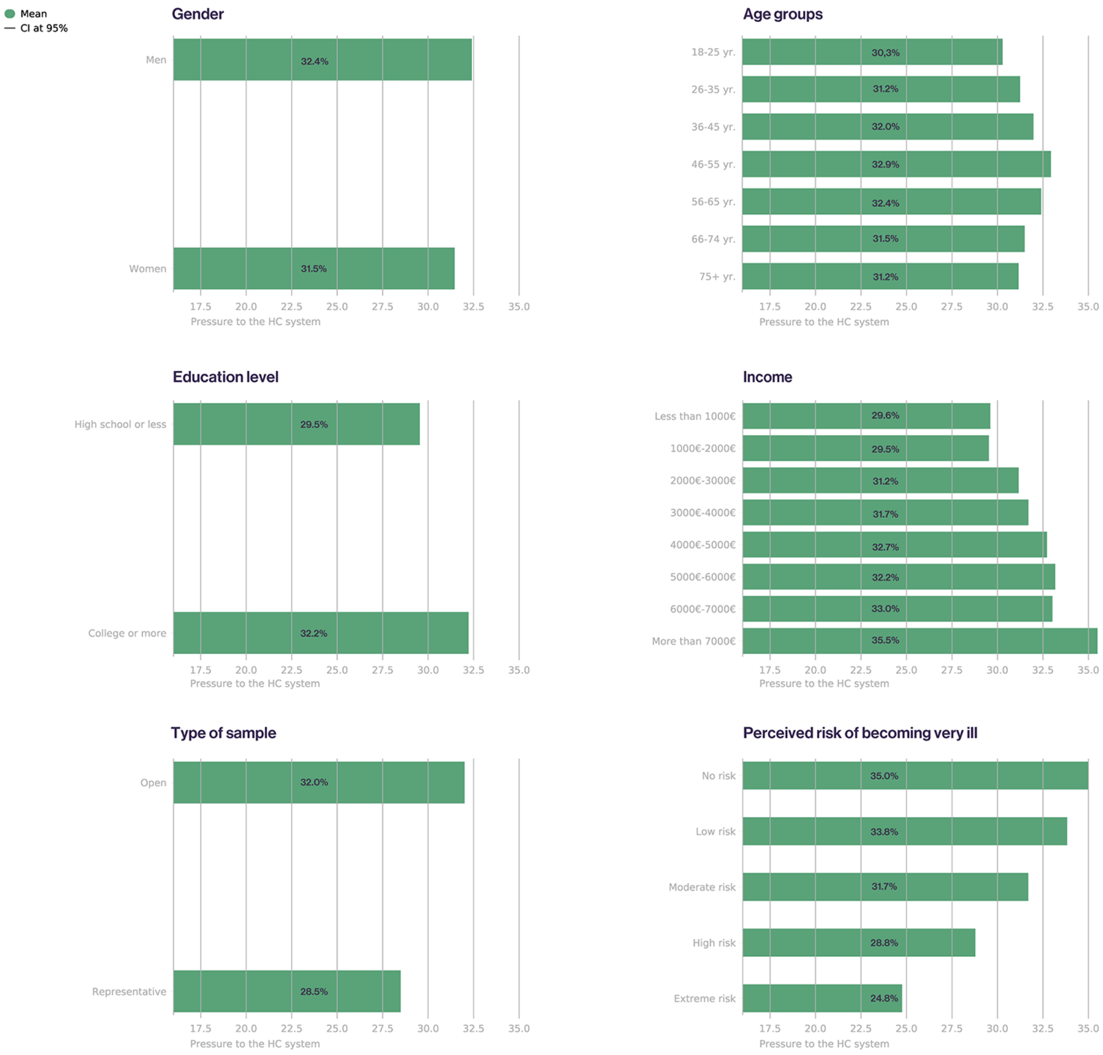


*Figure 2.2: Additional pressure of the health care system resulting from the recommended portfolio*

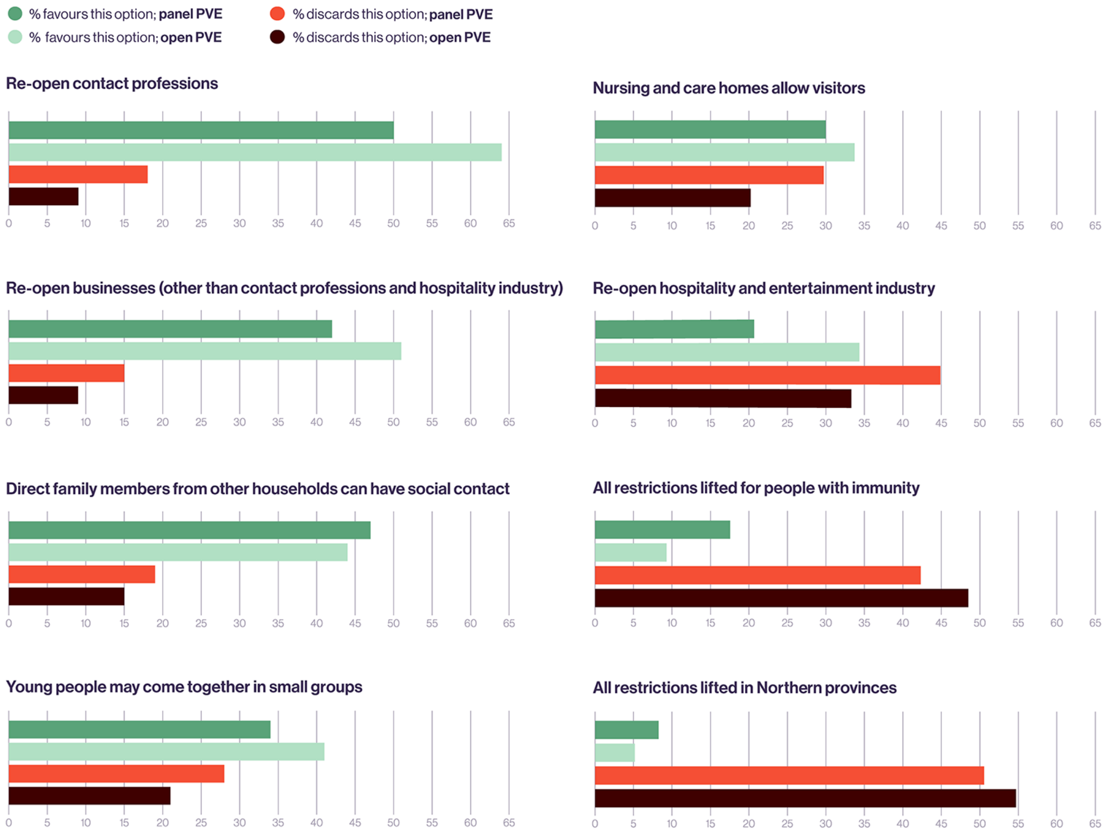


*Figure 2.3: Percentage of respondents who recommended or opposed the eight relaxation measures*

who do have a family look forward to hold their loved ones". On the other hand, many respondents said that the relaxation of this lockdown measure will not affect them as they already violated this rule.

| | No effect | Small effect | Medium effect | Large effect | Very large effect |
|---|---|---|---|---|---|
| Option 1<br>Nursing and care homes allow visitors | 71% | 13% | 6% | 6% | 4% |
| Option 2<br>Re-open businesses (other than contact professions and hospitality industry) | 14% | 27% | 28% | 21% | 10% |
| Option 3<br>Re-open contact professions | 6% | 30% | 36% | 20% | 8% |
| Option 4<br>Young people may come together in small groups | 9% | 24% | 29% | 24% | 14% |
| Option 5<br>All restrictions lifted for people with immunity | 45% | 20% | 15% | 11% | 9% |
| Option 6<br>All restrictions lifted in Northern provinces | 51% | 19% | 15% | 9% | 6% |
| Option 7<br>Direct family members from other households can have social contact | 69% | 13% | 7% | 5% | 6% |
| Option 8<br>Re-open hospitality and entertainment industry | 14% | 19% | 26% | 25% | 16% |

*Table 2.1: To what extent will lifting lockdown measures have an effect on your life?*

## 2.5.2. Quantitative results

This section presents the estimation results of the MDCEV-PVE model under two specifications. In the first specification, we estimate a simple model that accounts for the effects of impacts through taste parameters as well as policy-specific constants. The second specification includes sociodemographic variables for each relaxation option to uncover differences between different groups of individuals in terms of their preferences over certain relaxation options. We then provide the optimal portfolio of relaxation options for the first specification. All results provided in this section were calculated using the full available sample (i.e. combining responses from the open sample and the representative sample). Appendix 2.C provides the estimation results of the first specification of the MDCEV-PVE model for each sample separately.

### MDCEV-PVE model estimates

Table 2.2 summarises the MDCEV-PVE estimates for the model without sociodemographic variables, henceforth referred to as the “simple model”. The first set of estimates are the taste parameters. All estimates are statistically significant, except for the taste parameter associated with reductions in permanent mental injuries. The sign of the taste parameters indicates whether an increase in the associated impact makes a relaxation option more (un)attractive. Thus, any additional deaths and (permanent) physical injuries resulting from COVID-19 negatively impact the attractiveness of a relaxation option, while a reduction in the number of households experiencing income loss of greater than 15% increases that attractiveness. Using the taste parameters, it is also possible to establish the relative importance of the different impacts in defining the desirability of relaxation options. For instance, we can infer from the results that citizens consider a reduction of 100 deaths of persons below the age of 70 years and the reduction of 168 deaths of citizens older than 70 years (-0.8486 / -0.5084) equally attractive (in that they provide the same utility).

The second set of estimates correspond to the policy-specific constants. A higher value of these estimates reflects a stronger preference for the associated relaxation options irrespective of the impacts for which we estimated taste parameters.

Table 2.3 summarizes the estimates of an MDCEV-PVE model which includes a set of sociodemographic variables for each relaxation option. We included a variable to identify potential differences in the preferences of men and women, a variable to identify the extent to which the preferences of the youngest (19 to 25 years old) and oldest (above 65 years old) citizens differ from those in the middle age groups and a variable to analyse whether people with a high education level have different preferences than those with a lower education level. Finally, we analysed whether residents

| | Estimate | (Std. Err.) |
|---|---|---|
| Policy-specific constants: | | |
| 1: Nursing and care homes allow visitors | 2.6948$^{***}$ | (0.0273) |
| 2: Re-open businesses (other than contact professions and hospitality industry) | 2.6187$^{***}$ | (0.0208) |
| 3: Re-open contact professions | 3.1906$^{***}$ | (0.0243) |
| 4: Young people may come together in small groups | 1.8544$^{***}$ | (0.0127) |
| 5: All restrictions lifted for people with immunity | 1.6231$^{***}$ | (0.0200) |
| 6: All restrictions lifted in Northern provinces | 1.6617$^{***}$ | (0.0314) |
| 7: All restrictions lifted in Northern provinces | 2.5117$^{***}$ | (0.0278) |
| 8: Re-open hospitality and entertainment industry | 2.7032$^{***}$ | (0.0327) |
| Taste parameters: | | |
| Additional 10.000 deaths of people of +70 years | -0.5084$^{***}$ | (0.0802) |
| Additional 10.000 deaths of people of less than 70 years | -0.8486$^{***}$ | (0.1582) |
| Additional 10.000 people with permanent physical injury | -0.1082$^{***}$ | (0.0155) |
| Minus 10.000 people with permanent mental injury | 0.0006 | (0.0033) |
| Minus 10.000 households that have lost 15% of income | 0.0076$^{***}$ | (0.0022) |
| Observations | 29,651 | |
| Log-likelihood | -144,957.5115 | |
| AIC | 289,889.023 | |
| BIC | 289,781.1588 | |
| Statistical significance: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$ | | |

*Table 2.2: MDCEV model estimates*

of the Northern provinces have stronger preferences for lifting all restrictions in their own region.

Our results support the existence of varying preferences for relaxation options among different sociodemographic groups. We observe that the estimated parameters associated with sociodemographic variables are in general statistically significant. The sign of these parameters indicates whether individuals who belong to the sociodemographic group perceive the relaxation option as more (un)attractive.

We can illustrate this with a few examples from the results. In terms of gender differences, men perceive allowing visitors in nursing homes as less attractive than women do; at the same time, however, men are more positive about re-opening contact professions. With respect to age, people above 65 years old are most supportive of allowing visitors in nursing homes, while those between the ages of 19 and 25 are more receptive to a re-opening of the hospitality industry than are other age groups. In terms of education level, Dutch citizens with a higher level of education perceive re-opening the hospitality industry as more attractive than people with other educational backgrounds. Finally, residents of the Northern provinces perceive lifting restrictions in that region as more attractive than inhabitants of other provinces. One of the results that stands out is that the estimated parameters for the option "re-open contact professions" are consistently small regardless of socioeconomic grouping, while the policy-specific constant is the highest out of any option. This indicates a broad base of support throughout Dutch society. We also estimated an MDCEV-PVE simple model using a sample of residents of the Northern provinces and report the results in Appendix 2.D. Although citizens living in this region have a relatively positive view of the strategy which entails lifting the corona measures in the Northern provinces, this strategy is not included in the optimal portfolio.

| | Nursing homes | Businesses | Contact professions | Young people | People with immunity | Northern provinces | Direct family members | Hospitality industry |
|---|---|---|---|---|---|---|---|---|
| Parameters specific to each relaxation option | | | | | | | | |
| Constant | 2.8651*** | 2.1238*** | 3.0668*** | 1.6067*** | 1.6348*** | 1.5363*** | 2.6065*** | 2.4668*** |
| | (0.0384) | (0.0332) | (0.0332) | (0.0279) | (0.035) | (0.045) | (0.0385) | (0.0404) |
| Is Male | -0.4537*** | 0.4094*** | 0.1953*** | 0.0576* | 0.1414*** | 0.0884** | -0.1178*** | 0.2213*** |
| | (0.0233) | (0.0244) | (0.0263) | (0.0233) | (0.0259) | (0.0299) | (0.0236) | (0.0231) |
| Is above 65 years old | 0.4480*** | -0.0151 | -0.2021*** | 0.013 | 0.0552 | 0.2171*** | -0.0591 | -0.5659*** |
| | (0.0335) | (0.0348) | (0.037) | (0.0331) | (0.0358) | (0.0394) | (0.0339) | (0.0334) |
| Is between 19 and 25 years old | -0.3941*** | 0.01 | -0.2833*** | 0.4105*** | -0.1289** | -0.2468*** | -0.0612 | 0.1659*** |
| | (0.0399) | (0.041) | (0.0425) | (0.0409) | (0.0452) | (0.0558) | (0.0395) | (0.0391) |
| Has college degree (HBO or university) | 0.0872** | 0.4254*** | 0.1660*** | 0.2649*** | -0.1252*** | -0.0448 | -0.0321 | 0.2359*** |
| | (0.0271) | (0.0276) | (0.03) | (0.0271) | (0.0295) | (0.0339) | (0.0277) | (0.0269) |
| Lives in a Northern province | | | | | | 0.8371*** | | |
| | | | | | | (0.0418) | | |
| Taste parameters (common among all relaxation options) | | | | | | | | |
| Additional 10.000 deaths of people of +70 years | -0.5955*** | | | | | | | |
| | (0.0956) | | | | | | | |
| Additional 10.000 deaths of people of less than 70 years | -0.8803*** | | | | | | | |
| | (0.2022) | | | | | | | |
| Additional 10.000 people with permanent physical injury | -0.1148*** | | | | | | | |
| | (0.0165) | | | | | | | |
| Minus 10.000 people with permanent mental injury | 0.0034 | | | | | | | |
| | (0.0036) | | | | | | | |
| Minus 10.000 households that have lost 15% of income | 0.0091*** | | | | | | | |
| | (0.0024) | | | | | | | |
| Observations | 24,004 | | | | | | | |
| Log-likelihood | -115,791.0719 | | | | | | | |
| AIC | 231,490.1437 | | | | | | | |
| BIC | 231,118.1888 | | | | | | | |
| Statistical significance: $^{***}$p ¡ 0.001, $^{**}$p ¡ 0.01, $^{*}$p ¡ 0.05 | | | | | | | | |

*Table 2.3: Estimation results of MDCEV-PVE model (with covariates)*

### Optimal portfolios of relaxation options

Using the estimates of Table 2.2, we computed the optimal portfolio of relaxation options which respects the budget constraint of a maximum increase of the pressure to the healthcare system of 50%. This optimal portfolio is determined under the assumption that all individuals in society face the same impact levels and pressure on the healthcare system. We have taken these values from the average impact levels and pressure presented in the experiment (see Appendix 2.E). We include two additional scenarios for the purpose of sensitivity analysis. The first scenario is a pessimistic case, under the assumption that all individuals in society face the maximum levels of pressure to the healthcare system, the maximum levels of the impacts that have a negative taste parameter estimate, and the minimum levels of the impacts that have a positive taste parameter estimate. The second scenario is an optimistic case, in which all individuals in society face the minimum levels of pressure, the minimum levels of impacts with negative taste parameter estimates, and the maximum levels of impacts with positive taste parameter estimates. More information on the impact levels and pressure used to compute the optimal portfolios in these sensitivity analyses can also be found in Appendix 2.E.

Table 2.4 lists the optimal portfolio under each of the three scenarios. The optimal portfolio given an average-level scenario suggests that the most preferred bundle of relaxation options is to re-open contact professions re-open businesses (except the hospitality industry) and to allow social contact again between families. This bundle imposes an increase of the pressure to the healthcare system of 32%, still leaving a substantial amount of pressure without allocation. Under a pessimistic scenario, only allowing contact professions to re-open is included in the optimal portfolio, with a pressure to the healthcare system of 15%. Under an optimistic scenario, five out of eight relaxation options are part of the optimal portfolio, excluding re-opening the hospitality industry, lifting restrictions for individuals with immunity and lifting restrictions for the Northern provinces. Such bundle of relaxation policies results in an increase in the pressure to the healthcare system of 34%.

Section 2.5.2 revealed that highly educated respondents were over-represented in the sample. One of the strengths of PVE is that it is possible to control for this in the evaluation step through applying corrective weights (Mouter et al., 2021a). Appendix 2.F provides a description of this procedure and the computation of the corrected optimal portfolio. We found no differences between the optimal portfolio presented in table 2.4 and the corrected optimal portfolio.

| | Average | Pessimistic | Optimistic |
|---|---|---|---|
| 1: Nursing and care homes allow visitors | | | X |
| 2: Re-open businesses (other than contact professions and hospitality industry) | X | | X |
| 3: Re-open contact professions | X | X | X |
| 4: Young people may come together in small groups | | | X |
| 5: All restrictions lifted for people with immunity | | | |
| 6: All restrictions lifted in Northern provinces | | | |
| 7: All restrictions lifted in Northern provinces | X | | X |
| 8: Re-open hospitality and entertainment industry | | | |
| Added pressure onto the healthcare system | 32% | 15% | 34% |

*Table 2.4: Optimal portfolios of relaxation options.*

## 2.5.3. Qualitative results

Analysis of the written motivations of 2,237 randomly selected participants on why they preferred some relaxation options over others revealed four insights. First, it shed light on the arguments and concerns of proponents and opponents of each option; many of these had not come to our attention during our analysis of media content and the conversations that we had with policymakers when designing the PVE. We will summarize all arguments in Tables 6-13 and also show how many respondents of the 600 respondents from whom written motivations were analysed in the second round cited a certain argument. The second insight relates to the underlying principles that are at stake in relaxing lockdown measures. For example, participants consider it important that the relaxation leads to "unity" rather than "division". These principles seem to play a large role in the explanation of why certain relaxation options are not favoured by Dutch citizens (e.g. lifting restrictions for the Northern provinces or for Dutch people who are immune to COVID-19). The third insight relates to how Dutch citizens condition their preferences. Without being specifically asked, a large number of participants conditioned their relaxation preferences to, amongst other things, increased safety measures. These conditions also revealed ideas, how to solve dilemmas of relaxation options and improve the effectiveness of relaxation options. Finally, the fourth insight was hearing explicitly from participants that they had evaluated relaxation options in relation to each other. This supports the use of PVE as a preference elicitation technique over alternatives such as CV and DCE, as it is a key advantage of the former.

### Nursing and care homes allow visitors

Many participants who recommended this option argue that the quality of life of older people and those in their final stages of life is more important than increasing their life expectancy. In sections 2.5.1. and 2.5.2 we already showed that participants were divided about the attractiveness of this relaxation option and the written motivations also reflect strong differences in opinion among respondents regarding the desirability of this strategy. On the one side, many respondents refer to fundamental rights when arguing that inhabitants of nursing and care homes should be able to decide for themselves whether they want to allow visitors. On the other hand, some respondents who disfavour this option argue that old and vulnerable people should be shielded from the rest of society to ensure that the rest of the country can go back to normal. Moreover, various respondents argued that relaxation options that positively impact the economy should be prioritised, which suggests that they evaluated this relaxation option in relation to the others.

| | # respondents out of the quotes of 600 respondents analysed in the second round |
|---|---|
| *Arguments for* | |
| The risk of catching the coronavirus does not outweigh the risk of loneliness or dying alone | 77 |
| Elderly people in nursing and care homes are very much in need of visitors and social contacts | 73 |
| Being able to decide about whether or not family can visit is a fundamental right that should not be violated | 17 |
| Not being allowed to visit is also traumatic for family members | 13 |
| Lifting this measure is advantageous for healthcare personnel, because it enables extra care from visitors and creates a better atmosphere | 12 |
| These people are generally not hospitalised so it does not put that much pressure on ICU | 1 |
| *Arguments against* | |
| Allowing visits leads to more infections | 58 |
| Vulnerable people should be shielded from the rest of society to ensure that the rest of the country can go back to normal | 10 |
| This also endangers the health of others, not just residents | 7 |
| Relaxing measures that are good for the economy should be prioritised | 3 |
| *Conditions* | |
| Ensure that there is enough protective material | 16 |
| Implement tailor-made measures, such as splitting nursing and care homes into sections with and without visitors | 14 |

*Table 2.5: Nursing and care homes allow visitors: arguments for, arguments against and conditions*

### Re-open businesses other than contact professions and hospitality industry

Many participants indicated that they selected this option because of the benefits for the economy, which is an argument that was anticipated a priori based on our conversations with policymakers. Nevertheless, the relatively large number of people who revealed generally positive attitudes over working from home was quite surprising. In meetings with policy makers and the media we did not encounter the argument that opening-up the economy in developed countries will have positive impacts on people living in developing countries.

| | # respondents out of the quotes of 600 respondents analysed in the second round |
|---|---|
| *Arguments for* | |
| This option prevents substantial damage to the economy | 187 |
| Being able to work again has a positive effect on people's well-being and mental health | 59 |
| The impact on the number of infections will not be high | 13 |
| If we don't get the economy out of the muck quickly, we won't be able to pay for our expensive health care in the future. The money that is needed to finance the health care sector needs to be earned somewhere | 9 |
| When developed countries close their economy this will amplify poverty in developing countries | 1 |
| *Arguments against* | |
| This measure substantially increases the risk of infections as businesses bring large groups of people together and will also result in greater movement of persons throughout the Netherlands | 27 |
| Working from home is not so bad | 16 |
| *Conditions* | |
| Only when social distancing and/or isolated workplaces can be guaranteed at the office | 67 |
| There should also be an option for high-risk individuals to work from home | 22 |

*Table 2.6: Re-open businesses other than contact professions and hospitality industry: arguments for, arguments against and conditions*

### Re-open contact professions

This relaxation option was the most chosen option by respondents and section 2.5.2 shows that there was also widespread support for this option among participants from different socio-demographic groups. Table 2.7 shows that participants provided a range of arguments as to why this option should be prioritised by the government. From the written motivations it could be inferred that many participants in the PVE sympathised with preventing the bankruptcy of a large number of (generally small) businesses. Another argument cited was that contact professions should re-open because, unlike the hospitality industry, they lack alternative sources of income. This was an argument that was not raised in the media content that we analysed, nor in the conversations that we had with policymakers in the design stage of the PVE. The fact that respondents explicitly made a comparison with the circumstances in the hospitality industry provides evidence that participants valued relaxation options in relation to each other rather than separately. Various respondents argued that contact professions with a medical purpose should be prioritised.

| | # respondents out of the quotes of 600 respondents analysed in the second round |
|---|---|
| *Arguments for* | |
| Prevents the bankruptcy of large number of small companies/entrepreneurs | 159 |
| Contact professions often have a medical care function. Hence, this relaxation option is good for (medical and psychological) health and for the economy | 84 |
| It is good to start with this relaxation option. Risks are low. If this goes well, the government can relax other lockdown measures | 25 |
| People working in these professions are trained to take care of hygiene and protect themselves and their clients | 21 |
| Appearance is important for people's well-being | 10 |
| These are often professions in which you cannot easily work from home | 8 |
| This relaxation option will increase support for the continuation of the other measures | 8 |
| If you do not allow the contact professions to go back to work, there is a chance that they begin working in secret, which entails higher risks | 7 |
| For these (small) entrepreneurs it is almost impossible to come up with an alternative business model (this is to some extent possible for the hospitality industry) | 3 |
| *Arguments against* | |
| Relatively high risk of infections because people that work in contact professions help many people each day and they are in contact with a client over a relatively long period of time | 20 |
| It is not essential/necessary | 3 |
| *Conditions* | |
| Sufficient protective material | 90 |
| Contact professions with a medical function (e.g. osteopaths) should be given priority over contact professions without a medical function (e.g. tattoos) | 13 |
| Opening hours should be widened to ensure the spreading of customers | 7 |
| Provide additional protection for personnel belonging to high-risk groups. The government should provide financial support | 5 |

*Table 2.7: Re-open contact professions: arguments for, arguments against and conditions*

### Young people may come together again in small groups

Under this option, young people would still be required to respect the 1.5-metre distance rule when they meet older people. Supporters cited its relatively small effect on the spread of the virus, the low risk for young people and its positive effects for young people, while detractors saw problems around enforcement of this rule and its being seen as a form of age discrimination.

| | # respondents out of the quotes of 600 respondents analysed in the second round |
|---|---|
| *Arguments for* | |
| Young people play a minor role in the spread of the virus and their risk of getting sick is low | 136 |
| Social contact is relatively important for young people (to develop themselves) | 61 |
| For young people it is difficult not to violate the rules | 49 |
| Reduction of problematic psychological symptoms | 18 |
| Reduces the pressure on parents | 17 |
| Possibility to build up herd immunity | 10 |
| Increases support among young people for other lockdown measures | 5 |
| *Arguments against* | |
| Constitutes age discrimination which results in a dichotomy in society | 27 |
| Measures are difficult to enforce. Young people will also get in contact with other people | 23 |
| *Conditions* | |
| Young people should maintain 1.5m distance from those outside that group | 20 |

*Table 2.8: Young people may come together again in small groups: arguments for, arguments against and conditions*

### All restrictions lifted in the Northern provinces

Sections 2.5.1 and 2.5.2 reveal that there is little support among Dutch citizens for policy options that relax restrictions for one specific group of citizens. Many participants find it very important that the relaxation of lockdown measures leads to “unity” and not to “division”. They are afraid that the unity among Dutch people that currently exists – along with the support for corona-related government policies – will be lost if and when the Cabinet chooses to lift restrictions for a specific group of Dutch people (e.g. the North of the Netherlands, Dutch people who are immune to COVID-19). Below are several quotes that illustrate this point:

> *“By making a distinction between people who are immune and people who may still be infected, you create a very strange dividing line between two groups in the population. The same with all the restrictions lifted in the Northern provinces. It’s either the whole of the Netherlands without restrictions, or not. Making divisions between occupations or parts of daily life (such as hospitality vs. contact professions) to lift restrictions is about smaller steps and is easier to understand than exempting a whole part of the Netherlands”.*
>
> *“We have to overcome this crisis together, so it is not wise to create divisions”.*
>
> *“There should be no difference between people. We live in one country and all have to follow the same rules. We are all Dutch and that means equal treatment”.*

> *“We are a country of 17 million people, who should be treated equally. We fight for equality and against racism so you should not make a distinction between people that live in different parts of the country.”*

| | # respondents out of the quotes of 600 respondents analysed in the second round |
|---|---|
| *Arguments for* | |
| Low risk of transmission in these provinces. The impact of relaxation measures can be monitored relatively easily | 15 |
| Impact of relaxing lockdown measures can be monitored and this provides useful information for future decisions on relaxing lockdown measures | 9 |
| Boosts the economy in the North of the Netherlands | 7 |
| *Arguments against* | |
| Practically unfeasible because this is almost impossible to enforce. People will go to the North for entertainment and bring infections to these provinces | 122 |
| Solidarity will be undermined and this will not benefit the Netherlands as a whole | 113 |
| *Conditions* | |
| Enforceability of this measure should be guaranteed | 3 |
| Measures should be relaxed in small steps | 1 |

*Table 2.9: All restrictions lifted in the Northern provinces: arguments for, arguments against and conditions*

### All restrictions lifted for people with immunity

The relaxation option ”For people who are immune, all restrictions are lifted” can also count on little support from the Dutch population. Table 2.10 shows that people have various concerns about this option and also cite that they oppose this option because it might lead to a dichotomy in society.

### Direct family members from other households can have social contact

Some of the written motivations provided by respondents who advised this relaxation option were new and unexpected. For instance, various respondents argued that they selected this option because, in their view, this will increase compliance with other lockdown measures as it provides positive energy and optimism. Moreover, it is noteworthy that many respondents supported this option because, in their view, many people (sometimes including respondents themselves) already violated this lockdown measure. On the other hand, various participants disfavoured this option as they argued that, for them, seeing friends was more important than social contact with family.

| | # respondents out of the quotes of 600 respondents analysed in the second round |
|---|---|
| *Arguments for* | |
| These people pose no danger to their environment | 16 |
| These people can keep society and the economy going again | 10 |
| It is pointless to demand solidarity from these people if they are already immune. Doing so will lead to fierce protests | 9 |
| *Arguments against* | |
| Tests for immunity are not foolproof, and this increases the risk of new infections | 121 |
| Creates a dichotomy in society. People who are not immune can get annoyed by the behaviour of those who are allowed to resume normal life | 70 |
| Difficult to enforce | 60 |
| Potential confusion as immunity is not outwardly apparent | 18 |
| *Conditions* | |
| Only consider this option when you are 100% sure that immunity can be measured | 1 |

*Table 2.10: All restrictions lifted for people with immunity: arguments for, arguments against and conditions*

| | # respondents out of the quotes of 600 respondents analysed in the second round |
|---|---|
| *Arguments for* | |
| Improves the well-being of many Dutch people. Contact with family is important in times of crisis, and can alleviate psychological harm. Hence, in the longer term, this can reduce the need for mental care caused by psychological distress | 123 |
| People will behave responsibly to ensure that they do not infect their loved ones. Family members keep each other informed about their health | 46 |
| People already violate this rule so this relaxation option brings the rules more in sync with reality | 41 |
| This allows contact with only a small number of people ('social bubble') which has a relatively small impact on the risk of large-scale transmission of COVID-19 | 26 |
| This relaxation option ensures that citizens will comply with the lockdown measures. It provides positive energy and optimism over the future | 8 |
| Grandparents can take care of their grandchildren which reduces pressure on families | 5 |
| *Arguments against* | |
| This substantially increases the risk of infections | 16 |
| This measure is difficult to enforce | 10.0 |
| Focusing only on (direct) family is too limited. My friends are more important to me than my family | 7 |
| Measures that have an impact on the economy should be prioritised | 3 |
| *Conditions* | |
| Ensure that this rule is only applicable to direct family members | 6 |

*Table 2.11: Direct family members from other households can have social contact: arguments for, arguments against and conditions*

### Re-open hospitality and entertainment industry

Participants argued that opening up hospitality and entertainment is not only good for the economy and business, but they also considered it important for the well-being of the Dutch. That said, many participants were also concerned that this relaxation option would result in increased infections, particularly in situations where the consumption of alcohol had the potential to change perceived risks for individuals. Many participants argued that this relaxation option is less urgent than other options. For instance, one respondent argued that "nursing and care homes allow visitors" should be prioritised because the situation in nursing and care homes is much more poignant. Finally, some participants argued that the risk that this relaxation option contributes to new outbreaks of COVID-19 is relatively high and for this reason they think that this option should only be considered after other options had turned out to be successful.

| | # respondents out of the quotes of 600 respondents analysed in the second round |
|---|---|
| *Arguments for* | |
| This is good for our economy and business | 106 |
| It is good for people's well-being | 83 |
| This relaxation option will increase support for the continuation of the other measures | 7 |
| It is enforceable | 7 |
| People can take responsibility for themselves by staying away if they wish | 7 |
| We should preserve our cultural heritage and cannot risk bankruptcies in the cultural sector | 4 |
| Keeping these businesses closed is too big of a sacrifice for young people | 3 |
| In this way, we can build up herd immunity | 1 |
| If the hospitality industry is not re-opened people will do other things to relax which is also risky | 1 |
| *Arguments against* | |
| Risk of too many people gathering together, which helps to spread the virus | 83 |
| It is not necessary at the moment | 22 |
| When alcohol is consumed, people are more likely to underestimate risks and are less likely to comply with distancing measures | 11 |
| Opening up the hospitality and entertainment sectors should only be considered in the next phase if it appears that other adjustments have worked | 10 |
| Hospitality industry has a bad impact on society. Please keep it closed | 1 |
| *Conditions* | |
| There are many options for measures to be taken in hospitality and entertainment (including reducing alcohol consumption). Rely on the sector's creativity and sense of responsibility. | 40 |
| It is important to differentiate between different sectors (e.g. bars closed, museums open) | 14 |
| Re-open hospitality industry but restrict opening hours | 1 |

*Table 2.12: Re-open hospitality and entertainment industry: arguments for, arguments against and conditions*

### 2.5.4. The merits of the PVE as perceived by participants

The draft results of our study were shared on 4 May with the Ministry of Health and the Dutch National Institute for Public Health. The latter, in turn, chairs the central Outbreak Management Team which advises the government on COVID-19 policies. The final results were shared on 6 May. As their involvement and collaboration in the research showed, those experts were open to and cognisant of concerns and priorities from the public. We do not know whether and how our results affected political decisions on the relaxation of lockdown measures, but it is noteworthy that the Dutch government decided on 6 May to start with the relaxation of lockdown measures for contact professions which was in line with our result that re-opening contact professions would have broad support in society. Another example of the way that political decisions overlapped with our results is that the Dutch government, unlike other countries such as Germany, adopted a central approach in terms of imposing and relaxing lockdown measures without differentiating between regions.

In section 2.3, we proposed several hypotheses regarding the strengths of PVE. These related to enabling citizens to participate in multi-dimensional policy issues (normative rationale for participation) and letting citizens experience intricate government dilemmas so as to improve their understanding of relevant trade-offs and potentially improve future compliance (instrumental rationale for participation). Moreover, we discussed that a potential weakness of PVE is that the quality of preferences that people express is probably lower than preferences that they express after deliberation (which is where mini-publics have an advantage).

To explore the extent to which the hypothesised strengths and weaknesses were actually realised, we evaluated how participants experienced their participation in the PVE through asking them to respond to several propositions (see Figure 2.4) and we asked open questions to reflect on the strengths and weaknesses of the method. Table 2.13 provides an overview of the number of respondents that cited a certain strength out of the 600 respondents for whom written motivations were analysed in the second round of analysis.

A first perceived strength of PVE is that putting large groups of people in the shoes of a policymaker might raise their awareness of intricate government dilemmas. Figure 2.4 shows that around 60% of the participants felt that they learned more about the choices the government needed to make regarding the relaxation of lockdown measures through participating in the PVE, whereas around 20% disagreed with this proposition. Table 2.13 shows that awareness-raising about the consequences of relaxation options and the dilemmas the government faces was also cited by many participants as a strength of the method. Below, we list illustrative quotes of respondents that were

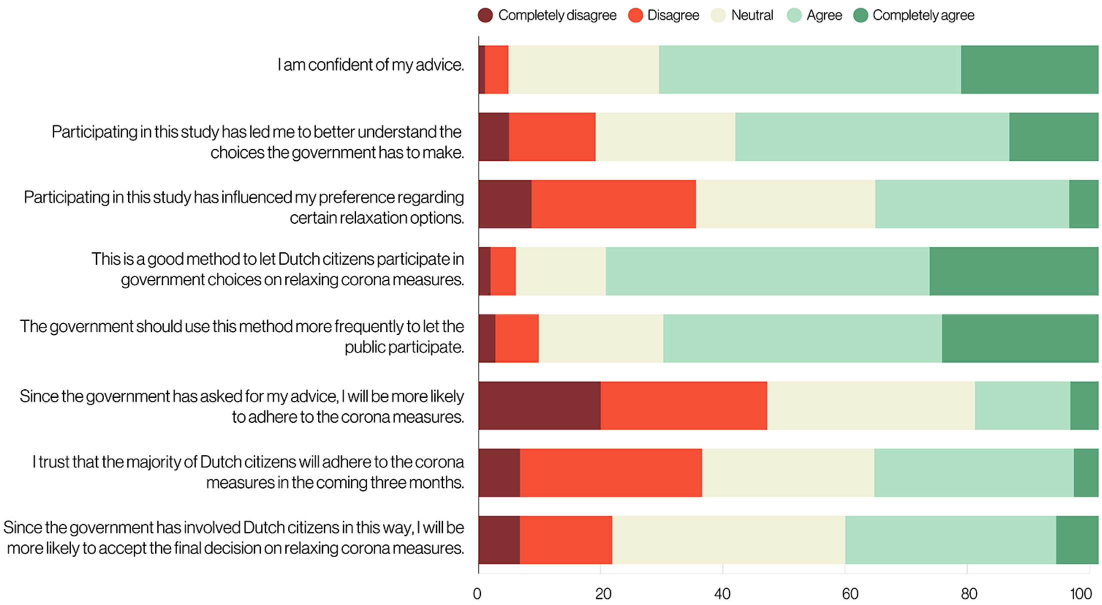


*Figure 2.4: Experiences of participants and their likelihood of adherence/acceptance of measures*

| Perceived strength | # respondents out of the quotes of 600 respondents analysed in the second round |
|---|---|
| The survey was very clear (clear instruction video and background information) | 88 |
| *Substantive rationale for participation* | |
| This is an informed advice to the government based on insights regarding the consequences of your advice | 76 |
| Provides lot’s opportunities to explain my advises and to add nuances | 49 |
| The constraint forces participants to make a choice (not possible to just choose everything) | 10 |
| The government gets an impression of citizens’ preferences regarding this topic | 4 |
| *Normative rationale for participation* | |
| Positive that the government consults its citizens | 52 |
| I had the feeling that my opinion counted | 4 |
| Positive that the consultation was accessible for all citizens. | 2 |
| Allowed me to provide a contribution to fighting the COVID-19 crisis | 1 |
| *Instrumental rationale for participation* | |
| Raised my awareness regarding (consequences of) relaxation options | 77 |
| Improves transparency regarding the dilemmas the government faces | 34 |
| Encourages me to reflect on my own opinions | 7 |
| Improves understanding and support for final decisions on relaxation of lockdown measures | 5 |

*Table 2.13: Number of times that perceived strengths of the PVE method were cited*

positive about the awareness-raising ability of PVE:

> *This gives me a better understanding of the choice that politicians and policymakers face.*
>
> *"Everything was well-explained. The people that designed this research succeeded in showing that this is a choice with multiple dimensions instead of a simple choice. Great achievement that such a research is designed in such a small amount of time. It provides you as a participant with insights into the complexity of government choices."*
>
> *"I liked how you get insight into the consequences of relaxation options and the way that decisions on relaxing lockdown measures are interrelated."*
>
> *"This study increases the transparency of the trade-off that the government faces. Participants are also confronted with the consequences of their advices."*
>
> *"It made me think of how difficult these kinds of dilemmas are."*
>
> *"You experience the responsibility that people in government also experience."*

Ideally, improved awareness improves the extent to which participants accept the final decision of the government and comply with government measures. To check this we asked respondents to evaluate the proposition: "Because the government involves me in this way, I am better able to accept the final decision of the government regarding the relaxation of lockdown measures between 20 May and 20 July." 40% of the respondents agreed with this proposition and slightly more than 20% disagreed with it. Only a few respondents explicitly cited this as a strength of PVE. We also asked respondents on their opinion regarding the proposition: "since the government has asked for my advice, I will be more likely to adhere to the corona measures". Our results show that only 18% thought that participating in a PVE would increase their compliance with lockdown measures.

Another potential benefit of PVE is that it provides an opportunity for participants to advise their government after experiencing a dilemma faced by policymakers. For reasons of limited space in the survey, we were not able to include a proposition which specifically asks participants how they perceived this specific characteristic of the PVE, but we asked them to respond to two propositions: "PVE is a good method for involving citizens in government decisions concerning the relaxation of government measures between 20 May and 20 July" and "The government should use this method more often for involving citizens in policymaking." Around 80% of the respondents agreed with the proposition that PVE is a good method for involving citizens regarding this topic and 75% said that the government should use this method more often. Less-educated Dutch citizens are slightly more positive about the method than their highly educated counterparts.

Various participants cited some characteristics of PVE to explain why they thought it was a good method to transmit preferences of citizens to the government. Participants liked the fact that citizens were asked to provide advice based on insights regarding the consequences and that they were forced to make a choice between relaxation options. Moreover, participants liked that there was ample room to add nuances. Below, we provide illustrative quotes.

> *"This setting allows participants to digest information about the consequences of government policies before they provide an advice. As a result, the outcomes are much more useful for government decision-makers than the preferences that people express on Facebook and Twitter."*
>
> *"You see the consequences of your advices. It is not a simple yes or no question without seeing the consequences like with the hopeless and useless idea of a referendum."*
>
> *"It is really good that people are asked to explain their choices because this ensures that people do not get away with pressing a few buttons based on their gut feeling."*
>
> *"The opportunity to provide written explanations. This allows you to express the nuances of your opinion that you cannot express with only making some choices between relaxation options."*

Respondents also said that they liked that the PVE demonstrated that the government was open to the ideas of citizens.

> *"I also like the fact that the government is open to the (good) ideas of its citizens. Thank you very much!"*
>
> *"Nice way to involve people more directly in politics."*
>
> *"This allows people to communicate their concerns and worries. Now they use social media for this purpose, but I think it is very important and really useful to have a more formal place where people can blow off steam in a more productive way."*

Only a handful of the 600 respondents mentioned as a strength that participating in the PVE gave them the feeling that their opinion counted. We also asked participants what weight politicians should assign to the outcomes of the PVE alongside the advice that politicians received from health experts. A minority of the participants (5%) thought that the advice given by citizens in the PVE should have a heavier weighting in the government's decision-making than the advice given by experts. Conversely, 69% of participants opined that the expert advice should weigh heavier. The remaining 28% felt that the government should give both types of advice equal weighting. We think that it is interesting that citizens who participated in the PVE – who must have an above-average interest in participating in government decision-making – believe that more weight should be given to scientific advice than to the advice of citizens.

One potential downside of PVE is that the quality of preferences that people express is probably lower than preferences that people express after deliberation (such as is the case in mini-publics). It is, of course, difficult to directly verify the quality of the preferences of respondents, but as a surrogate, we asked respondents whether they were convinced of their advice. More than 70% of them responded positively to this proposition. Moreover, we asked respondents whether they changed their opinion due to participating in the PVE (about a third of the participants said that this was the case). In addition, respondents were asked to mention weaknesses of the method (or aspects that can be improved) and we only found one argument among the written answers of the 600 respondents we analysed which referred to limitations in terms of the ability to transmit preferences to the government via a PVE. A handful of participants criticised the fact that they could only make a distinction between different subsector (e.g. bars should be closed, but museums should be opened) in the written motivations and not in the primary choice tasks of the PVE. Other weaknesses were mentioned by a larger number of respondents: not possible to conduct the experiment via a smartphone, the profiles of relaxation options varied across respondents on their impact levels and pressure to the healthcare system (see section 2.5.2) and some respondents found this suspicious, some respondents found the survey too complex and, finally, respondents argued that the research team should bring the experiment under the attention by more people via advertisement to ensure that more people participate.

## 2.6. Conclusion and discussion

This paper reports about an attempt that was made in the Netherlands to involve about 30,000 Dutch citizens in policy decisions regarding relaxing lockdown measures between 20 May and 20 July 2020 through a Participatory Value Evaluation (PVE).

Participants in the PVE were presented with eight possibilities for relaxing lockdown measures for this period, out of which they could make recommendations to the government. For each of these relaxation options, they received information regarding the option's societal impact (e.g., increase in pressure on the health care system, an increase in deaths among people younger than 70 years and a decrease in the number of households with a long-term loss of income). The constraint that participants faced in the PVE was the maximum capacity of the healthcare system. They were not able to recommend a bundle of relaxation options that resulted in a greater than 50% increase in the pressure on the healthcare system. Subsequently, participants were asked which of the eight relaxation options should not be considered by the government. We carried out the PVE with two different samples. First, a random selection of 3,358 Dutch adults, who were selected with a view to be representative for the Dutch population of 18 years and older. Second, we opened the PVE for the general public, which resulted in more than 26,000 participants within six days. The primary goal of this paper is to show what sorts of insights a PVE can provide to policymakers and other stakeholders who have to decide on COVID-19 policies. A secondary objective of this paper is to improve understanding towards the strengths and weaknesses of PVE in terms of involving citizens in crisis policymaking.

### 2.6.1. Main findings

Our results show that the majority of the participants in the PVE advised the government to relax lockdown measures, but not to the point at which the healthcare system becomes heavily overloaded. Participants in the 'open PVE' were inclined to support a somewhat more extensive relaxation of lockdown measures than the average Dutch citizen (participants in the panel PVE). From the choices respondents made in the PVE, we were able to infer the implicit trade-offs made by Dutch citizens between impacts of relaxation options. For instance, we find that a reduction of 100 deaths of persons below the age of 70 years and the reduction of 168 deaths of citizens older than 70 years are equally attractive. There is wide support among participants for re-opening contact professions and our results show that this option is popular in all segments of Dutch society. Conversely, we found little support for policy options that would relax restrictions for one specific group of citizens. The options "All restrictions lifted in Northern provinces" and "All restrictions lifted for people with immunity" can count on little support among the Dutch population at large. The low support for the option "All restrictions lifted in Northern provinces" is at odds with the message of a number of scientists who advocated this option in the weeks before we conducted the PVE (Klaassen, 2020). These scholars considered this a promising approach, since at the time that the PVE was conducted there were only a few infections in these provinces; this made it easier to keep infection levels low through testing and tracing. Participants had a negative stance towards these relaxation options because they found it very important that the relaxation of lockdown measures leads to "unity" and not to "division". They are afraid that the unity among Dutch people that existed at the time that we conducted the PVE – along with the support for corona-related government policies – would be lost if and when the Cabinet chooses to lift restrictions for a specific group of Dutch people. The importance of equal treatment is also identified in studies which examined Dutch citizens preferences regarding health policies before the outbreak of the coronavirus (Reckers-Droog et al., 2018; Wouters et al., 2017). However, a clear contribution of our study is that Dutch citizens seem to think that it is unfair to distinguish policies between different regions, age groups and people who are (not) immune to COVID-19 – various respondents even labelled this as 'discrimination'– whereas we did not identified any respondents who explicitly said that making distinctions between different sectors (contact professions, hospitality industry and other business) would be 'unfair'. Another result that stands out is that 71% of the respondents who recommended the relaxation option "Nursing and care homes allow visitors" say that they will not experience any impacts from the implementation of this option. This suggests that involving large numbers of citizens in determining crisis policies might also increase empathy between individuals and foster an exchange of perspectives regarding ethical trade-offs (Aldrich & Meyer, 2015).

The choices made by participants in the PVE can be used as input for behaviourally-informed choice models which analyse people's preferences for (the impacts of) relaxation policies. These preferences can, in turn, be used to rank options in terms of their desirability. We find that citizens consider a reduction of 100 deaths of persons below the age of 70 years to be equally attractive as a reduction of 168 deaths of citizens older than 70 years. We find that the optimal portfolio of relaxation policies consists of three strategies: re-open contact professions, re-open businesses (except the hospitality industry) and allow social contact between direct family members. An advantage of PVE is that sensitivity analyses can be conducted to explore how the desirability of policy options is affected by changes in impacts. These sensitivity analyses show that in a pessimistic scenario only re-opening contact professions is included in the optimal portfolio. In an optimistic scenario five out of eight relaxation policies are part of the optimal portfolio, excluding re-opening the hospitality industry, lifting restrictions for those with immunity and lifting restrictions for the Northern provinces.

In this paper, we listed various reasons why PVE could be an appealing participatory approach for involving citizens in policy decisions during a pandemic: 1) citizens can participate in a PVE online, which is appealing in times of social distancing; 2) a

PVE can be deployed rapidly, which is important during a pandemic as governments have to respond quickly to new developments; 3) the design of a PVE can adopt other constraints than only public budget; 4) PVE provides information to policymakers about the extent to which the desirability of policy options is affected by the impacts of the policy options; 5) PVE allows citizens to transmit new ideas, arguments, values and conditions to decision-makers; 6) PVE enables citizens to participate in multi-dimensional policy issues that do not lend themselves to a simple ‘yes’ or ‘no’ or the allocation of a constrained amount of scarce public resources; 7) PVE lets citizens experience intricate government dilemmas, increasing their understanding of the impacts of proposed measures and potentially increasing levels of acceptance and compliance.

In this paper, we establish that the first five potential benefits of the method were realised in this PVE. Citizens could participate online and the PVE was deployed rapidly (design process started 9 April, 2020 and results were shared with policymakers 6 May, 2020). The PVE adopted another constraint than the public budget (maximum pressure on the health care system), we showed that the PVE provided information to policymakers about the extent to which the desirability of policy options is affected by the impacts of the policy options and the PVE allowed citizens to transmit new ideas, arguments, values and conditions to policymakers. Policymakers can embed these new ideas, arguments, values and conditions in their policies, and the quantitative results produced by the PVE – such as the ranking of relaxation options – can inform their prioritisations. Moreover, the outcomes of the PVE provides policymakers with information about the effectiveness of existing policies. For instance, many respondents said that they themselves (or other people) were already violating the rule that family members from another household cannot have social contact.

We think that we can safely conclude that we partially realised the sixth and seventh appealing characteristics of PVE. Almost 60% of respondents said that they became more aware of the consequences of relaxation options and the dilemmas the government faces (instrumental rationale for participation). Almost 80% of participants stated that PVE is a good method to let citizens participate in government decision-making on lifting lockdown measures. Participants liked the fact that they were asked to provide advice while evaluating relaxation options in relation to each other and being informed about the consequences of the options. Participants also appreciated that they were forced to make a choice between relaxation options and that there was ample room to add nuances. That said, our results do not show convincingly that respondents would also comply to a higher level with public health measures simply because they participated in our study (only 18% said that this was the case) or that participation in the PVE would increase their acceptance of the lockdown policies of the government (only 40% argued that participation in the PVE would increase their acceptance).

A final result of our study is that only 5% of the participants thought that the advice given by citizens in the PVE should have a heavier weighting in the government's decision-making than the advice given by experts. Conversely, 69% of participants opined that the expert advice should weigh heavier. We think that it is surprising that citizens who participated in the PVE – who must have an above-average interest in participating in government decision-making – believe that more weight should be given to scientific advice than to the advice of citizens. The result that only a minority of the participants thinks that advice given by citizens should have a heavier weight in government decisions than expert advice is also observed in other PVEs (Mouter et al., 2021b; Spruit et al., 2020; Spruit & Mouter, 2020). Building upon the results presented in section 2.5.4 a possible explanation for this finding is that the participation of citizens in a PVE increases their awareness of the dilemmas the government faces and the complexity of government decisions which, in turn, leads citizens to the conclusion that layman's opinions should have a modest role in political decision-making when compared to expert opinions.

### 2.6.2. Limitations and further research

One major benefit of a PVE is that it can be deployed rapidly, but at the same time many limitations of our study were caused by the short timeframe. The study was designed in 20 days and the data was collected and analysed in 7 days. It goes without saying that the quality of our study would have been higher if we had had more time to design the study and analyse the data. Had this been the case, we probably would have analysed the written motivations of a larger number of respondents to provide policymakers with an even larger set of new ideas, conditions and values that the respondents aimed to transmit to their government. The fact that we were not able to analyse all the written motivations is problematic because participants cited the fact that PVE provides a lot of opportunities to explain their advises and to add nuances as a key strength of PVE. We believe that this shortcoming can be alleviated by analysing the qualitative data faster and more systematically through natural language processing (Liscio et al., 2021) and using a larger group of annotators. Another limitation of our study that was caused by time pressure is that we were not able to finalise a mobile version in time which might have resulted in a lower participation of younger individuals. Moreover, on the first day of our data collection, the PVE went offline due to lack of server capacity. With a mobile version and enough server capacity in place, we believe that the number of participants would have been substantially higher. Despite these limitations, we believe that the PVE can serve as an example for policymakers and academics of what can realistically be achieved in terms of involving the public in crisis policymaking (Pearse, 2020). PVE is probably a cheaper and more efficient alternative to live experimentation – that is, imposing policies on citizens and seeing what sticks (Pearse, 2020). Moreover, we computed that at least 10,000 respondents were needed to obtain significant parameters for the projects and attributes that were part of this study. Hence, the quality of the quantitative insights that were extracted from this PVE would not increase after 10,000 people participated. Of course, this does not hold for the quality of the qualitative insights and one could argue from a normative point of view that the success of a participatory process always increases when with a higher number of participants.

Another limitation of our study concerns its generalisation to other contexts. This research is only a temporary glance into Dutch citizens' preferences concerning the relaxation of lockdown measures in late April 2020. Citizens in different countries and cultures might have different preferences. Furthermore, preferences can shift as the severity of the pandemic, individual experiences and risk perceptions and the efficacy of pharmaceutical and non-pharmaceutical measures evolve over time. It would be interesting to repeat the PVE in different contexts (time, phase of the pandemic, and location) to explore its generalisability in terms of outcomes and the way that the method is perceived by participants. When this PVE would be repeated in another context we would recommend to also invite experts, policy makers and other stakeholders such as interest groups to conduct the experiment. One of the most striking results of this PVE is that 69% of participants opined that the expert advice should weigh heavier than the advice of citizens, but an omission of our study is that we do not know whether there is a distinction between advices of experts and advices of citizens.

Even though there is no point of comparison, we reason that the quality of preferences that people express in the PVE is probably lower than preferences that they express after deliberation (such as is the case in mini-publics). This is because citizens' interests, preferences and perceptions of a crisis situation are not fixed but subject to discursive challenges. In the PVE, respondents were provided with information on the policy alternatives that were on the government's table, but – as far as we know – most of them studied this information individually, without the opportunity to ask questions of experts, discuss implications with other groups of people, and so forth (Bartkowski & Lienhoop, 2017). Not only is preference formation an inherently social and dynamic process, so is the adherence to social distancing recommendations during the COVID-19 pandemic (Coroiu et al., 2020). Therefore, as mentioned earlier, various scholars argue that deliberation with others is decisive for preference formation (Bartkowski & Lienhoop, 2017; Dietz et al., 2009). When citizens deliberate, they can expand their knowledge, including both their own self-understanding and their collective understanding of what will best serve other affected groups (Warren & Mansbridge, 2013).

Moreover, empirical studies show that individuals interacting with one another generally outperform groups of unconnected individuals (Almaatouq, 2020). Hence, enriching PVE experiments with deliberative elements (e.g., group discussion, consulting expert witnesses or a forum) may contribute to well-formed preferences in the case of unfamiliar and complex government policies and may even increase adherence to subsequent government measures (Bartkowski & Lienhoop, 2017). Augmenting PVE with deliberative elements will allow participating citizens to learn from each other, to form reasoned opinions and to evaluate positions, thereby ironing out critiques of the individual approach to preference formation. It is important to investigate the extent to which the beneficial aspects of social interaction outweigh potential downsides such as social bias, herding and groupthink to ensure that social interaction leads to the 'wisdom of the crowd' instead of the 'madness of the mob' (Almaatouq et al., 2020). For the same reason, we believe that PVE is merely one of several ways to involve citizens in crisis policymaking, and might complement other public participation methods. In our view, PVE could be optimally used jointly with deliberative methods, such as mini-publics. For instance, a mini-public could be used in the design stage of the PVE (selecting relaxation options which are included in the PVE) and could then also be asked to translate the results of the PVE into policy recommendations. A related limitation of our study is that we do not know how their individual preferences were being shaped. That is, we do not know what sources of information – other than the information that was provided to them in the PVE – influenced their choice. Hence, it would be useful to ask the citizens participating in future PVEs which are their main sources of information: traditional media, social media, the internet, friends and relatives, school. This allows policy makers to understand how the public opinion is being shaped.

A final promising avenue for further research would be to study how the results of the PVE could (better) fit in political decision-making processes. In the context of this PVE, we had contact with civil servants, but we were not in contact with the Dutch parliament and we aren't even aware of whether they received our report. It would be interesting to study how a PVE should be institutionalized in a representative democracy, also considering the fact that only 5% of participants in the PVE itself demanded that their advice as citizens should count for more than that of experts.

# Acknowledgements

We are grateful to Kantar Public for charging us less than their usual rate per obtained respondent. We received valuable input from the following individuals dur-

ing the design of the PVE, the implementation of the PVE and the analysis of the data: Marion Collewet, Sjoerd Jenninga, Selma van Delft, Julia Kooijman, Lionel Kaptein, Lisa Volberda, Enrico Liscio, Pradeep Murukannaiah, Perry Borst, Shannon Spruit, Ardine de Wit, Adrienne Rotteveel, Mattijs Lambooij, Anita Suijkerbuijk, Paul van Gils, Toep van Dijk, Tomas Peeters, Suzanne Pietersma, Denny Borsboom, Tessa Blanken, Job van Exel, Sake de Vlas, Marcel Jansen, Vincent van Petten, Nienke van der Haak, Hans Heesterbeek, and Rob Kooij.

## 2.A. Experimental design process

The first part of the experimental design consisted of defining the possible impact levels and pressure on the healthcare system of each relaxation option, based on the feedback and information that was obtained in the PVE design process. Table 2.14 summarizes each possible impact including the increase in the pressure on the health care system caused by the relaxation option.

| Relaxation option | Pressure on the health-care system | Additional deaths of people of +70 years | Additional deaths of people of less than 70 years | Additional people with permanent physical injury | Minus people with permanent mental injury | Minus households that have lost 15% of income |
|---|---|---|---|---|---|---|
| 1: Nursing and care homes allow visitors | 10%<br>15%<br>20%<br>25% | 1500<br>2000<br>3000 | 30<br>50<br>100<br>150<br>300 | 100<br>500<br>1000 | 30000<br>60000 | 50<br>200 |
| 2: Re-open businesses (other than contact professions and hospitality industry) | 6%<br>8%<br>10%<br>15% | 200<br>400<br>600<br>1000 | 150<br>300<br>500<br>750 | 1000<br>2000<br>3000<br>5000<br>7500 | 1000<br>2000<br>5000<br>7500 | 10000<br>20000<br>50000<br>75000 |
| 3: Re-open contact professions | 8%<br>10%<br>15% | 200<br>400<br>600<br>1000 | 150<br>300<br>500<br>750,<br>1000 | 1000<br>2000<br>3000<br>5000<br>7500<br>10000 | 5000<br>7500<br>10000<br>15000 | 20000<br>50000<br>75000 |
| 4: Young people may come together in small groups | 4%<br>6%<br>8% | 50<br>200<br>400 | 50<br>100<br>150<br>300 | 500<br>1000<br>2000<br>3000<br>5000 | 2000<br>5000<br>7500<br>10000<br>15000 | 50<br>200<br>5000 |
| 5: All restrictions lifted for people with immunity | 10%<br>15%<br>20% | 400<br>600<br>1000<br>1500 | 300<br>500<br>750 | 2000<br>3000<br>5000 | 1000<br>2000<br>5000<br>7500 | 5000<br>10000<br>20000 |
| 6: All restrictions lifted in Northern provinces | 15%<br>20%<br>25%<br>30% | 600<br>1000<br>1500<br>2000 | 300<br>500<br>750<br>1000 | 5000<br>7500<br>10000 | 10000<br>15000<br>30000 | 20000<br>50000<br>75000 |
| 7: Direct family members from other households can have social contact | 6%<br>8%<br>10%<br>15% | 600<br>1000<br>1500<br>2000 | 300<br>500<br>750<br>1000 | 2000<br>3000<br>5000<br>7500<br>10000 | 30000<br>60000 | 50 |
| 8: Re-open hospitality and entertainment industry | 15%<br>20%<br>25% | 200<br>400<br>600<br>1000 | 300<br>500<br>750<br>1000 | 1000<br>2000<br>3000<br>5000<br>7500<br>10000 | 15000<br>30000<br>60000 | 50000<br>75000<br>100000 |

*Table 2.14: Possible impact levels and pressure on the healthcare system for each relaxation option.*

Ideally, an experimental design should consider all possible combinations of impact and pressure levels, in order to capture information from all the possible profiles of relaxation strategies from participants choices in the PVE. This is called in literature as a "full-factorial design". However, collecting data for such design is often non-tractable, because the number of combinations explodes even for small numbers of relaxation strategies, impacts, and impact/pressure levels. For this PVE, a full factorial design is composed of more than $1.59 * 10^{26}$ combinations.

There are several solutions to reduce the number of combinations of the experimental design. The first and most intuitive one is to take a random number of profiles (defined by the researcher) of the full factorial design. This is called in literature as a "fractional factorial design". A problem of this approach is that it artificially increases the correlation level between impact/pressure levels of the experimental design. In turn, this increased correlation has an impact on the possibility of extracting information related to preferences for impacts (i.e. taste parameters) using econometric models. In light of this, the construction of a reduced experimental design should ensure that the correlation between attributes is minimal.

The experimental design of this PVE aims to obtain a tractable number of profiles, at the same time that the correlation between impact/pressure levels is minimized. In particular, we aim to minimize the maximum value of the correlation matrix between impact/pressure levels of the design:

$$D^* = \arg\min[\rho_{MAX}], \qquad \rho_{MAX} = \max\left(Corr(\mathbf{X}, \mathbf{C})\right) \tag{2.3}$$

Where $D^*$ is the optimal design matrix, $\mathbf{X}$ is a $N \times J \times K$ matrix of impact levels for $N$ profiles, $J$ relaxation options, and $K$ impacts; $\mathbf{C}$ is a $N \times J$ matrix of pressure levels for each profile and relaxation option. We call this type of designs as "min-max correlation" designs.

We developed an algorithm that creates min-max correlation designs by iteratively selecting impact/pressure levels, evaluating on each step whether the max-correlation is reduced. This algorithm is described as follows:

- Step 0 (Definition of inputs): Define the set of possible impact and pressure levels for each relaxation option. Define $N$ as the number of required profiles.
- Step 1 (Definition of initial candidate design): Construct 10 designs with $N$ profiles each one, by taking random levels of impact and pressure levels from the set defined on Step 0. Keep the design with the smallest max-correlation value and call it as "(initial) candidate design".
- Step 2 (Replacement): Set a random impact/pressure from the candidate design, and replace its value with a random impact/pressure value from the set defined

in Step 0, and call it as the new “candidate design”.

- Step 3 (Evaluation): Compute the max-correlation value of the candidate design.
- Step 4 (Decision): If the max-correlation value of the new candidate design is reduced, the change is kept and go back to Step 2. Otherwise, the change is reverted and go back to Step 2.

This algorithm is conducted until a certain number of iterations without improvement is reached, or until a certain amount of time. Then, the last stored design is called as the optimal design. For this PVE, we ran the algorithm for 10 minutes, although we observed no further improvement after 3 minutes approximately. Finally, we introduced additional constraints to the replacement step in the algorithm in order to avoid that a possible impact/pressure level does not appear in the optimal design.

## 2.B. Descriptive information for sociodemographic variables used in Figure 2.3 and Table 2.3

| Gender | Open sample | Representative sample |
|---|---|---|
| Men | 10425 (49.21%) | 1447 (51.35%) |
| Women | 10705 (50.53%) | 1365 (48.44%) |
| Other | 56 (0.26%) | 6 (0.21%) |
| No answer | 5107 | 540 |
| Age group | | |
| 18-25 yr. | 1894 (8.94%) | 483 (17.14%) |
| 26-35 yr. | 3915 (18.48%) | 430 (15.26%) |
| 36-45 yr. | 3617 (17.07%) | 527 (18.7%) |
| 46-55 yr. | 4749 (22.42%) | 416 (14.76%) |
| 56-65 yr. | 4266 (20.14%) | 360 (12.78%) |
| 66-74 yr. | 2312 (10.91%) | 455 (16.15%) |
| 75+ yr. | 433 (2.04%) | 147 (5.22%) |
| No answer | 5107 | 540 |
| Maximum education level | | |
| No education | 25 (0.12%) | 20 (0.71%) |
| Primary school | 42 (0.2%) | 269 (9.55%) |
| Primary vocational school | 183 (0.86%) | 666 (23.63%) |
| Secondary vocational school | 663 (3.13%) | 410 (14.55%) |
| High school | 1171 (5.53%) | 33 (1.17%) |
| Junior college | 2221 (10.48%) | 174 (6.17%) |
| University of applied sciences | 7925 (37.41%) | 321 (11.39%) |
| University | 8956 (42.27%) | 925 (32.82%) |
| No answer | 5107 | 540 |
| Province | | |
| Groningen | 564 (2.66%) | 102 (3.62%) |
| Friesland | 402 (1.9%) | 157 (5.57%) |
| Drenthe | 358 (1.69%) | 339 (12.03%) |
| Overijssel | 2307 (10.89%) | 431 (15.29%) |
| Flevoland | 731 (3.45%) | 65 (2.31%) |
| Gelderland | 1908 (9.01%) | 218 (7.74%) |
| Utrecht | 2550 (12.04%) | 143 (5.07%) |
| Noord-Holland | 3397 (16.03%) | 76 (2.7%) |
| Zuid-Holland | 6019 (28.41%) | 83 (2.95%) |
| Zeeland | 228 (1.08%) | 225 (7.98%) |
| Noord-Brabant | 1991 (9.4%) | 567 (20.12%) |
| Limburg | 731 (3.45%) | 412 (14.62%) |
| No answer | 5107 | 540 |
| Perceived risk of becoming very ill | | |
| No risk | 531 (2.46%) | 950 (33.12%) |
| Low risk | 8354 (38.74%) | 594 (20.71%) |
| Moderate risk | 8606 (39.91%) | 92 (3.21%) |
| High risk | 3347 (15.52%) | 1017 (35.46%) |
| Extreme risk | 725 (3.36%) | 215 (7.5%) |
| No answer | 4730 | 490 |
| Total sample | 26293 | 3358 |

*Table 2.15: Frequency of sociodemographic variables used in Figure 2.3 and Table 2.3*

## 2.C. MDCEV estimates and optimal portfolio for separate samples

| | Open sample | Representative sample |
|---|---|---|
| Policy-specific constants: | | |
| 1: Nursing and care homes allow visitors | 2.6865*** | 2.7219*** |
| | (0.0297) | (0.0764) |
| 2: Re-open businesses (other than contact professions and hospitality industry) | 2.6451*** | 2.4132*** |
| | (0.0233) | (0.0556) |
| 3: Re-open contact professions | 3.2382*** | 2.8500*** |
| | (0.0276) | (0.0631) |
| 4: Young people may come together in small groups | 1.8825*** | 1.6317*** |
| | (0.0142) | (0.0357) |
| 5: All restrictions lifted for people with immunity | 1.5608*** | 1.9986*** |
| | (0.0211) | (0.0533) |
| 6: All restrictions lifted in Northern provinces | 1.5954*** | 2.0641*** |
| | (0.0342) | (0.0809) |
| 7: Direct family members from other households can have social contact | 2.4893*** | 2.6784*** |
| | (0.0294) | (0.0748) |
| 8: Re-open hospitality and entertainment industry | 2.7346*** | 2.4078*** |
| | (0.0376) | (0.0857) |
| Taste parameters: | | |
| Additional 10.000 deaths of people of +70 years | -0.4123*** | -1.1009*** |
| | (0.0945) | (0.2308) |
| Additional 10.000 deaths of people of less than 70 years | -0.9295*** | -0.4503 |
| | (0.1933) | (0.438) |
| Additional 10.000 people with permanent physical injury | -0.1033*** | -0.1481*** |
| | (0.0174) | (0.0434) |
| Minus 10.000 people with permanent mental injury | 0.0023 | -0.0121 |
| | (0.0037) | (0.0094) |
| Minus 10.000 households that have lost 15% of income | 0.0094*** | -0.0042 |
| | (0.0026) | (0.006) |
| Observations | 26,293 | 3,358 |
| Log-likelihood | -127,928.8123 | -16,499.3413 |
| AIC | 255,831.6246 | 32972.6826 |
| BIC | 255,725.3229 | 32,893.1343 |

Note: Standard errors in parenthesis. **Statistical significance:** ***p < 0.001, **p < 0.01, *p < 0.05

*Table 2.16: MDCEV model estimates. Separate samples*

| | **Open sample** | | | **Representative sample** | | |
|---|---|---|---|---|---|---|
| | Avg. | Pessim. | Optim | Avg. | Pessim. | Optim |
| 1: Nursing and care homes allow visitors | | | X | | | X |
| 2: Re-open businesses (other than contact professions and hospitality industry) | X | | X | | | X |
| 3: Re-open contact professions | X | X | X | X | | X |
| 4: Young people may come together in small groups | | | X | | | X |
| 5: All restrictions lifted for people with immunity | | | | | | |
| 6: All restrictions lifted in Northern provinces | | | | | | |
| 7: Direct family members from other households can have social contact | X | | X | X | | X |
| 8: Re-open hospitality and entertainment industry | | | X | | | |
| **Pressure to the healthcare system** | **31.6%** | **15%** | **49%** | **21.8%** | **0%** | **34%** |

*Table 2.17: Optimal portfolios of relaxation options. Separate samples*

## 2.D. Quantitative results and impact/pressure levels used for sensitivity analysis. Sample of provinces of Friesland, Groningen and Drenthe (the Northern provinces)

| | **Estimates** |
|---|---|
| **Baseline utility of relaxation strategies:** | |
| 1: Nursing and care homes allow visitors | 2.7526*** |
| | (0.1085) |
| 2: Re-open businesses (other than contact professions and hospitality industry) | 2.3516*** |
| | (0.0797) |
| 3: Re-open contact professions | 2.9320*** |
| | (0.0923) |
| 4: Young people may come together in small groups | 1.7161*** |
| | (0.0511) |
| 5: All restrictions lifted for people with immunity | 1.4744*** |
| | (0.0803) |
| 6: All restrictions lifted in Northern provinces | 2.2522*** |
| | (0.1139) |
| 7: Direct family members from other households can have social contact | 2.3574*** |
| | (0.1072) |
| 8: Re-open hospitality and entertainment industry | 2.4172*** |
| | (0.1219) |
| **Impact effects:** | |
| Additional 10.000 deaths of people of +70 years | -0.7926* |
| | (0.3298) |
| Additional 10.000 deaths of people of less than 70 years | -0.7957 |
| | (0.6428) |
| Additional 10.000 people with permanent physical injury | -0.0492 |
| | (0.0616) |
| Minus 10.000 people with permanent mental injury | 0.0042 |
| | (0.0132) |
| Minus 10.000 households that have lost 15% of income | 0.0007 |
| | (0.0085) |
| Observations | 1,645 |
| Log-likelihood | -8,073.5348 |
| AIC | 16,121.0695 |
| BIC | 16,050.7981 |

*Table 2.18: MDCEV model estimates. Sample for individuals who live in the Northern provinces of Friesland, Groningen and Drenthe*

| | **Averages** | **Pessimistic** | **Optimistic** |
|---|---|---|---|
| 1: Nursing and care homes allow visitors | | | X |
| 2: Re-open businesses (other than contact professions and hospitality industry) | | | X |
| 3: Re-open contact professions | X | X | X |
| 4: Young people may come together in small groups | | | X |
| 5: All restrictions lifted for people with immunity | | | |
| 6: All restrictions lifted in Northern provinces | | | |
| 7: Direct family members from other households can have social contact | | | X |
| 8: Re-open hospitality and entertainment industry | | | |
| **Pressure to the healthcare system** | **11.4%** | **15%** | **34%** |

*Table 2.19: Optimal portfolios of relaxation options. Sample for individuals who live in the Northern provinces of Friesland, Groningen and Drenthe*

| Impact | Relaxation strategy | Average | Conservative | Optimistic |
|---|---|---|---|---|
| Additional deaths of people of +70 years | Nursing and care homes allow visitors | 2185.41 | 3000 | 1500 |
| | Re-open businesses (other than contact professions and hospitality industry) | 712.58 | 1000 | 200 |
| | Re-open contact professions | 591.85 | 1000 | 200 |
| | Young people may come together in small groups | 251.52 | 400 | 50 |
| | All restrictions lifted for people with immunity | 1060.49 | 1500 | 400 |
| | All restrictions lifted in Northern provinces | 1218.18 | 2000 | 600 |
| | Direct family members from other households can have social contact | 1127.54 | 2000 | 600 |
| | Re-open hospitality and entertainment industry | 567.29 | 1000 | 200 |
| Additional deaths of people of less than 70 years | Nursing and care homes allow visitors | 130.7 | 300 | 30 |
| | Re-open businesses (other than contact professions and hospitality industry) | 468.33 | 750 | 150 |
| | Re-open contact professions | 576.66 | 1000 | 150 |
| | Young people may come together in small groups | 161.25 | 300 | 50 |
| | All restrictions lifted for people with immunity | 554.89 | 750 | 300 |
| | All restrictions lifted in Northern provinces | 641.82 | 1000 | 300 |
| | Direct family members from other households can have social contact | 605.74 | 1000 | 300 |
| | Re-open hospitality and entertainment industry | 570.67 | 1000 | 300 |
| Additional people with permanent physical injury | Nursing and care homes allow visitors | 619.88 | 1000 | 100 |
| | Re-open businesses (other than contact professions and hospitality industry) | 4049.24 | 7500 | 1000 |
| | Re-open contact professions | 4612.16 | 10000 | 1000 |
| | Young people may come together in small groups | 2636.17 | 5000 | 500 |
| | All restrictions lifted for people with immunity | 3430.4 | 5000 | 2000 |
| | All restrictions lifted in Northern provinces | 7500 | 10000 | 5000 |
| | Direct family members from other households can have social contact | 5422.8 | 10000 | 2000 |
| | Re-open hospitality and entertainment industry | 4859.27 | 10000 | 1000 |
| Reduction of people with permanent mental injury | Nursing and care homes allow visitors | 41069.91 | 30000 | 60000 |
| | Re-open businesses (other than contact professions and hospitality industry) | 4031 | 1000 | 7500 |
| | Re-open contact professions | 9575.99 | 5000 | 15000 |
| | Young people may come together in small groups | 6765.96 | 2000 | 10000 |
| | All restrictions lifted for people with immunity | 3676.6 | 1000 | 7500 |
| | All restrictions lifted in Northern provinces | 17054.71 | 10000 | 30000 |
| | Direct family members from other households can have social contact | 46778.12 | 30000 | 60000 |
| | Re-open hospitality and entertainment industry | 42246.2 | 15000 | 60000 |
| Reduction of households that have lost 15% of income | Nursing and care homes allow visitors | 141.19 | 50 | 200 |
| | Re-open businesses (other than contact professions and hospitality industry) | 38270.52 | 10000 | 75000 |
| | Re-open contact professions | 50434.65 | 20000 | 75000 |
| | Young people may come together in small groups | 1371.37 | 50 | 5000 |
| | All restrictions lifted for people with immunity | 12167.17 | 5000 | 20000 |
| | All restrictions lifted in Northern provinces | 50206.69 | 20000 | 75000 |
| | Direct family members from other households can have social contact | 50 | 50 | 50 |
| | Re-open hospitality and entertainment industry | 75516.72 | 50000 | 100000 |

*Table 2.20: Impact levels used for optimal portfolio computation for three scenarios. Sample for individuals who live in the Northern provinces of Friesland, Groningen and Drenthe*

| Relaxation strategy | Average | Conservative | Optimistic |
|---|---|---|---|
| Nursing and care homes allow visitors | 18.02 | 25 | 10 |
| Re-open businesses (other than contact professions and hospitality industry) | 9.87 | 15 | 6 |
| Re-open contact professions | 11.49 | 15 | 8 |
| Young people may come together in small groups | 6.59 | 8 | 4 |
| All restrictions lifted for people with immunity | 15.19 | 20 | 10 |
| All restrictions lifted in Northern provinces | 22.23 | 30 | 15 |
| Direct family members from other households can have social contact | 10.11 | 15 | 6 |
| Re-open hospitality and entertainment industry | 19.09 | 25 | 15 |

*Table 2.21: Pressure to the healthcare system used for optimal portfolio computation for three scenarios. Sample for individuals who live in the Northern provinces of Friesland, Groningen and Drenthe*

## 2.E. Impact/pressure levels used for sensitivity analysis, for each type of sample.

| Impact | Relaxation strategy | Average | Conservative | Optimistic |
|---|---|---|---|---|
| Additional deaths of people of +70 years | Nursing and care homes allow visitors | 2215.27 | 3000 | 1500 |
| | Re-open businesses (other than contact professions and hospitality industry) | 714.07 | 1000 | 200 |
| | Re-open contact professions | 593.74 | 1000 | 200 |
| | Young people may come together in small groups | 248.61 | 400 | 50 |
| | All restrictions lifted for people with immunity | 1063.01 | 1500 | 400 |
| | All restrictions lifted in Northern provinces | 1220.99 | 2000 | 600 |
| | Direct family members from other households can have social contact | 1120.51 | 2000 | 600 |
| | Re-open hospitality and entertainment industry | 569.73 | 1000 | 200 |
| Additional deaths of people of less than 70 years | Nursing and care homes allow visitors | 129.62 | 300 | 30 |
| | Re-open businesses (other than contact professions and hospitality industry) | 471.64 | 750 | 150 |
| | Re-open contact professions | 573.65 | 1000 | 150 |
| | Young people may come together in small groups | 161.98 | 300 | 50 |
| | All restrictions lifted for people with immunity | 548.56 | 750 | 300 |
| | All restrictions lifted in Northern provinces | 638.18 | 1000 | 300 |
| | Direct family members from other households can have social contact | 597.84 | 1000 | 300 |
| | Re-open hospitality and entertainment industry | 569.27 | 1000 | 300 |
| Additional people with permanent physical injury | Nursing and care homes allow visitors | 614.3 | 1000 | 100 |
| | Re-open businesses (other than contact professions and hospitality industry) | 3961.32 | 7500 | 1000 |
| | Re-open contact professions | 4611.11 | 10000 | 1000 |
| | Young people may come together in small groups | 2592.76 | 5000 | 500 |
| | All restrictions lifted for people with immunity | 3404.78 | 5000 | 2000 |
| | All restrictions lifted in Northern provinces | 7522.15 | 10000 | 5000 |
| | Direct family members from other households can have social contact | 5540.2 | 10000 | 2000 |
| | Re-open hospitality and entertainment industry | 4945.23 | 10000 | 1000 |
| Reduction of people with permanent mental injury | Nursing and care homes allow visitors | 40728.71 | 30000 | 60000 |
| | Re-open businesses (other than contact professions and hospitality industry) | 4059.84 | 1000 | 7500 |
| | Re-open contact professions | 9532.19 | 5000 | 15000 |
| | Young people may come together in small groups | 6756.25 | 2000 | 10000 |
| | All restrictions lifted for people with immunity | 3721.41 | 1000 | 7500 |
| | All restrictions lifted in Northern provinces | 17249.46 | 10000 | 30000 |
| | Direct family members from other households can have social contact | 46466.74 | 30000 | 60000 |
| | Re-open hospitality and entertainment industry | 42606.21 | 15000 | 60000 |
| Reduction of households that have lost 15% of income | Nursing and care homes allow visitors | 140.75 | 50 | 200 |
| | Re-open businesses (other than contact professions and hospitality industry) | 39210.82 | 10000 | 75000 |
| | Re-open contact professions | 49987.83 | 20000 | 75000 |
| | Young people may come together in small groups | 1408.73 | 50 | 5000 |
| | All restrictions lifted for people with immunity | 12051.5 | 5000 | 20000 |
| | All restrictions lifted in Northern provinces | 49546.08 | 20000 | 75000 |
| | Direct family members from other households can have social contact | 50 | 50 | 50 |
| | Re-open hospitality and entertainment industry | 75771.12 | 50000 | 100000 |

*Table 2.22: Impact levels used for optimal portfolio computation for three scenarios. Open sample*

| **Relaxation strategy** | **Average** | **Conservative** | **Optimistic** |
|---|---|---|---|
| Nursing and care homes allow visitors | 17.92 | 25 | 10 |
| Re-open businesses (other than contact professions and hospitality industry) | 9.87 | 15 | 6 |
| Re-open contact professions | 11.51 | 15 | 8 |
| Young people may come together in small groups | 6.54 | 8 | 4 |
| All restrictions lifted for people with immunity | 15.15 | 20 | 10 |
| All restrictions lifted in Northern provinces | 22.25 | 30 | 15 |
| Direct family members from other households can have social contact | 10.22 | 15 | 6 |
| Re-open hospitality and entertainment industry | 18.97 | 25 | 15 |

*Table 2.23: Pressure to the healthcare system used for optimal portfolio computation for three scenarios. Open sample*

| Impact | Relaxation strategy | Average | Conservative | Optimistic |
|---|---|---|---|---|
| Additional deaths of people of +70 years | Nursing and care homes allow visitors | 2223.65 | 3000 | 1500 |
| | Re-open businesses (other than contact professions and hospitality industry) | 706.37 | 1000 | 200 |
| | Re-open contact professions | 588.92 | 1000 | 200 |
| | Young people may come together in small groups | 250.07 | 400 | 50 |
| | All restrictions lifted for people with immunity | 1062.63 | 1500 | 400 |
| | All restrictions lifted in Northern provinces | 1221.11 | 2000 | 600 |
| | Direct family members from other households can have social contact | 1110.75 | 2000 | 600 |
| | Re-open hospitality and entertainment industry | 570.4 | 1000 | 200 |
| Additional deaths of people of less than 70 years | Nursing and care homes allow visitors | 129.01 | 300 | 30 |
| | Re-open businesses (other than contact professions and hospitality industry) | 471.86 | 750 | 150 |
| | Re-open contact professions | 569.57 | 1000 | 150 |
| | Young people may come together in small groups | 161.49 | 300 | 50 |
| | All restrictions lifted for people with immunity | 549.24 | 750 | 300 |
| | All restrictions lifted in Northern provinces | 639.9 | 1000 | 300 |
| | Direct family members from other households can have social contact | 597.29 | 1000 | 300 |
| | Re-open hospitality and entertainment industry | 569.88 | 1000 | 300 |
| Additional people with permanent physical injury | Nursing and care homes allow visitors | 604.05 | 1000 | 100 |
| | Re-open businesses (other than contact professions and hospitality industry) | 4010.42 | 7500 | 1000 |
| | Re-open contact professions | 4625.67 | 10000 | 1000 |
| | Young people may come together in small groups | 2580.85 | 5000 | 500 |
| | All restrictions lifted for people with immunity | 3410.66 | 5000 | 2000 |
| | All restrictions lifted in Northern provinces | 7510.42 | 10000 | 5000 |
| | Direct family members from other households can have social contact | 5498.66 | 10000 | 2000 |
| | Re-open hospitality and entertainment industry | 4864.35 | 10000 | 1000 |
| Reduction of people with permanent mental injury | Nursing and care homes allow visitors | 40640.26 | 30000 | 60000 |
| | Re-open businesses (other than contact professions and hospitality industry) | 4032.61 | 1000 | 7500 |
| | Re-open contact professions | 9586.81 | 5000 | 15000 |
| | Young people may come together in small groups | 6745.09 | 2000 | 10000 |
| | All restrictions lifted for people with immunity | 3725.58 | 1000 | 7500 |
| | All restrictions lifted in Northern provinces | 17374.93 | 10000 | 30000 |
| | Direct family members from other households can have social contact | 46340.08 | 30000 | 60000 |
| | Re-open hospitality and entertainment industry | 42480.64 | 15000 | 60000 |
| Reduction of households that have lost 15% of income | Nursing and care homes allow visitors | 140.54 | 50 | 200 |
| | Re-open businesses (other than contact professions and hospitality industry) | 38941.33 | 10000 | 75000 |
| | Re-open contact professions | 50064.03 | 20000 | 75000 |
| | Young people may come together in small groups | 1466.24 | 50 | 5000 |
| | All restrictions lifted for people with immunity | 12182.85 | 5000 | 20000 |
| | All restrictions lifted in Northern provinces | 50023.82 | 20000 | 75000 |
| | Direct family members from other households can have social contact | 50 | 50 | 50 |
| | Re-open hospitality and entertainment industry | 75707.27 | 50000 | 100000 |

*Table 2.24: Impact levels used for optimal portfolio computation for three scenarios. Representative sample*

| **Relaxation strategy** | **Average** | **Conservative** | **Optimistic** |
|---|---|---|---|
| Nursing and care homes allow visitors | 17.99 | 25 | 10 |
| Re-open businesses (other than contact professions and hospitality industry) | 9.91 | 15 | 6 |
| Re-open contact professions | 11.49 | 15 | 8 |
| Young people may come together in small groups | 6.52 | 8 | 4 |
| All restrictions lifted for people with immunity | 15.17 | 20 | 10 |
| All restrictions lifted in Northern provinces | 22.21 | 30 | 15 |
| Direct family members from other households can have social contact | 10.30 | 15 | 6 |
| Re-open hospitality and entertainment industry | 19.02 | 25 | 15 |

*Table 2.25: Impact levels used for optimal portfolio computation for three scenarios. Open sample*

## 2.F. Corrected optimal portfolio for representative education level groups

Dekker et al. (2019) provides a general expression of the expected utility of a portfolio for different sociodemographic groups. Let $g = 1, \ldots, G$ be a sociodemographic group of the population with its own expected utility equal to $EU_g$. Then the expected utility of society is the weighted sum of the expected utility of each sociodemographic group:

$$EU_{corrected} = \sum_{g=1}^{G} Q_g EU_g = \sum_{g=1}^{G} Q_g \left( y_0 E\left[\Psi_{ng0}\right] + \sum_{j=1}^{J} y_{nj} E\left[\Psi_{ngj}\right] \right), \tag{2.4}$$

Where $Q_g$ represents the proportion of individuals that belong to sociodemographic group $g$. Notice that $\Psi_{ng0}$ depends of each sociodemographic group, as a difference with the expression of section 4.3, which is independent of $g$. This implies that a corrected optimal portfolio requires different parameter estimates of the MDCEV model for the different sociodemographic groups. Then, the computation of a corrected optimal portfolio involves the evaluation of $EU_{corrected}$ for all feasible combinations of policy options (i.e. the combinations that satisfy the resource constraint), and values for $Q_g$ taken from external sources, such as census data. S6 Table 1 summarizes the results of the MDCEV model with a low education effect incorporated to allow the computation of corrected optimal portfolios. These effects are included as additional policy-specific constants present in respondents with low education levels (VMBO, MAVO, Mulo or lower).

S6 Table 2 summarizes the corrected optimal portfolios using the estimates of the MDCEV model and using a correction rate of 28,5% for low educated individuals, according to the Dutch census data. We observe no differences with the uncorrected optimal portfolio of Table 5.

| | **Estimates** | **Low education effects** |
|---|---|---|
| **Policy-specific constants:** | | |
| 1: Nursing and care homes allow visitors | 2.7623*** | -0.0356 |
| | (0.0320) | (0.0490) |
| 2: Re-open businesses (other than contact professions and hospitality industry) | 2.6799*** | -0.5540*** |
| | (0.0272) | (0.0488) |
| 3: Re-open contact professions | 3.2557*** | -0.3297*** |
| | (0.0313) | (0.0527) |
| 4: Young people may come together in small groups | 1.9070*** | -0.4114*** |
| | (0.0148) | (0.0493) |
| 5: All restrictions lifted for people with immunity | 1.6116*** | 0.1422** |
| | (0.0274) | (0.0526) |
| 6: All restrictions lifted in Northern provinces | 1.6634*** | 0.0845 |
| | (0.0399) | (0.0580) |
| 7: Direct family members from other households can have social contact | 2.5272*** | -0.0311 |
| | (0.0351) | (0.0502) |
| 8: Re-open hospitality and entertainment industry | 2.7437*** | -0.4511*** |
| | (0.0399) | (0.0489) |
| **Taste parameters:** | | |
| Additional 10.000 deaths of people of +70 years | -0.5904*** | |
| | (0.0993) | |
| Additional 10.000 deaths of people of less than 70 years | -0.9304** | |
| | (0.2942) | |
| Additional 10.000 people with permanent physical injury | -0.1137*** | |
| | (0.0174) | |
| Minus 10.000 people with permanent mental injury | 0.0012 | |
| | (0.0037) | |
| Minus 10.000 households that have lost 15% of income | 0.0085*** | |
| | (0.0025) | |
| Observations | 24,004 | |
| Log-likelihood | -117,305.5958 | |
| AIC | 234,569.1916 | |
| BIC | 234,399.3861 | |

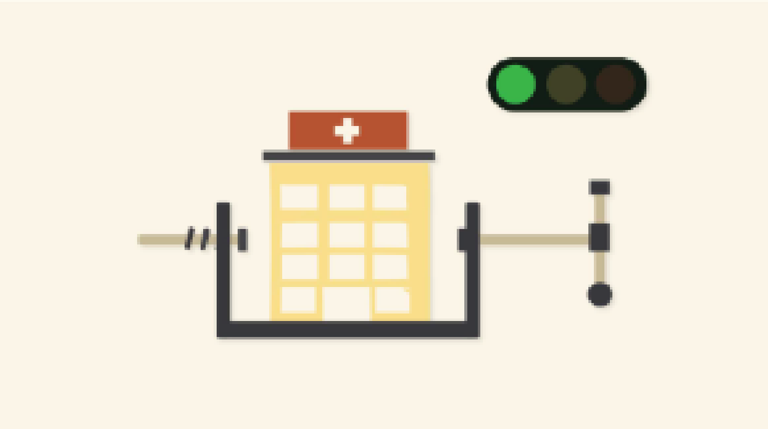
*Table 2.26: MDCEV estimation results with low education effects.*

| | Average | Pessimistic | Optimistic |
|---|---|---|---|
| 1: Nursing and care homes allow visitors | | | X |
| 2: Re-open businesses (other than contact professions and hospitality industry) | X | | X |
| 3: Re-open contact professions | X | X | X |
| 4: Young people may come together in groups | | | X |
| 5: All restrictions lifted for people with immunity | | | |
| 6: All restrictions lifted in Northern provinces | | | |
| 7: Direct family members from other households can have social contact | X | | X |
| 8: Re-open hospitality and entertainment industry | | | |
| **Added pressure onto the healthcare system** | **32%** | **15%** | **34%** |

*Table 2.27: Corrected optimal portfolios of relaxation options.*

## 2.G. Survey instrument

## Participatory Value Evaluation about relaxing COVID-19 lockdown measures



### Instruction part

Welcome to this online consultation about the relaxation of corona lockdown measures from 20th May – 20th July, 2020.

Following the outbreak of the new coronavirus, COVID-19, in the Netherlands, the government has taken various measures to control the spread of the virus, to protect high-risk groups, such as the elderly and people in a precarious state of health, and to prevent various parts of the healthcare system from becoming overloaded.

Now that the measures appear to be effective, a number of these measures can be relaxed. Would you like certain measures to be relaxed between 20th May and 20th July? And if so, which measures should be relaxed first? The government would like to receive advice from a large group of Dutch citizens about which of these 'relaxation options' are preferred.

The research is being done by researchers at Delft University of Technology in collaboration with researchers from other universities and researchers from the Dutch National Institute for Public Health and Environment (RIVM). Policy staff from the Ministry of Health, Welfare and Sport and the Ministry of Finance also participated. The results of the research will be shared with the RIVM and other researchers who think along with and advise the government about corona.

We would like to thank you very much, in advance, for participating in this consultation!

This consultation is in two parts:

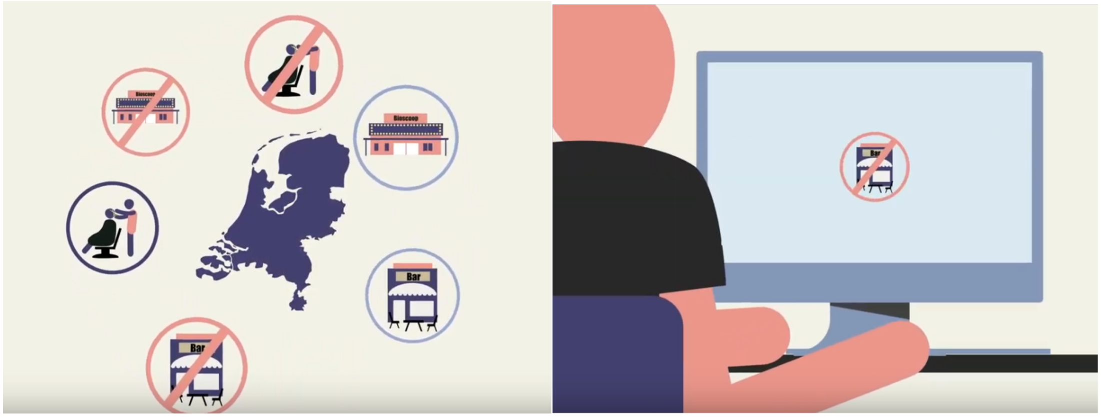
**Part 1: Advice on the relaxation options (takes: 15 – 20 minutes)** We present a number of ways in which corona measures can be relaxed over the next two months (‘relaxation options’). Thereafter, we ask you to advise the government. Do you think the government should introduce relaxation options from 20th May – 20th July, 2020, and if so, which relaxation options should be chosen?

**Part 2: Other advice and rationale (takes: 5 – 10 minutes)** We ask you if there are any relaxation options that you feel should not be considered. We then ask you to explain your choices. We are interested to know why you feel that certain measures should or should not be relaxed. Lastly, we ask a number of general questions about you (gender, age, where you live, profession).

**Rules**

- The research has been approved by TU Delft’s Ethics Committee.
- Your answers will be saved at TU Delft on a secure server until no later than 27th April, 2030.
- Only citizens over 18 years old may participate in this research.

## Participating interface: The policy measures



### Nursing and care homes allow visitors

Care homes allow one visitor 2 times a week. The elderly and people with a mental or physical disability who live in a care home can receive visitors again. Visitors must be healthy.

**The most important positive effects:**

- Because the elderly and people with a mental or physical handicap can receive visitors, there is a reduction in psychological complaints (especially from the people living in care homes, but also from the people who can visit them again).
- For care home staff the situation is much more pleasant as people are not as lonely.

**The most important negative effects:**

- Increase in the mortality rate among care home residents (especially people in the 75+ category). Residents who become ill due to corona are often not sent to Intensive Care, but they die in the care home.
- Higher work pressure for care home workers.

**The effects of this relaxation option:** Below you will find more information about the effects of this relaxation option. Click here for more information about the uncertainties surrounding the estimations of these effects. When we talk about impacts, we mean the differences between relaxing the measures within two months (20th May – 20th July, 2020) and relaxing the measures after two months (from 20th July, 2020).

### Employees in contact professions (e.g. hairdressers) go back to work

When you choose this relaxation option, employees in contact professions will be able to work again in the next two months. Hairdressers, beauticians, make-up artists, pedicurists, manicurists, driving instructors and tattoo artists can all re-open their shops. This will be done in stages. For example, hairdressers will be able to re-open their salons before tattoo artists. Employees will have to try to keep contact with their customers to a minimum and should wear as much protective clothing as possible. Employers can choose to allow staff who fall under a risk category (e.g. people with pulmonary disease, diabetes or a chronic heart condition) to work from home. People are advised to work from home if they have symptoms that could indicate corona (e.g. a runny nose, cough or fever). Public transport may be used but the 1.5 meter distance rule applies here too, so there is a chance of increased travel time.

**The most important positive effects:**

- The economy will start to get moving again. Therefore, the number of bankruptcies and job losses will decrease compared to the situation if these professions are only allowed to re-start after 20th July. This means that there will be fewer people (15%) whose income is badly affected for a period of more than three years.
- Decrease in sustained psychological complaints from people whose psychological complaints were caused by their job or business being at risk.

**The most important negative effects:**

- More people will become ill. This is because the risk of getting infected is relatively high within the contact profession group and because employees in contact professions, in general, are constantly meeting other people. If they are ill, they can infect large groups of people, who in turn spread the virus across the region.

**The effects of this relaxation option:** Below you will find more information about the effects of this relaxation option. Click here for more information about the uncertainties surrounding the estimations of these effects. When we talk about impacts, we mean the differences between relaxing the measures within two months (20th May – 20th July, 2020) and relaxing the measures after two months (from 20th July, 2020).

### Social contact is allowed again for direct family members from other households

Currently, only people from within the same household don't have to keep a 1.5 meter distance from one another. When you choose this relaxation option, physical contact between first and second tier family members will be allowed. For example, grandparents may reestablish contact with their grandchildren, and brothers and sisters can visit one another again. Direct contact with third tier family members (uncles, aunts, cousins, great-grandparents) are still not allowed. Only 50 people are allowed to attend weddings if the 1.5 meter distance rule is respected by everyone who isn't a direct family member.

**Positive effects:**

- Relaxing this measure will lead to a decrease in loneliness.
- Happy moments (birthdays and weddings) and sad moments (deaths) can be commemorated in a more enjoyable way.

**Negative effects:**

- The number of people that have direct contact (within 1.5 meters) with one another will vastly increase, which means that the number of sick people and the mortality rate will increase. There are two reasons why the number of deaths will still be relatively restricted: 1) Expectations are that, especially at first, many people will put off visiting family members who are in one of the higher risk groups; 2) It will still be the same circle of people who have contact with one another. This ensures that the virus can't spread as quickly as when people constantly come into contact with a different group of people.

**The effects of this relaxation option:** Below you will find more information about the effects of this relaxation option. Click here for more information about the uncertainties surrounding the estimations of these effects. When we talk about impacts, we mean the differences between relaxing the measures within two months (20th May – 20th July, 2020) and relaxing the measures after two months (from 20th July, 2020).

**Businesses open again (except the hospitality industry and contact professions)**

At the moment, Dutch citizens have to work from home if at all possible. Therefore, millions of Dutch citizens are currently working from home (consider the employees of large corporates, universities and the civil service). When you choose this relaxation option, employees can return to their offices when it can be guaranteed that they will be able to keep a 1.5 meter distance. Hospitality industry workers and contact professionals (such as hairdressers and physiotherapists) are not included in this relaxation option. Employers can choose to allow staff who fall under a risk category (e.g. people with pulmonary disease, diabetes or a chronic heart condition) to work from home. People are advised to work from home if they have symptoms that could indicate corona (e.g. a runny nose, cough or fever). Public transport may be used but the 1.5 meter distance rule applies here too, so there is a chance of increased travel time.

**The most important positive effects:**

- The economy will start to get moving again. Therefore, the number of bankruptcies and job losses will decrease compared to the situation if these professions are only allowed to re-start after 20th July. This means that there will be fewer people (15%) whose income is badly affected for a period of more than three years.

- Decrease in the number of sustained psychological complaints caused by working from home. Having to work from home can lead to loneliness or signs of a burn-out, for example.

**The most important negative effects:**

- When the commuter traffic increases, the virus will spread relatively quickly. The number of people who become sick will increase. The number of deaths and the number of sustained physical health problems will also increase because of this.

**The effects of this relaxation option:** Below you will find more information about the effects of this relaxation option. Click here for more information about the uncertainties surrounding the estimations of these effects. When we talk about impacts, we mean the differences between relaxing the measures within two months (20th May – 20th July, 2020) and relaxing the measures after two months (from 20th July, 2020).

### All restrictions are lifted for people who are immune

When you choose this relaxation option, people who have had the coronavirus can continue living their normal life without any restrictions. Tests can show if someone has enough antibodies in their blood. When this is the case, citizens will receive a corona letter and they should always carry this around with them, as a kind of passport. In the next two months, these tests will not yet be sufficiently reliable, so it could be that someone who has tested positive, could, in fact, not be immune (due to a flawed positive test) and could become ill through this. At the moment, it appears that approximately 4% of all Dutch citizens have had corona (approximately 650,000 people), but the percentage of Dutch citizens who have built up sufficient antibodies is not yet known.

**Positive effects:**

- People who are immune can again come and go wherever they please. They can go back to work and this means that there will be fewer people with a serious (15%) drop in income for a period of more than three years.

- People who are immune can visit family and friends again. Through this, the number of people who feel lonely will decrease.

**Negative effect:**

- In the next two months, the immunity tests could still be unreliable, so people who appear to be immune could still become infected. Under normal circumstances, it can take years to develop a good immunity test. Also, people who only have a few antibodies in their blood, can probably become infected again, even if they have already had the virus. As a consequence, these people can infect a lot of other people if they have contact with people from across the country. This short film provides more information about the problems when testing for immunity.
- The instructions explained that you should assume that the testing capacity has vastly improved, but the number of people that can be tested per day remains limited. Some people who are immune will have to wait for a long time before they are tested.
- People who aren't immune could have more problems sticking to the measures now that they see that people who are immune can come and go as they please. The virus will spread quickly when the number of people that no longer adhere to the rules increases, and there will even be people who catch the virus on purpose so that they will be given a corona letter. It is unclear if this measure can be easily regulated.

**The effects of this relaxation option:** Below you will find more information about the effects of this relaxation option. Click here for more information about the uncertainties surrounding the estimations of these effects. When we talk about impacts, we mean the differences between relaxing the measures within two months (20th May – 20th July, 2020) and relaxing the measures after two months (from 20th July, 2020).

### Hospitality and entertainment sectors open again

When you choose this relaxation option, the hospitality and entertainment sector (consider restaurants, cafés, gyms, theme parks, museums, theaters and movie theaters) can re-open if it can be guaranteed that people will be able to keep to the 1.5 meter distance rule. The 1.5 meter distance rule applies to staff in relation to one another, as well as to staff in relation to the customers. Employers can choose to allow staff who fall under a risk category (e.g. people with pulmonary disease, diabetes or a chronic heart condition) to work from home. People are advised to work from home if they have symptoms that could indicate corona (e.g. a runny nose, cough or fever). Public transport may be used but the 1.5 meter distance rule applies here too, so there is a chance of increased travel time.

**The most important positive effects:**

- The economy will start to get moving again. Therefore, the number of bankruptcies and job losses will decrease compared to the situation if these professions are only allowed to re-start after 20th July. This means that there will be fewer people (15%) whose income is badly affected for a period of more than three years.
- Decrease in sustained psychological complaints from people whose psychological complaints were caused by their job or business being at risk. The number of psychological complaints will also decrease because people will be able to enjoy going to a café, restaurant, bar or some other entertainment.

**The most important negative effects:**

- Despite the distancing rule, there will be an increase in the number of people who become ill. It will also be the cause of an increase in mortality rates and the number of sustained physical health problems.

**The effects of this relaxation option:** Below you will find more information about the effects of this relaxation option. Click here for more information about the uncertainties surrounding the estimations of these effects. When we talk about impacts, we mean the differences between relaxing the measures within two months (20th May – 20th July, 2020) and relaxing the measures after two months (from 20th July, 2020).

### Restrictions lifted in Friesland, Groningen and Drenthe

When you choose this relaxation option, all the restrictive measures will be lifted for the provinces of Friesland, Groningen and Drenthe because the coronavirus is very much under control in these regions. Citizens in these provinces account for just 3% of the number of hospital admissions through corona in the Netherlands. There are two exceptions. The restrictive measures will not be lifted for people in high risk groups (the elderly over 75, and people with pulmonary disease, diabetes or a chronic heart condition). For the next two months, events and meetings with more than 50 attendees will still be forbidden. Hospitals across the country will accommodate any patients from these regions. In the first few weeks, people from these Northern provinces may only enter or leave the area if they have a valid reason. Checks will be held on the access roads. This can restrict the spread of the virus to other regions and it will restrain any ‘hospitality tourism’. Depending on how the pressure on the healthcare system is affected, this measure could be relaxed or tightened. The percentage shown at the top of this page is the target rate for pressure on the healthcare system.

**The most important positive effects:**

- The economy in Friesland, Groningen and Drenthe will start to get moving again. There will be limited economic damage caused by the corona crisis in this region.

- Fewer complaints of loneliness and other physical complaints, such as a burn-out and depression among the citizens of Friesland, Groningen and Drenthe.

**The most important negative effects:**

- The number of infected people, deaths, and cases of sustained physical health problems will increase, but because the testing capacity will have greatly increased, people who are infected can be quickly isolated.

- People who live in the other provinces, outside Friesland, Groningen and Drenthe, will have more problems sticking to the measures now that they see that people in another part of the Netherlands can come and go as they please. The virus will spread quickly when the number of people that no longer adhere to the rules increases. It is unclear if this can be easily regulated.

- The checks on the access roads will create increased travel time.

**The effects of this relaxation option:** Below you will find more information about the effects of this relaxation option. Click here for more information about the uncertainties surrounding the estimations of these effects. When we talk about impacts, we mean the differences between relaxing the measures within two months (20th May – 20th July, 2020) and relaxing the measures after two months (from 20th July, 2020).

### Young people may meet up in groups

Currently, young people up to 18 years of age are allowed to take part in organized sport. When you choose this relaxation option, the corona measure will be lifted in phases between 20th May and 20th July for young people up to and including 25 years old, in that they will be allowed to meet in groups and they will no longer have to respect the 1.5 meter distance rule. It is very important that they do have to respect the 1.5 meter distance rule for older people and this will be upheld.

Young people carry a much smaller risk of becoming seriously ill from the coronavirus than the elderly (although there are exceptions). Just 1% - 1.5% of corona patients that have been admitted to hospital were under 25 years old.

In the first phase, children under 12 years old can meet up again in groups of 10. They can do contact sports in a group of 10, for example. If this appears to have little effect on the number of corona infections, and it appears that young people still respect the 1.5 meter distance rule when they are with older people, then the relaxation of the measure will be gradually phased in and extended to bigger groups (25, 50) and to older age groups (12 – 18 years old; 18 – 25 years old). So, in an ideal situation, by the end of June, Dutch citizens who are 25 years old or younger will be able to meet up again in groups of 50, but it could also be that it is kept to the under 12s in groups of no more than 10 for quite a while.

Contact between young people and the elderly must still always be avoided, of course.

**The most important positive effects:**

- There will be a decrease in sustained psychological complaints among young people. Also, there will be more breathing space within families now that the children can meet up with their friends again.

**The most important negative effects:**

- There will be a slight increase in the number of young people admitted to hospitals and the mortality rate, but this will be relatively limited because young people carry a much lower risk of becoming seriously ill from the coronavirus. People who have contact with young people will assume a relatively big risk.
- It will become more difficult for young people to respect the 1.5 meter distance rule between themselves and older people when they no longer have to respect this rule among themselves.

**The effects of this relaxation option:** Below you will find more information about the effects of this relaxation option. Click here for more information about the uncertainties surrounding the estimations of these effects. When we talk about impacts, we mean the differences between relaxing the measures within two months (20th May – 20th July, 2020) and relaxing the measures after two months (from 20th July, 2020).

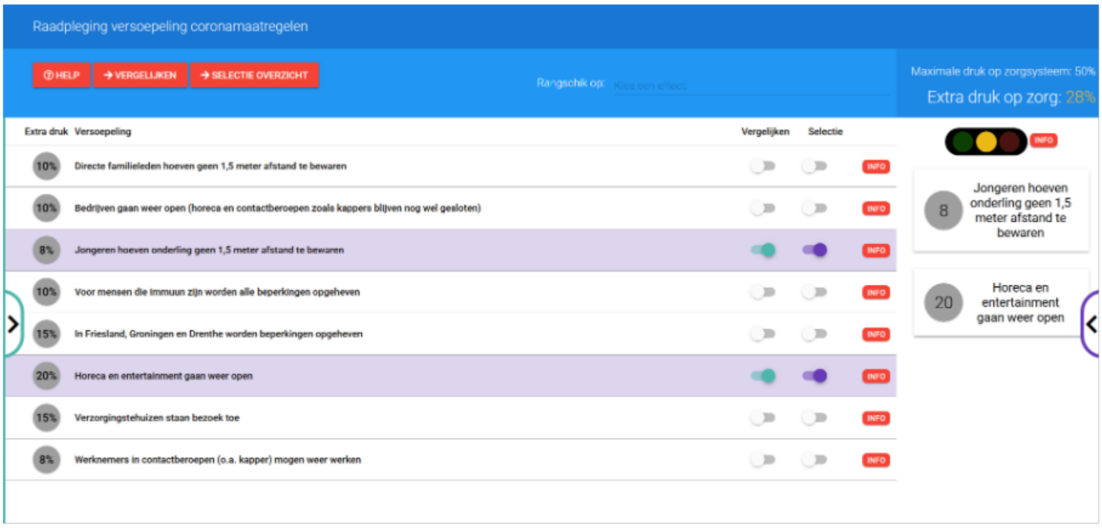
## You can compare your selected relaxation options.

**Comparison options:**

- Increased pressure on the healthcare system
- Increase in the number of deaths 70+
- Increase in the number of deaths under 70 years old

- Increase in the number of people with permanent physical health problems
- Decrease in the number of people with permanent psychological health problems
- Decrease in the number of households with long-term loss of income

## Screenshot of the webtool (In Dutch)



Maximale druk op zorgsysteem: 50%

Extra druk op zorg: 28%

| Extra druk | Versoepeling | Vergelijken | Selectie | |
|---|---|---|---|---|
| 10% | Directe familieleden hoeven geen 1,5 meter afstand te bewaren | ○ | ○ | INFO |
| 10% | Bedrijven gaan weer open (horeca en contactberoepen zoals kappers blijven nog wel gesloten) | ○ | ○ | INFO |
| 8% | Jongeren hoeven onderling geen 1,5 meter afstand te bewaren | ● | ● | INFO |
| 10% | Voor mensen die immuun zijn worden alle beperkingen opgeheven | ○ | ○ | INFO |
| 15% | In Friesland, Groningen en Drenthe worden beperkingen opgeheven | ○ | ○ | INFO |
| 20% | Horeca en entertainment gaan weer open | ● | ● | INFO |
| 15% | Verzorgingstehuizen staan bezoek toe | ○ | ○ | INFO |
| 8% | Werknemers in contactberoepen (o.a. kapper) mogen weer werken | ○ | ○ | INFO |

8 Jongeren hoeven onderling geen 1,5 meter afstand te bewaren

20 Horeca en entertainment gaan weer open

## You have chosen these relaxation options.

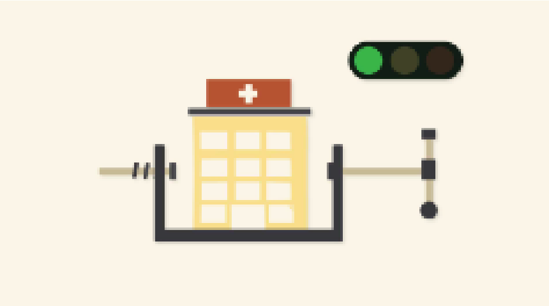
Below you can see the relaxation options that you have recommended.

In this survey, you can only recommend a limited number of relaxation options. The pressure on the healthcare system may not increase by more than 50%. You may recommend that the government should not relax any measures between 20th May and 20th July.

If you would like to amend your advice, click on the red ‘back’ button. If you are happy with your advice, click on the ‘send’ button below. We will then ask you a few more brief questions.

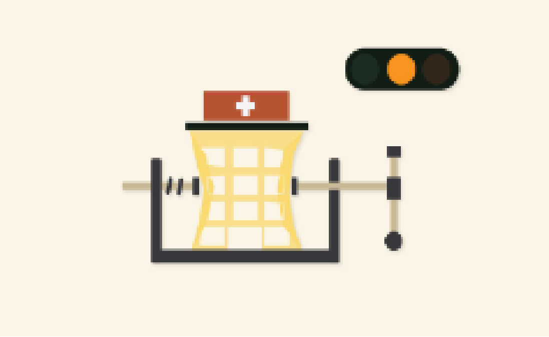
## Pressure on the healthcare system

**Green (0% - 25% increased pressure)** The healthcare system is not overstretched. People in the healthcare sector do not have to work overtime. The healthcare sector can catch its breath, and the chance of employees leaving the healthcare sector in the short and long term is more or less the same as in the period before the corona crisis (at the start of 2020). There is sufficient room to be able to handle treatments other than corona.

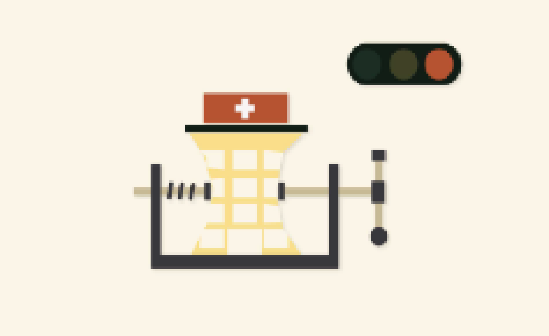


**Yellow (26% - 40% increased pressure)** The healthcare system is overstretched. People in the healthcare sector work an extra 6 hours per week, on average. The healthcare sector can catch its breath somewhat but there is still a chance that employees in the healthcare sector will leave in the short and long term. Some of the treatments other than corona have to be postponed.

**Red (41% - 50% increased pressure)** The health care system is heavily overstretched. People in the healthcare sector work an extra 12 hours per week, on average. The healthcare sector is unable to catch its breath and there is a big chance that employees in the health care sector will leave in the short and long term. All treatments other than corona that are not absolutely necessary have to be postponed. There is a possible shortage of protective material. Nurses and doctors who would normally work on another ward (e.g. Oncology, Cardiology and Neurology) must now work on the corona intensive care ward. So, in fact, for a long time, these healthcare workers will



have to do a different job. This can be tough as healthcare workers will have to work on a different team, so there could be some doubt about if the right choice has been made.



**An increase in pressure of more than 50% is not possible**

In this survey it is not possible to choose an increase in pressure on the healthcare system of more than 50%, because it would then no longer be possible to treat all patients that would have a chance of recovery. The government wants to avoid this scenario.

## Part 2: Additional questions

We would like to ask you a few further, general questions.

1. Please motivate your choice (participants are asked to provide verbal explanations for the options they selected):

2. Are there any relaxation options that you think the government should not consider?

3. Can you explain why you feel that these relaxation options should not be considered?
4. Can you indicate which of the relaxation options would have a big effect on your life? For each policy option:
   *a*) No effect
   *b*) Small effect
   *c*) Reasonable effect
   *d*) Big effect
   *e*) Very big effect
5. As well as consulting a large group of citizens, the government will also be consulting a number of researchers. In your opinion, how much value should the government put on this advice from the citizens and the researchers?
   *a*) Only follow the advice of citizens
   *b*) More value to advice of citizens than academics
   *c*) Equal value to advice of citizens and academics
   *d*) More value to advice of academics than citizens
   *e*) Only follow the advice of academics
6. Could you explain your answer to the previous question?
7. You will now be shown a number of statements. For each statement, please indicate how strongly you agree or disagree with this statement?
   *a*) I am certain that my advice is right
      1) Totally disagree
      2) Disagree
      3) Neutral
      4) Agree
      5) Totally agree
   *b*) By taking part in this research I have learned more about the choices that the government has to make
      1) Totally disagree

2) Disagree
3) Neutral
4) Agree
5) Totally agree

*c*) Participating in this research has influenced by opinion about the appropriateness of certain relaxation options

1) Totally disagree
2) Disagree
3) Neutral
4) Agree
5) Totally agree

*d*) This is a good way of involving Dutch citizens in decisions that the government has to make about relaxing corona measures between 20th May and 20th July

1) Totally disagree
2) Disagree
3) Neutral
4) Agree
5) Totally agree

*e*) The government should use this method more often to include citizens in government policy making.

1) Totally disagree
2) Disagree
3) Neutral
4) Agree
5) Totally agree

*f*) Now that the government has asked for my advice, I am more inclined to comply with the corona measures

1) Totally disagree
2) Disagree
3) Neutral
4) Agree
5) Totally agree

*g*) I am confident that most Dutch citizens will adhere to the corona measures in the next three months

1) Totally disagree
2) Disagree
3) Neutral
4) Agree
5) Totally agree

*h*) Because the government is involving citizens in this way, it will be easier for me to accept the government's final decision concerning the relaxation of corona regulations between 20th May and 20th July

1) Totally disagree
2) Disagree
3) Neutral
4) Agree
5) Totally agree

8. Would you like to pass on any ideas to the government for when they are considering the relaxation of the corona measures? Include your advice below:

9. Have you been infected by the corona virus?

*a*) No, tested and negative

*b*) Probably not, but haven't been tested

*c*) Probably, but haven't been tested

*d*) Yes, tested and positive

*e*) I don't want to answer this questions

10. Are there people in your direct environment (family in your household, other family, friends) who are (or have been) infected by the coronavirus?

*a*) No, tested and negative

*b*) Probably not, but haven't been tested

*c*) Probably, but haven't been tested

*d*) Yes, tested and positive

*e*) I don't want to answer this questions

11. How would you estimate the following risks for yourself?

    *a*) Getting infected with the coronavirus

    1) Low risk
    2) Reasonable risk
    3) High risk
    4) Extremely high risk

    *b*) Becoming very ill after being infected by the coronavirus

    1) Low risk
    2) Reasonable risk
    3) High risk
    4) Extremely high risk

    *c*) Having to be admitted to hospital after being infected by the coronavirus

    1) Low risk
    2) Reasonable risk
    3) High risk
    4) Extremely high risk

    *d*) Dying through being infected by the corona virus

    1) Low risk
    2) Reasonable risk
    3) High risk
    4) Extremely high risk

12. How do you estimate the risk for at least one person in your direct environment (family in your household, other family, friends):

    *a*) .

    1) Low risk
    2) Reasonable risk
    3) High risk
    4) Extremely high risk

    *b*) Becoming very ill after being infected by the coronavirus

    1) Low risk

  2) Reasonable risk
  3) High risk
  4) Extremely high risk

*c*) Having to be admitted to hospital after being infected by the coronavirus

  1) Low risk
  2) Reasonable risk
  3) High risk
  4) Extremely high risk

*d*) Dying through being infected by the corona virus

  1) Low risk
  2) Reasonable risk
  3) High risk
  4) Extremely high risk

13. How old are you?

  *a*) 18 - 25 years old
  *b*) 26 - 35 years old
  *c*) 36 - 45 years old
  *d*) 46 - 55 years old
  *e*) 56 - 65 years old
  *f*) 66 - 74 years old
  *g*) Above 75 years old

14. What is you highest level of education?

  *a*) No formal education
  *b*) Junior school/primary education
  *c*) Lower Vocational Education (trade school, domestic science school, lower technical school, lower economics and admin education, etc.)
  *d*) Pre-vocational Secondary Education / Advanced Primary Education (MAVO, VMBO, MULO)
  *e*) Higher General Secondary Education / Preparatory Academic Education (HAVO/VWO)

*f*) Secondary Vocational Education (MBO)

*g*) Higher Vocational Education (HBO)

*h*) University

15. What is your current living arrangement?

*a*) I live alone

*b*) I live with my partner

*c*) I live with my partner and child(ren)

*d*) I live with a child / children

*e*) I live with roommates

*f*) Others

16. Which province do you live in?

*a*) Groningen

*b*) Friesland

*c*) Drenthe

*d*) Overijssel

*e*) Flevoland

*f*) Gelderland

*g*) Utrecht

*h*) North-Holland

*i*) South-Holland

*j*) Zeeland

*k*) North-Brabant

*l*) Limburg

17. What is the net monthly income of your household? This is the total amount from salary, benefits, grants and pensions that your household receives every month.

*a*) Less than 1000 Euros

*b*) Between 1000 and 2000 Euros

  *c*) Between 2000 and 3000 Euros
  *d*) Between 3000 and 4000 Euros
  *e*) Between 4000 and 5000 Euros
  *f*) Between 5000 and 6000 Euros
  *g*) Between 6000 and 7000 Euros
  *h*) More than 7000 Euros
  *i*) I would rather not answer this question
  *j*) I don't know

18. How do you expect your household's net income to change in 2020?

    *a*) I expect our net income to strongly decrease
    *b*) I expect our net income to decrease
    *c*) I expect our net income to remain the same
    *d*) I expect our net income to increase
    *e*) I expect our net income to strongly increase

19. What is your current work situation (more than one answer is possible)?

    *a*) Paid work on a fixed contract
    *b*) Paid work on a temporary contract
    *c*) Freelancer / business owner
    *d*) Voluntary worker
    *e*) Pensioner
    *f*) Out of work / looking for work
    *g*) Unfit for work
    *h*) Receiving welfare benefits
    *i*) Housewife / House husband
    *j*) I am following a course / studying
    *k*) None of the above

20. What is your profession? (only answer this question if you chose one of the top three options listed above as an answer to the previous question)

21. What is your current work situation like? (only answer this question if you filled in a profession for the previous question)

    *a*) My work is currently at a standstill

    *b*) Currently, I have less work

    *c*) My work continues as normal

    *d*) I currently have more work

22. What did you feel were the strong points about this method?

23. What do you think could be done to improve this method?

## Thank you

Thank you very much for taking part in this consultation!

If you would like more information about this consultation and the method that we use for these consultations, visit www.tudelft.nl/covidexit/. On this website we also publish the results of this research.

You can send any feedback about this research by email to n.mouter@tudelft.nl

# Bibliography

Acemoglu, D., J. Robinson (2012) *Why Nations Fail: The Origins of Power, Prosperity, and Poverty*, Currency.

Aldrich, D., M. Meyer (2015) Social capital and community resilience, *The American behavioral scientist*, 59(2), pp. 254–69.

Allcott, H., L. Boxell, J. Conway, M. Gentzkow, M. Thaler, D. Yang (2020) Polarization and public health: Partisan differences in social distancing during the Coronavirus pandemic, *Journal of public economics*.

Almaatouq, A. (2020) Towards stable principles of collective intelligence under an environment-dependent framework.

Almaatouq, A., A. Noriega-Campero, A. Alotaibi, P. Krafft, M. Moussaid, A. Pentland (2020) Adaptive social networks promote the wisdom of crowds, *Proceedings of the National Academy of Sciences*, 117(21), pp. 11379–86.

Aragones, E., S. Sanchez-Pages (2009) A theory of participatory democracy based on the real case of Porto Alegre", *European Economic Review*, 53, pp. 56–72.

Authority, D. H. C. (2020) Heropstart van zorg: eenvoudige rekensommetjes, maar lastige afwegingen.

Bartkowski, B., N. Lienhoop (2017) Democracy and valuation: A reply to Schläpfer (2016, *Ecological economics*, 131, pp. 557–60.

Bernier, H. (2014) Lessons learned from implementing a multi-year, multi-project public engagement initiative to better inform governmental public health policy decisions, *J Participat Med May*, 22(68).

Bhat, C. R. (2008) The multiple discrete-continuous extreme value (MDCEV) model: Role of utility function parameters, identification considerations, and model extensions, *Transportation Research Part B: Methodological*, 42(3), pp. 274–303.

Blendon, R., L. Koonin, J. Benson, M. Cetron, W. Pollard, E. Mitchell (2008) Public response to community mitigation measures for pandemic influenza, *Emerging infectious diseases*, 14(5), pp. 778–86.

Bol, D., M. Giani, A. Blais, P. J. Loewen (2021) The effect of COVID-19 lockdowns on political support: Some good news for democracy?, *European Journal of Political Research*, 60(2), pp. 497–505.

Cabannes, Y. (2004) Participatory budgeting: A significant contribution to participatory democracy", *Environment and Urbanization*, 16(1), pp. 27–46.

Cairney, P. (2016) *The Politics of Evidence Based Policymaking*, Palgrave, London.

Capano, G., M. Howlett, D. Jarvis, M. Ramesh, N. Goyal (2020) Mobilizing policy (in)capacity to fight COVID-19: Understanding variations in state responses, *Policy & society*, 39(3), pp. 285–308.

Carson, R. (2012) *Contingent Valuation: A Comprehensive Bibliography and History*, Edward Elgar Publishing.

Carson, R., T. Groves (2007) Incentive and informational properties of preference questions, *Environ Resour Econ (Dordr*, 37(1), pp. 181–210.

Carson, R., W. Hanemann (2005) Contingent valuation, in: *Handbook of Environmental Economics. Elsevier*, pp. 821–936.

Carson, R., R. Mitchell, M. Hanemann, R. Kopp, S. Presser, P. Ruud (2003) Contingent valuation and lost passive use: Damages from the Exxon Valdez oil spill, *Environmental and Resource Economics*, 25(3), pp. 257–286.

Centers for Disease Control (2020) The Public engagement project on community control measures for pandemic influenza; findings and recommendations from citizen and stakeholder deliberation days.

Chorus, C., E. Sandorf, N. Mouter (2020) Diabolical dilemmas of COVID-19: An empirical study into Dutch society's trade-offs between health impacts and other effects of the lockdown, *PLoS One*, 15(9).

Chuang, Y.-C., Y.-L. Huang, K.-C. Tseng, C.-H. Yen, L.-H. Yang (2015) Social capital and health-protective behavior intentions in an influenza pandemic, *PLoS One*, 10(4).

Coroiu, A., C. Moran, T. Campbell, A. Geller (2020) Barriers and facilitators of adherence to social distancing recommendations during COVID-19 among a large international sample of adults, *PloS one*, 15(10).

Czajkowski, M., N. Hanley, J. LaRiviere (2015) The effects of experience on preferences: Theory and empirics for environmental public goods, *American journal of agricultural economics*, 97(1), pp. 333–51.

Dekker, T., P. Koster, N. Mouter (2019) The economics of participatory value evaluation.

Delgado, A., K. Lein Kjølberg, F. Wickson (2011) Public engagement coming of age: From theory to practice in STS encounters with nanotechnology, *Public understanding of science (Bristol, England)*, 20(6), pp. 826–45.

Deliberative democracy consortium. (2020) Deliberative democracy consortium, https://deliberative-democracy.net/.

Dietz, T., P. Stern, A. Dan (2009) How deliberation affects stated willingness to pay for mitigation of carbon dioxide emissions: An experiment, *Land economics*, 85(2), pp. 329–47.

Dostal, J. (2020) Governing under pressure: German policy making during the Coronavirus crisis, *The Political quarterly*, 91(3), pp. 542–52.

Driscoll, J., K. Sonin, J. Wilson, A. L. Wright (2020) Poverty and Economic Dislocation Reduce Compliance with Covid-19 Shelter-in-Place Protocols.

Dryzek, J., A. Bächtiger, S. Chambers, J. Cohen, J. Druckman, A. Felicetti (2019) The crisis of democracy and the science of deliberation, *Science (New York, N.Y.)*, 363(6432), pp. 1144–6.

Dryzek, J., S. Niemeyer (2008) Discursive representation, *American political science review*, 120(4), pp. 481–492.

Dynes, R. (2006) Social capital: Dealing with community emergencies, *Homeland Security Affairs*, 2(2), pp. 1–26.

Esaiasson, P., M. Gilljam, M. Persson (2017) Responsiveness beyond policy satisfaction: Does it matter to citizens?, *Comparative political studies*, 50(6), pp. 739–65.

Farrell, H. (2009) *The Political Economy of Trust: Institutions, Interests, and Inter-Firm Cooperation in Italy and Germany*, Cambridge University Press.

Fiorino, J. (1990) Citizen participation and environmental risk: a survey of institutional mechanisms, *Science, Technology and Human Values*, 15, pp. 226–243.

Frey, B., A. Stutzer (2000) Happiness, economy and institutions, *Econ J (London*, 110(466), pp. 918–38.

Fung, A., O. Wright (2003) Thinking about empowered participatory governance.

Gilbert, M., M. Dewatripont, E. Muraille, J.-P. Platteau, M. Goldman (2020) Preparing for a responsible lockdown exit strategy, *Nature medicine*, 26(5), pp. 643–4.

Goodin, R. (2000) Democratic deliberation within, *Philosophy & public affairs*, 29(1), pp. 81–109.

Goodin, R., J. Dryzek (2006) Deliberative impacts: The macro-political uptake of mini-publics, *Politics and society*, 34(2), pp. 219–244.

Guardian, T. (2020) Surge in domestic violence during COVID-19 crisis.

Gutmann, A., D. Thompson (2004) *Why Deliberative Democracy?*, Princeton University Press.

Haab, T. C., K. E. McConnell (2002) *Valuing Environmental and Natural Resources: The Econometrics of Non-Market Valuation*, Edward Elgar Publishing.

Halkos, G., A. Leonti, E. Sardianou (2020) Assessing the preservation of parks and natural protected areas: A review of Contingent Valuation studies, *Sustainability*, 12(11).

Hanley, N., S. Mourato, R. Wright (2001) Choice modelling approaches: A superior alternative for environmental valuation?, *Journal of Economic Surveys*, 15, pp. 435–62.

Harari, Y. (2020) The world after coronavirus, *Financial Times*.

Hartley, K., D. Jarvis (2020) Policymaking in a low-trust state: Legitimacy, state capacity, and responses to COVID-19 in Hong Kong, *Policy & society*, 39(3), pp. 403–23.

Hendriks, C., J. Lees-Marshment (2019) Political leaders and public engagement: The hidden world of informal elite–citizen interaction, *Politische Studien*, 67(3), pp. 597–617.

Hensher, D. A., J. M. Rose, J. M. Rose, W. H. Greene (2005) *Applied Choice Analysis: A Primer*, Cambridge University Press.

Johnston, R., K. Boyle, A. W, J. Bennett, R. Brouwer, T. Cameron (2017) Contemporary guidance for stated preference studies, *Journal of the Association of Environmental and Resource Economists*, 4(2), pp. 319–405.

Kingsley, P. (2020) Serbia protests meet violent response in Europe's 1st major virus unrest.

Klaassen, N. (2020) Mag het noorden als eerste uit de lockdown?

Koskimaa, V., L. Rapeli (2020) Fit to govern? Comparing citizen and policymaker perceptions of deliberative democratic innovations, *Policy & Politics*, 48(4), pp. 637–652.

Lancaster, K. J. (1966) A New Approach to Consumer Theory, *Journal of Political Economy*, 74(2), pp. 132–157.

Lavazza, A., M. Farina (2020) The role of experts in the Covid-19 pandemic and the limits of their epistemic authority in democracy, *Frontiers in public health*, 8(356).

Lee, S., C. Hwang, M. Moon (2020) Policy learning and crisis policy-making: Quadruple-loop learning and COVID-19 responses in South Korea, *Policy & society*, 39(3), pp. 363–81.

Lewis-Kraus, L. (2020) Why have some Asian countries controlled their outbreaks so well? It's because authorities have earned their citizens' confidence.

Liscio, E., M. Meer, L. Siebert, C. Jonker, N. Mouter, P. Murukannaiah (2021) Identifying and evaluating context-specific values, in: *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021*, vol. 10, pp. , 3–7,.

Liu, Y., A. Boin (2020) Framing a mega-disaster: Political rhetoric and the Wenchuan earthquake, *Safety science*, 125(104621).

Lupia, A., J. Matsusaka (2004) Direct Democracy: New approaches to old questions, *Annual Review of Political Science, Vol 13*, 7(1), pp. 463–82.

Mansbridge, J. (1997) Taking coercion seriously, *Constellations (Oxford, England)*, 3(3), pp. 407–16.

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

Moon, M. (2020) Fighting against COVID-19 with agility, transparency, and participation: Wicked policy problems and new governance challenges, *Public Administration Review*.

Mouter, N., M. Cabral, T. Dekker, S. Cranenburgh (2019) The value of travel time, noise pollution, recreation and biodiversity: A social choice valuation perspective, *Res Transp Econ*, 76(100733).

Mouter, N., P. Koster, T. Dekker (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, pp. 54–73.

Mouter, N., R. M. Shortall, S. L. Spruit, A. V. Itten (2021b) 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, p. 101965.

Mudde, C. (2020) Wartime" coronavirus powers could hurt our democracy – without keeping us safe.

NIBUD (2020) Een vijfde van de Nederlanders ervaart inkomensterugval.

NLTimes (2019) Over 700,000 Dutch domestic violence victims in 5 years, *February*.

OECD (2020) The territorial impact of COVID-19: Managing the crisis across levels of government, https://www.oecd.org/coronavirus/policy-responses/the-territorial-impact-of-COVID-19-managing-the-crisis-across-levels-of-government-d3e314e1/.

Offe, C. (2017) Referendum vs. Institutionalized deliberation: What democratic theorists can learn from the 2016 Brexit decision, *Daedalus*, 146(3), pp. 14–27.

Pal, M. (2012) The promise and limits of citizens' assemblies: Deliberation, institutions and the law of democracy, *Queen's LJ*, 38(259).

Partnership, O. G. (2020) Collecting open government approaches to COVID-19, *Opengovpartnership.org*.

Pearse, H. (2020) Deliberation, citizen science and covid-19, *The Political quarterly*, 91(3), pp. 571–7.

Povoledo, E., R. Minder, I. Kwai (2020) Protesters in Italy and Spain clash with police as they call for 'freedom' from virus restrictions.

Rabobank (2020) Knowledge website, https://economics.rabobank.com/publications/2020/april/dutch-economy-to-contract-this-year-morethan-in-2009/.

Reckers-Droog, V., J. van Exel, W. Brouwer (2018) Who should receive treatment? An empirical enquiry into the relationship between societal views and preferences concerning healthcare priority setting, *PLOS ONE*, 13(6), p. e0198761.

Roth, K. (2020) How authoritarians are exploiting the covid-19 crisis to grab power [Internet, *Nybooks.com*.

Rowe, G., L. Frewer (2005) *A Typology of Public Engagement Mechanisms*, vol. 30, Science, Technology, & Human Values, 251–290 pp.

Ryan, M., S. G. (2014) *Deliberative Mini-Publics: Involving Citizens in the Democratic Process, 9-26*, Ecpr Press.

Sabat, I., S. Neuman-Böhme, N. Varghese, P. Barros, W. Brouwer, J. Exel (2020) United but divided: Policy responses and people's perceptions in the EU during the COVID-19 outbreak, *Health Policy [Internet*.

Schoch-Spana, M., A. Chamberlain, C. Franco, J. Gross, C. Lam, A. Mulcahy (2006) Disease, disaster, and democracy: The public's stake in health emergency planning, *Biosecurity and bioterrorism : biodefense strategy, practice, and science*, 4(3), pp. 313–9.

Schoch-Spana, M., C. Franco, J. Nuzzo, C. Usenza (2007) Working Group on Community Engagement in Health Emergency Planning. Community engagement: Leadership tool for catastrophic health events, *Biosecurity and bioterrorism : biodefense strategy, practice, and science*, 5(1), pp. 8–25.

Sintomer, Y., C. Herzberg, A. Rocke (2008) Participatory budgeting in europe: Potentials and challenges", *International Journal of Urban and Regional Research*, 32(1), pp. 164–178.

Spruit, S., N. Mouter (2020) 1795 inwoners over het toekomstige energiebeleid van Regio Foodvalley.

Spruit, S. L., N. Mouter, L. Kaptein, P. Ytsma, W. Gommans, M. Collewet, N. Van Schie, A. Karmat, M. Knip (2020) 1376 inwoners van Súdwest-Fryslân over het toekomstige energiebeleid van hun gemeente: De uitkomsten van een raadpleging.

SteelFisher, G., R. Blendon, J. Ward, R. Rapoport, E. Kahn, K. Kohl (2012) Public response to the 2009 influenza A H1N1 pandemic: A polling study in five countries, *The Lancet. Infectious diseases*, 12(11), pp. 845–50.

Study, T. B. C. (2020) The Big Corona Study.

Train, K. E. (2009) *Discrete Choice Methods with Simulation*, Cambridge University Press.

Tufekci, Z. (2020) How hong kong did it.

Vaughan, E., T. Tinker (2009) Effective health risk communication about pandemic influenza for vulnerable populations, *American journal of public health*, 2(S2).

Warren, M., J. Mansbridge (2013) *Deliberative Negotiation*, American Political Science Association.

Webster, N. (2020) Public discussions on COVID-19 lockdown in scotland, https://medium.com/participo/public-discussions-on-COVID-19-lockdown-in-scotland-8f34a586c69c.

Weible, C., D. Nohrstedt, P. Cairney, D. Carter, D. Crow, A. Durnová (2020) COVID-19 and the policy sciences: Initial reactions and perspectives, *Policy sciences*, 53(2), pp. 1–17.

Wouters, S., J. Exel, W. Brouwer, R.B. (2017) Priority to end of life treatments? Views of the public in the Netherlands, *Value in Health*, 20(1), pp. 107–117.

Xafis, V. (2020) What is Inconvenient for You is Life-saving for Me': How Health Inequities are playing out during the COVID-19 Pandemic, *Asian Bioethics Review*, 1.

# Chapter 3

# Data-driven methods to assist choice models for Participatory Value Evaluation experiments

- 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.

We propose three procedures based on association rules (AR) learning and random forests (RF) to support the specification of a portfolio choice model applied in data from complex choice experiment data, specifically a Participatory Value Evaluation (PVE) choice experiment. In a PVE choice experiment, respondents choose a combination of alternatives, subject to a resource constraint. We combine a methodological-iterative (MI) procedure with AR learning and RF models to support the specification of parameters of a portfolio choice model. Additionally, we use RF model predictions to contrast the validity of the behavioural assumptions of different specifications of the portfolio choice model. We use data of a PVE choice experiment conducted to elicit the preferences of Dutch citizens for lifting COVID-19 measures. Our results show model fit and interpretation improvements in the portfolio choice model, compared with conventional model specifications. Additionally, we provide guidelines on the use of outcomes from AR learning and RF models from a choice modelling perspective.

## 3.1. Introduction

In the last years, Participatory Value Evaluation (PVE) choice experiments have become an alternative to capture more complex and realistic forms of human decision making in diverse fields (Mouter et al., 2020; Rotteveel et al., 2022; Mulderij et al., 2021). PVE is a preference elicitation framework based in a portfolio choice experiment (Wiley & Timmermans, 2009), in which respondents select their preferred set of alternatives, subject to one or more resource constraints (Mouter et al., 2021). In the PVE choice experiment, respondents face a set of available alternatives, the attributes and costs of each alternative, and the available resources. Then, respondents must choose a combination of alternatives (if any), without violating the constraints. As in recently developed experiments (Caputo & Lusk, 2022; Carson et al., 2022; Neill & Lahne, 2022), a PVE choice experiment is an extension of the discrete choice experiment (DCE) approach that provides a more realistic experimental setting for choice situations where a multiple choice subject to constraints is required (e.g., policy makers deciding to fund certain policies with a scarce budget).

While PVE choice experiments offer a more realistic experimental setting than a conventional DCE, specifying choice models to analyse data from such experiments is challenging. Hitherto, choice models developed to analyse PVE choice experiments data (Dekker et al., 2019; Bahamonde-Birke & Mouter, 2019) have been built to address multiple-discrete (portfolio) choices, the presence of resource constraints and interaction effects when two or more alternatives are chosen together (Bahamonde-Birke & Mouter, 2019). However, the specification process of these models usually relies on prior knowledge from the analyst concerning, for example, how respondents derive utility, how attributes interact (e.g., linear-in-parameters specification), the respondents' decision rule, what interactions between alternatives are relevant to include, etc. Furthermore, finding a proper model specification usually involves a trial-and-error procedure, in which several candidate specifications are tested and the most parsimonious or plausible model is chosen. This process is already cumbersome for discrete choice models (Ortelli et al., 2021), but for more complex choice models, and models for PVE choice experiments data in particular, even more so. The presence of considerably more variables, possible combinations of chosen alternatives, and potential interactions effects between alternatives impose more complexity in the specification of a choice model for PVE choice experiments data, with the consequently longer estimation times than for a discrete choice model, namely from the range of minutes for a simple specification, to an hour in more complex cases.

In the last years, there has been an increasing interest on assisting the specification of choice models with data-driven methods. Data-driven methods (e.g., machine learning, data mining) are methodological approaches that aim to identify relevant patterns and/or learn the underlying data-generating process (DGP) directly from the data. Recent studies have shown that data-driven methods can complement the toolbox of choice modellers (see, for example van Cranenburgh et al., 2022; Sifringer et al., 2020; Wang et al., 2020), or provide further insights for researchers, without explicitly using choice models (e.g., Keuleers et al., 2001; van Cranenburgh & Kouwenhoven, 2020). Furthermore, specific approaches based in data-driven methods to assist the specification of discrete choice models have been recently proposed in literature (Ortelli et al., 2021; Hillel et al., 2019; Shiftan & Bekhor, 2020). However, to the authors' knowledge, no studies have explored methods to assist the specification of choice models to analyse data from more complex type of choice experiments than a DCE, and particularly from PVE choice experiments, or they explored potential insights obtained from using this methods with PVE choice experiments data.

In this paper, we propose three procedures to assist the specification of choice models for PVE choice experiments based in two data-driven methods, and we provide insights on the interpretation of the outcomes of such methods a choice modelling perspective. The first method is Association Rules (AR) learning (Agrawal et al., 1993); a data mining approach used to identify frequent interactions between the variables of a dataset in terms of a set of empirical relational statistics. Applications of AR learning in areas where choice models are standard methods can be seen in the works of (Keuleers et al., 2001), Geurts et al. (2003) and (Kaur & Kang, 2016), but solely focused on gathering association rules between explanatory variables of choice data. We use AR learning to gather association rules between chosen alternatives of the PVE choice experiment that can be interpreted as relevant interactions made by respondents. The second method is a Random Forest (RF) model (Breiman, 2001); a predictive machine learning model built from an ensemble of decision tree models. RF models can model complex relationships from the data, while yet providing a degree of interpretability through the computation of variable importances. We build upon the works of Hillel et al. (2019), Yao & Bekhor (2020) and Shiftan & Bekhor (2020), and we propose two methodological-iterative (MI) procedures to specify the parameters of the utility functions of a portfolio choice model applied in PVE experiments data (Bahamonde-Birke & Mouter, 2019) in a structured way, based on the outcomes of AR learning and RF models, respectively. Finally, we propose a procedure to test the validity of the behavioural assumptions of different specifications of the portfolio choice model, based on comparing their ranking of combinations of alternatives with highest choice probability with the ranking obtained from a RF model.

For our analyses we use data from a PVE choice experiment to elicit the preferences of Dutch citizens for relaxing COVID-19 restrictions after the first wave of the

Coronavirus pandemic (Mouter et al., 2021). In this experiment, respondents were asked to choose their preferred package of COVID-19 restrictions to be relaxed from eight options, such that a constraint of pressure to the healthcare system is not violated. On the one hand, relaxing COVID-19 restrictions may lead to increasing deaths due to COVID-19; on the other hand, it can provide psychological relief and reduce economic losses. Interactions between individual relaxations are reasonably expectable in this PVE choice experiment, as well as differences in terms of the preferences for an impact among different options. Furthermore, it is reasonable to expect the existence of complex interactions that are difficult to uncover from a choice model. In fact, analyses of written arguments to make a choice in this PVE choice experiment suggest the existence of semi-compensatory and lexicographic choice behaviour in a significant amount (Mouter et al., 2021). In that sense, a more agnostic approach (in terms of behavioural assumptions), such as a RF model can be more appropriate for prediction purposes than a choice model.

This paper is organised as follows. Section 3.2 details the PVE choice experiment data preparation and data description. Section 3.3 formalises AR learning, RF models, the portfolio choice mode and the procedure to assist the specification of choice models. Section 3.4 presents the results. Section 3.5 concludes and provides a discussion of our findings and further research directions.

## 3.2. Data

### 3.2.1. The COVID-19 PVE choice experiment data

We use data from a PVE choice experiment conducted to elicit the preferences of Dutch citizens to relax COVID-19 measures in the Netherlands (Hernandez et al., 2021)[1], henceforth the COVID-19 PVE choice experiment. In this experiment, respondents were asked to choose which COVID-19 restrictions should be relaxed, without surpassing a maximum level of pressure increase to the healthcare system. Respondents faced eight relaxation options (alternatives):

1. Nursing and care homes allow visitors (NH),
2. Re-open businesses, other than contact professions and hospitality industry (RB),
3. Re-open contact professions (RC),
4. Young people may come together in small groups (YP),

[1] The dataset is available from https://doi.org/10.4121/14413958.v1

5. All restrictions lifted for people with immunity (LI),

6. All restrictions lifted in Northern provinces (LN),

7. Direct family members from other households can have social contact (DF),

8. Re-open hospitality and entertainment industry (RH).

Choosing an alternative generated an additional (percentage) pressure to the healthcare system. Respondents cannot surpass an increase of 50% of pressure to the healthcare system. In addition, each alternative was characterised by five attributes: a) additional deaths of people with 70 or more years old, b) additional deaths of people with less than 70 years old, c) additional cases of (permanent) physical injury, d) reduction of cases of (permanent) mental injury, and e) reduction of households with severe income loss. A more detailed description of the design of this experiment is presented in the work of Mouter et al. (2021). The choices of a PVE choice experiment can be represented in a matrix of size $N \times J$, as illustrated in table 3.1. Each row is a choice situation (respondent) from 1 to $N$, while each column is an alternative from 1 to $J$. A choice in a PVE choice experiment is a combination of choices among the $J$ alternatives, represented by ones (if chosen) and zeros (if not chosen).

| Participant ID | NH | RB | RC | YP | ... | RH |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 0 | ... | 1 |
| 2 | 1 | 1 | 1 | 0 | ... | 1 |
| 3 | 0 | 0 | 0 | 1 | ... | 1 |
| ... | | | | | | |
| N | 0 | 0 | 0 | 0 | ... | 0 |

*Table 3.1: Example of a choice matrix of a PVE choice experiment*

## 3.2.2. Data preparation and description

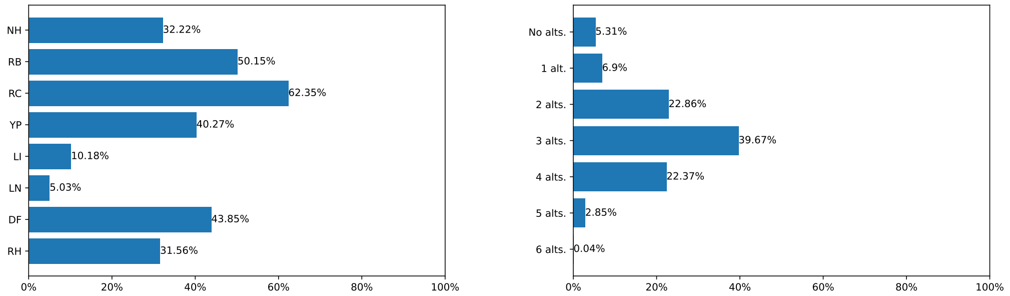
The COVID-19 PVE choice experiment dataset contains 29,669 responses and 57 variables. Table 3.2 provides a definition of the variables of this dataset. First, we define the choice indicators in two forms: eight individual indicators per alternative (`Choice_` from 1 to 8) used by AR learning and the choice model, and a single variable that uniquely identifies a chosen combination of alternatives used by the RF model (Y_index) ranging from 0 to 181. Second, we define the attributes of each alternative as a set of numeric variables, per alternative and per attribute (Pressure_ to Minus_HH_incloss_ 1 to 8). Except for pressure to the healthcare system, all attributes are scaled by 10,000 to avoid numerical overflow issues in the estimation/-training routines of the portfolio choice model.

| Variable | Description |
|---|---|
| Choice_ 1 to 8 | Binary choice indicator (= 1 if alternative is chosen) |
| Y_index | Unique index of chosen combination of alternatives (from 0 to 181) |
| Pressure_ 1 to 8 | Additional pressure to the healthcare system |
| Deaths_70plus_ 1 to 8 | Additional deaths of people of 70+ years old |
| Deaths_less70_ 1 to 8 | Additional deaths of people of less than 70 years old |
| Plus_physical_injury_ 1 to 8 | Additional people with (permanent) physical injury |
| Minus_mental_injury_ 1 to 8 | Decrease of people with (permanent) mental injury |
| Minus_HH_incloss_ 1 to 8 | Decrease of househols with severe income loss |

*Table 3.2: Variables of the COVID-19 PVE choice experiment dataset*

Figure 3.1 summarises the market shares (a) and the distribution of the number of chosen alternatives (b) the dataset. The most chosen alternatives are re-opening contact professions (RC) and other businesses (RB), with 62.4% and 50.1%, respectively; in contrast, lifting all restrictions for immune people (LI) and in the Northern provinces (LN) are the least chosen alternatives with 10.2% and 5% respectively. The vast majority of respondents choose between two and four alternatives (more than 80% of respondents). As expected, no respondents choose more than six alternatives due to the existence of a resource constraint. On the other hand, 5.3% of respondents choose no alternative at all (no choice). While this percentage is rather low if taken as a dropout measure (i.e., respondents who did not answer the choice experiment), it is considerably higher than the probability of randomly choosing any combination of alternatives of the dataset ($1/182$).

In addition to the empirical data, we create four datasets with pseudo-synthetic choices. Pseudo-synthetic datasets are generated to corroborate if our proposed methods are able to recover the true data-generating process and/or identify interactions included *a priori* in the data. For instance, we use pseudo-synthetic data to test whether the metrics of AR learning are aligned with the inclusion of explicit interactions. Pseudo-synthetic datasets are generated by using the experimental design of the the COVID-19 PVE choice experiment data to generate synthetic choices, assum-



*(a) Market shares* *(b) Number of chosen alternatives.*

*Figure 3.1: Descriptive statistics of choice variables, COVID-19 PVE choice experiment.*

ing a previously known DGP and "true" parameters. We provide a detail of the dataset generation process and parametrisation in appendix 3.A.

# 3.3. Methods

## 3.3.1. Association rules learning

AR learning is a data mining method that aims to identify frequent relationships between variables of a transactions dataset (Agrawal et al., 1993). In an AR learning application, an algorithm scans combinations of variables in the dataset named as *itemsets*, keeping only the combinations that satisfy a minimum *support* (relative frequency) threshold defined by the analyst. Then, a set of association rules of the form $A \Rightarrow B$, with $A$ and $B$ itemsets, are constructed, considering only those rules with a *confidence* (conditional frequency) higher than a threshold defined by the analyst. By doing so, AR learning rules out combinations of alternatives that scarcely appear in the dataset. In this paper, we use the seminal approach of AR learning provided by Agrawal et al. (1993) to identify association rules between combinations of discrete alternatives using the "Apriori" algorithm.

We proceed to formalise AR learning. Consider a transactions dataset $D$ with $N$ rows and $J$ columns. Each row $n \in \{1,...,N\}$ of the dataset is a transaction over $J$ items (columns). Each transaction is represented as a vector $y_n = \{y_{n1}, y_{n2}, \ldots, y_{nJ}\}$, in which each variable $y_{nj}$ is a binary indicator that is equal to one if item $j \in \{1,...,J\}$ is selected, and zero otherwise. Some examples of transaction datasets are supermarket purchase data, accesses to webpages, etc. In this paper, we treat the choice data of the PVE choice experiment as a transaction dataset, in which each transaction is a choice situation over $J$ alternatives.

Define an itemset as a subset of items of the dataset. For example, $A = \{y_{n1}, y_{n2}\}$, $B = \{y_{n3}, y_{n4}, y_{n5}\}$ and $C = \{y_{n1}, y_{n2}, y_{n4}\}$ are itemsets of the dataset $D$. An association rule between itemsets $A$ and $B$ is a directional implication of the form $A \Rightarrow B$, with $A \cap B = \emptyset$, in which $A$ is defined as the *antecedent* and $B$ is the *consequent*. If itemsets $A$ and $B$ are two (disjoint) combinations of alternatives, the rule $A \Rightarrow B$ can be interpreted as "if combination of alternatives $A$ is chosen, then combination of alternatives $B$ is chosen".

The problem of AR learning is to find all the itemsets that satisfy a minimum *support* threshold, and then generate all the association rules that satisfy a minimum *confidence* threshold. The *support* $supp(A)$ of an itemset $A$ as the relative frequency that $A$ appears in dataset $D$. The support of an itemset $A$ can be interpreted as the probability $P(A)$ on the domain of the dataset.

The *confidence* of an association rule $A \Rightarrow B$ is defined as:

$$conf(A \Rightarrow B) = \frac{supp(A \cup B)}{supp(A)} \tag{3.1}$$

Confidence is the percentage of transactions of itemset $A$ that also contain itemset $B$. In the context of PVE choice experiments data, a support of $A$ equal to $s$ is interpreted as "$s * 100\%$ of the choices of the dataset involve the combinations described in $A$", whereas a confidence of the rule $A \Rightarrow B$ equal to $c$ is interpreted as "$c * 100\%$ of the choices that involve $A$ also involve $B$". The confidence of $A \Rightarrow B$ can be interpreted as the conditional probability of $B$ given $A$.

While support and confidence measure how often an itemset or rule appear in the dataset, they do not provide information about the degree of dependence of the components of a rule. Thus, AR learning can generate trivial rules with high confidence and support for itemsets $A$ And $B$, even if such itemsets have a small or no dependence (Keuleers et al., 2001). In light of this, computing the *lift* of the association rules is recommended. The lift of an association rule $A \Rightarrow B$ measures the degree of dependence between $A$ and $B$ as:

$$lift(A \Rightarrow B) = \frac{supp(A \cup B)}{supp(A) \cdot supp(B)}. \tag{3.2}$$

The lift is a ratio between the support of $A$ and $B$ together, divided by the independent supports of each itemset. If $lift(A \Rightarrow B) > 1$ then $A$ and $B$ are more often to be found in the dataset than if $A$ and $B$ were independent, and viceversa for

$lift(A \Rightarrow B) < 1$.

The final outcome of an AR learning application is a list of association rules described by their support, confidence and lift. The analyst can refine this list according to their research needs. For instance, the analyst may be interested only in rules that contain a certain set of variables in the antecedent, and other sets of variables in the consequent; or finding the rules that are more frequent to appear than expected, by sorting them in terms of lift.

### 3.3.2. Random forests

A RF model (Breiman, 2001) is a supervised machine learning method that benefits of the strengths of three techniques from machine learning: an ensemble of decision tree (DT) models, bootstrap aggregation or bagging, and random feature selection. The process of constructing a RF model is illustrated in figure 3.2. Firstly, a set of $R$ DT models is estimated. A DT is a supervised machine learning method based on partitioning the space of the explanatory variables into finite regions to construct a tree structure that best describes the response variable (Friedman et al., 2001). Secondly, to allow variability among trees, each DT model is trained (estimated) using a bootstrapped sample of the data. Thirdly, the partitions of each individual DT model are done using a random subset of explanatory variables, in order to reduce the correlation between trees. Finally, each tree generates a set of predictions that are averaged to provide the final prediction of the RF model.

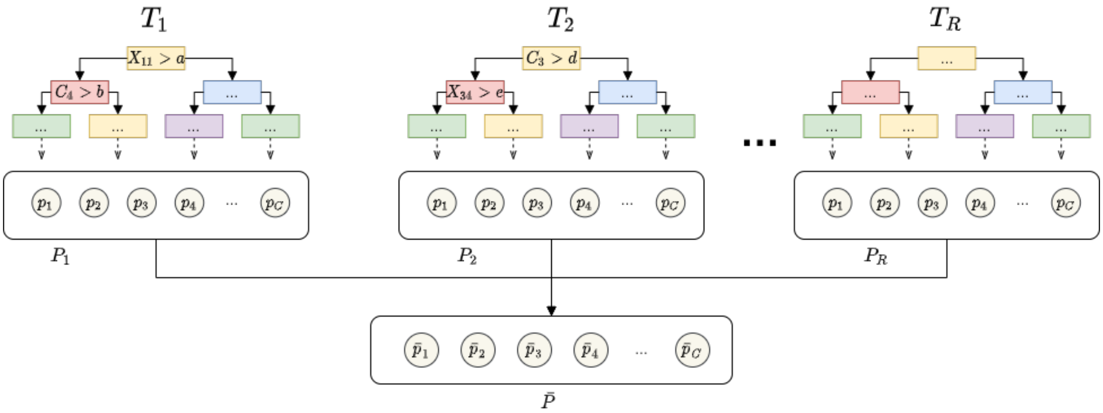


*Figure 3.2: Example of a RF model*

We proceed to formalise the RF model used in this paper. Let $y_n$ be a response variable that uniquely identifies a choice over $J$ alternatives for respondent $n$. $X_n$ is a set of explanatory variables, such as attributes and costs. Define $(\mathbf{y}, \mathbf{X})$ as a dataset of size $N$ where $\mathbf{y} = \{y_1, \ldots, y_N\}$ and $\mathbf{X} = \{X_1, \ldots, X_n\}$. The goal of the RF model is to construct an ensemble of DT models that predicts $\mathbf{y}$ as a function of $\mathbf{X}$.

The RF algorithm is described in figure 3.3. On each DT model $r$, a bootstrapped sample $(\mathbf{X^r}, \mathbf{y^r})$ of the original data is drawn and used to train the individual tree. Each split of the DT model $r$ is done using a random subset of the variables contained in $\mathbf{X^r}$. Finally, each trained DT model $r$ is stored. To make predictions with a RF model, a sample $(\mathbf{X}^*, \mathbf{y}^*)$ -that is different from the sample used for training- is used to predict the choice probabilities of each combination of alternatives among the $R$ trees. The final choice probabilities of the RF model are computed by averaging the predictions of all trees.

**RF Model algorithm:**

**Training:** Let $(\mathbf{X}, \mathbf{y})$ be the training sample.

**1. For** $r = \overline{1,R}$:

- **1a.** Draw a bootstrapped sample $(\mathbf{X^r}, \mathbf{y^r})$ from $(\mathbf{X}, \mathbf{y})$.
- **1b.** Train a DT model using $(\mathbf{X^r}, \mathbf{y^r})$, selecting random $k$ variables from $\mathbf{X^r}$ to do each split of the tree.
- **1c.** Store the DT model as $T_r$.

**Prediction:** Let $(\mathbf{X}^*)$ be a subset of the original explanatory variables, different from the training data.

**1. For** $r = \overline{1,R}$, predict the choice probabilities on each decision tree.

**2.** Compute the final choice probabilities as the average among the $R$ trees.

*Figure 3.3: RF model algorithm*

In addition, RF models can be used to determine the importance of the explanatory variables used on the training process. This is done by computing the mean decrease of impurity of each explanatory variable among the splits (Friedman et al., 2001). The decrease of impurity is the contribution of an explanatory variable on reducing misclassifications in terms of the Gini index (Cheng et al., 2019) on a split of a DT:

$$G(X_i) = \sum_{j=1}^{J} P(X_i = L_j)(1 - P(X_i = L_j)), \tag{3.3}$$

where $X_i$ is the candidate variable for making a split in the RF model, with a possible number of categories $L_1, ...L_J$, and $P(X_i = L_j)$ is the predicted probability of $X_i = L_j$. The mean decrease of impurity of a RF model is the average contribution of each explanatory variable on reducing misclassifications on each tree, and among trees of the RF model. Therefore, higher values of the mean decrease of impurity for a variable $X_i$ imply a major importance of such variable in the RF model, and viceversa.

We identify two considerations on the training and use of results of RF models. The first consideration is the proper selection of so-called hyperparameters. The hyperparameters of the RF model (i.e., the number of DT models, the maximum depth of each tree and the number of variables used per split) can have a significant impact on the final predictions. In light of this, we followed a grid search process to determine the best hyperparameters of the RF model applied to our data. We describe in detail such procedure in appendix 3.B. The second consideration is that variable importances are computed from the training data, which can lead to difficulties to generalise their interpretations. In light of this, the importance measures employed in this paper are computed by using a cross-validation process based on training 100 RF models using random samples (with replacement) of the original data, and averaging the obtained importances of each explanatory variable among all repetitions.

### 3.3.3. The portfolio choice model for PVE choice experiments data

The model we aim to assist its specification is a portfolio choice model for PVE choice experiments data proposed by Bahamonde-Birke & Mouter (2019). This model is an extension to the joint choice model (Lerman, 1976), modified to only consider the choice probabilities of combinations that do not violate the resource constraint. In addition, this model can incorporate interaction parameters that address increases (decreases) of utility when two or more alternatives are chosen at once, interpreted as positive/negative synergies.

We proceed to formalise the portfolio choice model. Let be $N$ respondents of a PVE choice experiment with $J$ alternatives and an available amount of resources of $B$. Each alternative $j \in \{1, \ldots, , J\}$ that respondent $n \in \{1, \ldots, N\}$ is characterised by the unitary cost of resources $c_{nj}$ and the vector of $K$ attributes $X_{nj}$. Each respondent perceives utility from their choice of a combination of alternatives $p$, where $p$ is a number from one to $2^J - U_n$, i.e., the number of possible combinations between alternative choices, minus the number of unfeasible combinations. Additionally, each respondent perceives utility from the amount of non-spent resources.

Following Bahamonde-Birke & Mouter (2019) and assuming only interactions between two alternatives, the utility of choosing a combination $p$ for respondent $n$ is defined by equation (3.4):

$$U_{np} = \begin{cases} \sum_{j=1}^{J} y_{nj} \cdot U_{nj} + \delta_0 \cdot \left(B - \sum_{j=1}^{J} y_{nj} \cdot c_{nj}\right) + \sum_i \sum_j \theta_{ij} y_i y_j + \varepsilon_{np} & \text{, if } \left(B - \sum_{j=1}^{J} y_{nj} \cdot c_{nj}\right) \geq 0 \\ -\infty & \text{, if } \left(B - \sum_{j=1}^{J} y_{nj} \cdot c_{nj}\right) < 0 \end{cases} \tag{3.4}$$

where $y_{nj}$ are binary indicators that are equal to one if respondent $n$ selected alternative $j$ and zero otherwise, $U_{nj}$ is the utility of each individual alternative, $\delta_0$ is a parameter that captures the preference for not spending resources, $\theta_{ij}$ is a parameter that captures the increase (or reduction) of utility when alternatives $i$ and $j$ are chosen together with $i \neq j$, and $\varepsilon_{np}$ is a stochastic error term with an Extreme Value distribution. $U_{nj}$ is defined by equation (3.5):

$$U_{nj} = \delta_j + \beta' X_{njk}, \tag{3.5}$$

where $\delta_j$ are alternative-specific constants and $\beta$ is a vector of parameters associated with the attributes of each alternative.

Assuming that each individual choose the combination of alternatives that maximise his/her utility, the probability of choosing alternative $i$ by respondent $n$ is defined by equation (3.6):

$$P_i = P(U_{ni} \geq U_{np}, \forall p \neq i) = \frac{exp(V_{ni})}{\sum_p exp(V_{np})}, \tag{3.6}$$

where $V_{ni}$ is the observed (non-stochastic) part of the utility function $U_{ni}$. Notice that the choice probabilities take the form of the MNL function, since the utility of a combination of alternatives incorporates an additive Extreme Value stochastic term. Furthermore, the choice probability of an unfeasible combination of alternatives collapse to zero, since $V_{ni} = -\infty$.

### 3.3.4. Assisted specification of the portfolio choice model: methodological-iterative approaches

Shiftan & Bekhor (2020) and Yao & Bekhor (2020)[2] propose a methodological-iterative (MI) approach to assist the specification of a discrete choice model using the

[2] We appreciate the suggestion of one anonymous reviewer on considering this work.

variable importances of a RF model. We build upon these works, and we propose two separate variations of their MI approach, in which we assist the specification of the parameters of a portfolio choice model using the outcomes of AR learning and RF models, respectively.

The first approach, named as MI/AR and detailed in figure 3.4, aims to use the set of association rules with highest and lowest lift values to specify alternative interaction parameters in the portfolio choice model. In the first step, we apply AR learning in the PVE choice experiment data, and we select the *N* rules with highest and lowest lift, with *N* chosen by the analyst. We name the set of rules with highest lift as "group 1", and the set of rules with lowest lift is named as "group 2". The algorithm starts by estimating a portfolio choice model with all the interactions of group 1 specified in the utility functions. Then, the algorithm selects the interaction with the lowest lift value of group 1 and evaluates whether the estimated parameter associated to such interaction is statistically significant. If the parameter is non-significant, the interaction is discarded from the model specification and a new portfolio choice model without the interaction is estimated, otherwise the interaction is kept and the process is repeated until all the interactions of group 1 are considered, in an increasing order in terms of lift. After all interactions of group 1 are considered, the process is repeated for group 2. The algorithm stops when all interactions of both groups are considered.

The second approach, named as MI/RF and illustrated in figure 3.5, aims to use the variable importances of a trained RF model to evaluate the inclusion/exclusion of attribute-specific parameters in the portfolio choice model. In the first step, we train the RF model with the PVE choice experiment data, we calculate the variable importances and we sort them in descending order. Then, we separate the variable importances in two groups: "group 1" contains the attributes with highest variable importances, and "group 2" contains the attributes with lowest variable importances. The algorithm starts by estimating a portfolio choice model with all the attribute-specific parameters of group 1 specified in the utility functions. Then, the exclusion of attribute-specific parameters is determined by their statistical significance, starting by the attributes with lowest importance from group 1. After all the attributes of group 1 are evaluated, the process is repeated for group 2. The algorithm stops when all the attributes of both groups are considered.

### 3.3.5. Using the RF model predictions to test the behavioural assumptions of the portfolio choice model

The final method proposed in this paper is a procedure to test the behavioural assumptions and specification of a portfolio choice model based on the predictions of a

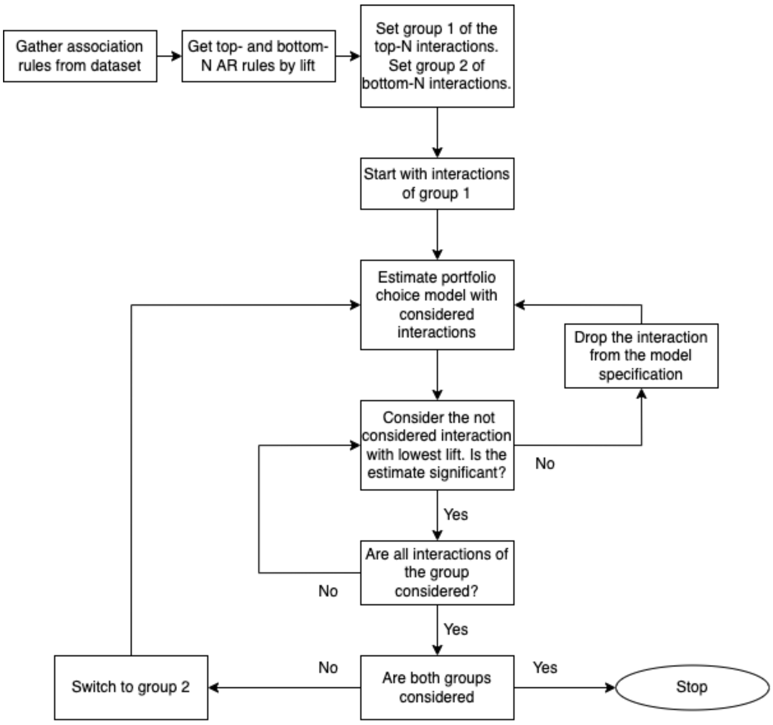

Gather association rules from dataset
Get top- and bottom-N AR rules by lift
Set group 1 of the top-N interactions. Set group 2 of bottom-N interactions.
Start with interactions of group 1
Estimate portfolio choice model with considered interactions
Drop the interaction from the model specification
Consider the not considered interaction with lowest lift. Is the estimate significant?
No
Yes
Are all interactions of the group considered?
No
Yes
Switch to group 2
No
Are both groups considered
Yes
Stop

*Figure 3.4: Methodological-Iterative algorithm for AR learning, based on Shiftan & Bekhor (2020) and Yao & Bekhor (2020).*

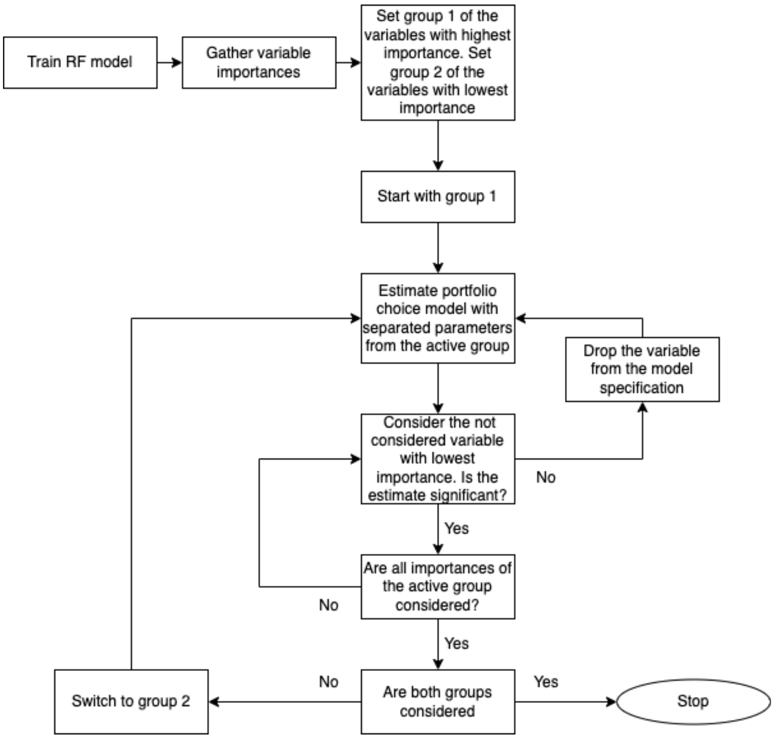

Train RF model
Gather variable importances
Set group 1 of the variables with highest importance. Set group 2 of the variables with lowest importance
Start with group 1
Estimate portfolio choice model with separated parameters from the active group
Drop the variable from the model specification
Consider the not considered variable with lowest importance. Is the estimate significant?
No
Yes
Are all importances of the active group considered?
No
Yes
Switch to group 2
No
Are both groups considered
Yes
Stop

*Figure 3.5: Methodological-Iterative algorithm for RF models, based on Shiftan & Bekhor (2020) and Yao & Bekhor (2020).*

RF model. Specifically, we train the RF model with the PVE choice experiment data, and we compute the ranking of combinations of alternatives with the highest choice probability. This ranking consists of predicting the choice probabilities of all possible combinations of alternatives and sort them by their choice probability in decreasing order. The same ranking is computed from the predictions of a portfolio choice model under different utility specifications, and we evaluate the (dis)similarity between these rankings and the ranking obtained from the RF model.

Figure 3.6 details the procedure to compute the ranking of combinations of alternatives with highest choice probability using a RF model. The procedure is equivalent to compute the ranking using a portfolio choice model, but replacing the RF model by the portfolio choice model. We consider a RF model trained with a sample of the original data, and a prediction (test) sample that differs from the training data. First, we predict the choice probabilities for each combination of alternatives using the prediction sample using the trained RF model. Then, we average the choice probabilities among each choice situation to obtain a single vector of choice probabilities per combination of alternatives. This vector is sorted in descending order and the ranking is constructed by matching each combination of alternatives with their corresponding choice probability. Additionally, the cost of each combination is reported, and in case the portfolio choice experiment considers resource constraints, then combinations that violate such resource constraint are discarded.

To evaluate the (dis)similarity between predicted rankings, we use the Kendall's Tau correlation coefficient (Kendall, 1945). This statistic measures the correlation between two pairs of ranked lists. The statistic is defined as:

$$\tau = \frac{P-Q}{\sqrt{(P+Q+T)(P+Q+U)}} \tag{3.7}$$

where $P$ is the number of concordant pairs between lists, $Q$ is the number of discordant pairs, $T$ the number of ties of the first list and $U$ the number of ties in the second list.

**Ranking of combinations of alternatives with highest choice probability using RF models**

**Start:** Consider a trained RF model. Let $(\mathbf{X}^*, \mathbf{y}^*)$ be a prediction sample.

**1.** Predict the choice probabilities $\hat{\mathbf{y}}$ with the prediction sample using the trained RF model.

**2.** Average the choice probabilities among choice situations (rows).

**3.** Sort the resulting average choice probabilities in descending order.

**4.** Construct the portfolio ranking by matching each combination with their respective choice probability.

**5.** Output the ranked combinations, their choice probabilities, and their total cost of resources.

*Figure 3.6: Algorithm for computing the ranking of combinations of alternatives with highest choice probability*

## 3.4. Results

### 3.4.1. Association rules learning

**Gathering and interpreting association rules from the PVE choice experiment data**

Table 3.3 summarises the found association rules from the COVID-19 PVE choice experiment data. We set low threshold values for support and confidence in order to avoid discarding rules that can be a sign of negative interactions between alternatives. Specifically, we set $minsupport \approx 0$ and $minconfidence = 0$ as support and confidence thresholds, respectively[3]. In total, we find 2100 association rules, with an average confidence of 9%, and average lift of 0.46, with a range between 0.01 and 2 approximately. We expected a low average confidence and large range of lift due to the low thresholds we specify for support an confidence. Additionally, we confirmed that confidence and lift values of association rules are aligned with the inclusion of interactions and unobserved correlation between chosen alternatives in the pseudo-synthetic data. We present the results of these analyses in appendix 3.C

| | Value |
|---|---:|
| Number of rules | 2100 |
| Mean confidence | 0.09 |
| Confidence range | [0.0, 1.0] |
| Mean lift | 0.458 |
| Lift range | [0.013, 1.994] |

*Table 3.3: Summary of found association rules.*

For the purposes of an easier interpretation, we focus on binary (one antecedent and one consequent) association rules, and we discard rules with swapped antecedent and consequent, since their lift is the same than the kept rules. Table 3.4 summarises the support, confidence and lift of the top- and bottom-10 binary association rules sorted by lift. The interpretation of confidence for the found rules is the extent that the consequent is found in the antecedent. For instance, the confidence value of 0.74 of the second association rule means that 74% of respondents that choose to re-open the hospitality sector (RH) also choose to re-open contact professions (RC). In contrast,

[3] We used a value of $minsupport = 1 * 10^{-16}$ since the Apriori algorithm only accepts support thresholds above zero.

only 3% of respondents that choose re-opening the hospitality sector (RH) also choose to lift restrictions in Northern provinces (LN). The interpretation of lift is in terms of the extent that two alternatives are more (less) prone to be chosen together than each alternative separately, compared with the other rules after applying filter criteria. For instance, we find that choosing together to lift restrictions in the Northern provinces (LN) and for immune people (LI) is more prone to be chosen than independently, compared with the rest of binary association rules. Conversely, choosing to re-open the hospitality sector (RH) and lift restrictions in the Northern provinces (LN) together is found to be lower than choosing them independently, compared with the other binary rules.

*Table 3.4: Top- and bottom-10 binary association rules ordered by lift.*

| Antecedents | Consequents | Support | Confidence | Lift |
|---|---|---|---|---|
| LN | LI | 0.0067 | 0.1342 | 1.3192 |
| RH | RC | 0.2343 | 0.7423 | 1.1906 |
| YP | DF | 0.2082 | 0.5171 | 1.1792 |
| RB | YP | 0.2184 | 0.4354 | 1.0814 |
| LI | YP | 0.0419 | 0.4113 | 1.0216 |
| YP | RC | 0.2548 | 0.6328 | 1.0149 |
| RC | RB | 0.3130 | 0.5021 | 1.0011 |
| DF | RC | 0.2720 | 0.6202 | 0.9947 |
| LI | DF | 0.0443 | 0.4352 | 0.9924 |
| RB | DF | 0.2121 | 0.4230 | 0.9645 |
| RH | YP | 0.1110 | 0.3518 | 0.8738 |
| LN | RB | 0.0206 | 0.4107 | 0.8190 |
| DF | LN | 0.0170 | 0.0388 | 0.7729 |
| RH | DF | 0.1058 | 0.3353 | 0.7645 |
| RC | LN | 0.0226 | 0.0362 | 0.7201 |
| LI | RH | 0.0228 | 0.2237 | 0.7090 |
| LN | NH | 0.0097 | 0.1933 | 0.5998 |
| RH | NH | 0.0602 | 0.1907 | 0.5917 |
| LI | NH | 0.0190 | 0.1863 | 0.5781 |
| RH | LN | 0.0064 | 0.0202 | 0.4020 |

### Using association rules to assist the specification of the portfolio choice model

We use three different model specifications of the portfolio choice model. The first model is a baseline specification with alternative-specific constants, attribute parameters that do not vary across alternatives, and no alternative interaction parameters. The second model considers all interactions described in table 3.4. The third model is specified with the MI/AR approach to discard non-significant alternative interaction parameters.

Table 3.5 details the estimation results of the portfolio choice model under the three different specifications. We find that the models that include interaction terms (last two columns) outperform the baseline specification in terms of log-likelihood and Akaike/Bayesian information criteria (AIC and BIC). Furthermore, the model specified using the MI/AR approach outperforms the other two specifications in terms of information criteria, which means that this model is more parsimonious. All interactions associated with high lift values are statistically significant and have a positive sign, in line with out expectations. On the other hand, we find that two of the interactions associated with lower lift (Interaction: ['YP', 'RH'] and Interaction: ['RH', 'LI']) have a positive sign and are statistically significant, against our expectations. However, we also observe that the inclusion of alternative interaction parameters induce a change of sign of some of the alternative-specific constants of the models. Furthermore, it is easy to see that the utility of choosing 'LI' and 'RH' together is lower in the model specified with the MI/AR approach than in the baseline model (-1.87 against -1.16, respectively). Thus, the interpretation of positive or negative interactions does not rely solely on the alternative interaction parameters, but also on the combination of such parameters with the alternative-specific constants.

## 3.4.2. Random forests

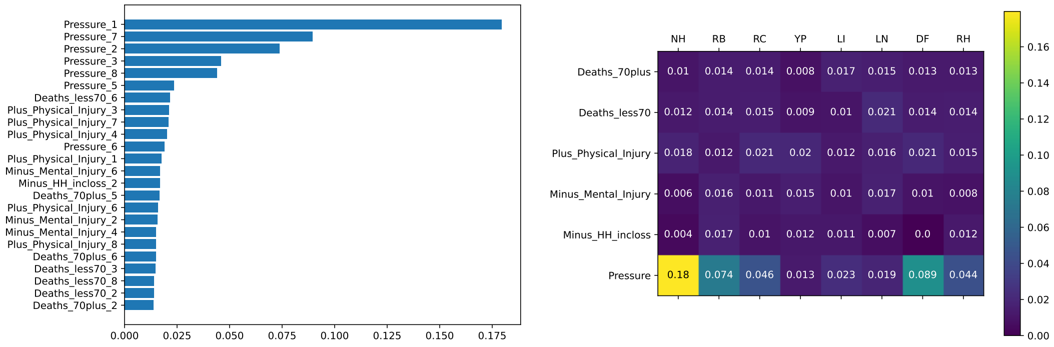
### Obtaining and interpreting variable importances

Figure 3.7 presents the top-half of variables ordered by their importance (a), as well as a heatmap view of the importance of all variables (b). We observe that the constrained attributes of the PVE choice experiment (i.e., pressure to the healthcare system) are among the most important variables. Visual inspection of the heatmap view allows to confirm that attributes other than pressure to the healthcare system play a rather minor role in terms of importance. In light of these results, we may expect that parameters associated with pressure to the healthcare system will predominate in a portfolio choice model specified with the MI/RF approach, and that most of the discards are focused on the remaining attribute-specific parameters.

| | Baseline model | All interactions | MI/AR |
|---|---|---|---|
| Remaining pressure | 0.0442 (0.0014)$^{***}$ | 0.0422 (0.0014)$^{***}$ | 0.0421 (0.0013)$^{***}$ |
| Constant of NH | 0.5185 (0.0415)$^{***}$ | 0.5959 (0.0477)$^{***}$ | 0.5932 (0.0471)$^{***}$ |
| Constant of RB | 0.6956 (0.0311)$^{***}$ | 0.3510 (0.0382)$^{***}$ | 0.3604 (0.0340)$^{***}$ |
| Constant of RC | 1.2643 (0.0351)$^{***}$ | 0.5193 (0.0431)$^{***}$ | 0.5197 (0.0397)$^{***}$ |
| Constant of YP | 0.1102 (0.0172)$^{***}$ | −0.6477 (0.0339)$^{***}$ | −0.6497 (0.0319)$^{***}$ |
| Constant of LI | −0.9674 (0.0354)$^{***}$ | −1.2422 (0.0505)$^{***}$ | −1.2427 (0.0479)$^{***}$ |
| Constant of LN | −1.1536 (0.0547)$^{***}$ | −1.2205 (0.0807)$^{***}$ | −1.2166 (0.0563)$^{***}$ |
| Constant of DF | 0.5025 (0.0389)$^{***}$ | 0.0779 (0.0472) | 0.0903 (0.0433)$^{*}$ |
| Constant of RH | 0.4653 (0.0528)$^{***}$ | −0.2946 (0.0616)$^{***}$ | −0.2998 (0.0577)$^{***}$ |
| Additional deaths 70 y.o. or more | −0.0707 (0.0965) | −0.0890 (0.1245) | −0.0873 (0.1204) |
| Additional deaths less than 70 y.o. | −0.7498 (0.2242)$^{***}$ | −0.7653 (0.1884)$^{***}$ | −0.7686 (0.2174)$^{***}$ |
| Additional physical injury | −0.1109 (0.0220)$^{***}$ | −0.1049 (0.0213)$^{***}$ | −0.1049 (0.0208)$^{***}$ |
| Reduction of psychological injury | 0.0204 (0.0046)$^{***}$ | 0.0179 (0.0046)$^{***}$ | 0.0180 (0.0045)$^{***}$ |
| Reduction of income losses | 0.0211 (0.0032)$^{***}$ | 0.0229 (0.0030)$^{***}$ | 0.0229 (0.0029)$^{***}$ |
| Interaction: ['LI', 'LN'] | | 0.8907 (0.0853)$^{***}$ | 0.8942 (0.0852)$^{***}$ |
| Interaction: ['RH', 'RC'] | | 1.2155 (0.0280)$^{***}$ | 1.2171 (0.0288)$^{***}$ |
| Interaction: ['YP', 'DF'] | | 0.6693 (0.0253)$^{***}$ | 0.6701 (0.0248)$^{***}$ |
| Interaction: ['YP', 'RB'] | | 0.4431 (0.0251)$^{***}$ | 0.4454 (0.0246)$^{***}$ |
| Interaction: ['LI', 'YP'] | | 0.3420 (0.0401)$^{***}$ | 0.3419 (0.0390)$^{***}$ |
| Interaction: ['YP', 'RC'] | | 0.2710 (0.0259)$^{***}$ | 0.2708 (0.0258)$^{***}$ |
| Interaction: ['RC', 'RB'] | | 0.2515 (0.0248)$^{***}$ | 0.2508 (0.0244)$^{***}$ |
| Interaction: ['RC', 'DF'] | | 0.2836 (0.0252)$^{***}$ | 0.2824 (0.0250)$^{***}$ |
| Interaction: ['LI', 'DF'] | | 0.2841 (0.0404)$^{***}$ | 0.2822 (0.0412)$^{***}$ |
| Interaction: ['DF', 'RB'] | | 0.0221 (0.0243) | |
| Interaction: ['YP', 'RH'] | | 0.1255 (0.0270)$^{***}$ | 0.1269 (0.0272)$^{***}$ |
| Interaction: ['LN', 'RB'] | | 0.0089 (0.0582) | |
| Interaction: ['LN', 'DF'] | | 0.0230 (0.0569) | |
| Interaction: ['RH', 'DF'] | | −0.1504 (0.0268)$^{***}$ | −0.1523 (0.0273)$^{***}$ |
| Interaction: ['LN', 'RC'] | | 0.0066 (0.0540) | |
| Interaction: ['LI', 'RH'] | | 0.2838 (0.0493)$^{***}$ | 0.2861 (0.0494)$^{***}$ |
| Interaction: ['LN', 'NH'] | | −0.2198 (0.0672)$^{**}$ | −0.2147 (0.0633)$^{***}$ |
| Interaction: ['RH', 'NH'] | | −0.1082 (0.0317)$^{***}$ | −0.1043 (0.0316)$^{***}$ |
| Interaction: ['LI', 'NH'] | | −0.3192 (0.0550)$^{***}$ | −0.3186 (0.0495)$^{***}$ |
| Interaction: ['LN', 'RH'] | | −0.0611 (0.0854) | |
| Log-likelihood | -124,119.02 | -122,336.75 | -122,337.59 |
| AIC | 248,266.03 | 244,741.50 | 244,733.18 |
| BIC | 248,382.19 | 245,023.61 | 244,973.80 |
| Rho-squared | 0.0981 | 0.1110 | 0.1110 |

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

*Table 3.5: Estimation results of portfolio choice models.*



*(a) Top-ranked variable importances* *(b) Heatmap view*

*Figure 3.7: RF variable importances, empirical data*

### Using variable importances to assist the specification of the portfolio choice model

Table 3.6 summarises the log-likelihood, rho-squared and information criteria of the baseline portfolio choice model detailed in table 3.5 (first column), a model with all attribute-specific parameters separated per alternative (second column), and a model with attribute-specific parameters specified with the MI/RF approach (third column).We observe that the model fit measures of the three models are similar, with slightly better performance in the MI/RF specification. This result can be explained by the low contribution of attributes to explain the portfolio choice model, other than pressure to the healthcare system.

| | Baseline model | Separated parameters | MI/RF |
|---|---:|---:|---:|
| Log-likelihood | -124,119.02 | -124,003.24 | -124,014.63 |
| AIC | 248,266.03 | 248,116.48 | 248,093.27 |
| BIC | 248,382.19 | 248,572.83 | 248,358.78 |
| Rho-squared | 0.0981 | 0.0989 | 0.0988 |

*Table 3.6: Model fit metrics of portfolio choice models.*

Table 3.7 summarises the estimation results of the portfolio choice model specified with the MI/RF approach. We observe that all alternative-specific constants are positive and statistically significant. The parameters associated with pressure to the healthcare system suggest mixed effects depending of each individual alternative. Notice that for this model specification (and the model with all attribute-specific parameters separated per alternative), pressure to the healthcare system is included as an additional attribute, hence it should be interpreted in terms of pressure increases, instead of remaining pressure such as in the results of table 3.5. With respect to the remaining attributes, all but one of the attribute-specific parameters have the expected sign. Furthermore, the estimated parameters of additional deaths of people of 70 years old or older become statistically significant and have a negative sign. Finally, we find that the estimate of additional physical injury by choosing to allow visitors in nursing homes has a positive sign, against our expectations.

| | NH | RB | RC | YP | LI | LN | DF | RH |
|---|---|---|---|---|---|---|---|---|
| Constant | 0.0383$^{***}$ | 0.0476$^{***}$ | 0.0418$^{***}$ | 0.0850$^{***}$ | 0.0511$^{***}$ | 0.0427$^{***}$ | 0.0677$^{***}$ | 0.0190$^{***}$ |
| | (0.0023) | (0.0036) | (0.0043) | (0.0078) | (0.0054) | (0.0052) | (0.0038) | (0.0033) |
| Additional pressure | 0.3043$^{***}$ | 0.6364$^{***}$ | 1.2694$^{***}$ | 0.4563$^{***}$ | $-0.9228^{***}$ | $-0.9068^{***}$ | 0.7593$^{***}$ | 0.1239 |
| | (0.0666) | (0.0427) | (0.0885) | (0.0542) | (0.0774) | (0.1541) | (0.0987) | (0.1061) |
| Additional deaths 70 y.o. or more | | | | $-2.1694^{***}$ | | | $-0.6655^{*}$ | 1.0220$^{*}$ |
| | | | | (0.6454) | | | (0.2797) | (0.4918) |
| Additional deaths less than 70 y.o. | | | $-1.1370^{**}$ | $-3.1763^{***}$ | | | | $-1.7083^{**}$ |
| | | | (0.4034) | (0.8106) | | | | (0.5999) |
| Additional physical injury | 0.8199$^{*}$ | | $-0.1210^{*}$ | | | $-0.3540^{**}$ | $-0.1324^{**}$ | $-0.1014^{*}$ |
| | (0.3810) | | (0.0524) | | | (0.1363) | (0.0463) | (0.0437) |
| Reduction of psychological injury | 0.0284$^{**}$ | | | | | | 0.0240$^{*}$ | |
| | (0.0091) | | | | | | (0.0096) | |
| Reduction of income losses | | 0.0261$^{***}$ | 0.0235$^{***}$ | | | | | 0.0143$^{*}$ |
| | | (0.0049) | (0.0063) | | | | | (0.0064) |
| Log-likelihood | -124,014.63 | | | | | | | |
| AIC | 248,093.27 | | | | | | | |
| BIC | 248,358.78 | | | | | | | |
| Rho-squared | 0.0988 | | | | | | | |

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

*Table 3.7: Estimation results, portfolio choice model specified with MI/RF approach.*

#### Testing the behavioural assumptions of the portfolio choice model

Table 3.8 summarises the model fit measures of the trained RF model compared with three specifications of the portfolio choice model: the baseline model, the specification based on the MI/AR approach used in table 3.5 and the specification based on the MI/RF approach detailed in table 3.7. We find that the RF model outperforms all the other choice modelling approaches, at least in terms of log-likelihood and rho-squared. We previously verified that the RF model is able to recover the true DGP of different specifications of the portfolio choice model, and that the RF model can approximate the ranking of combinations of chosen alternatives with highest choice probability. A more detailed descriptions of such tests can be found in appendix 3.D.

| | Baseline | MI/AR | RF/AR | RF |
|---|---|---|---|---|
| Log-likelihood | -124,119.02 | -122,337.59 | -124,014.63 | -120,197.17 |
| Rho-squared | 0.0981 | 0.1110 | 0.0988 | 0.1266 |

*Table 3.8: Model fit measures, RF model compared with portfolio choice models*

Table 3.9 details the Kendall's Tau value obtained from contrasting the top-5 and top-10 rankings of combinations of alternatives with highest choice probability of the RF model with their respective top-5 and top-10 rankings obtained from the baseline portfolio choice model and the specifications specified with the MI/AR and MI/RF approaches. Among all the contrasts, the only case in which the hypothesis of no-correlation is rejected (at 90%) of confidence is between the top-10 rankings of the RF model and the portfolio choice model specified with the MI/AR approach.

| | RF vs. Baseline | RF vs. MI/AR | RF vs. MI/RF |
|---|---|---|---|
| Top-5 | -0.4000 | -0.2000 | -0.4000 |
| Top-10 | 0.3778 | $0.4667^{+}$ | 0.3778 |

P-values of Kendalls Tau: $^{+}$ : $p < 0.1$

*Table 3.9: Comparison of Kendall's Tau of rankings of combinations of alternatives with highest choice probability.*

Finally, table 3.10 details the top-10 ranking of combinations of alternatives with highest choice probability of the RF model, the baseline portfolio choice model and the portfolio choice model specified with the MI/AR approach. We observe that not choosing any alternative is the combination with the highest probability from the predictions of a RF model. For the baseline portfolio choice model, the combination with highest choice probability is to re-open businesses, re-open contact professions and allow contact between direct family members of different households, which aligns with the results of Mouter et al. (2021) despite a different choice model was used i.e., a Multiple Discrete-Continuous Extreme Value (MDCEV) model. The combination of alternatives with highest choice probability in the MI/AR portfolio choice model is to re-open businesses, re-open contact professions, allow young people to come together in groups and allow contact between direct family members of different households. None of the portfolio choice models include not choosing any alternative among the top-10 probability ranking, whereas the ranking of the RF model ranks as third-best the combination that ranks the first in the model specified with the MI/AR approach.

## 3.5. Conclusion and discussion

In this paper, we propose procedures based on AR learning and RF models to support the specification of a portfolio choice model applied in data from a PVE choice experiment, and we provide insights on the interpretation of the outcomes of the proposed models from a choice modelling perspective. We use data from a PVE choice experiment conducted to elicit the preferences of Dutch citizens to lift COVID-19 restrictions during the first wave of the Coronavirus pandemic in 2020. On the one hand, AR learning is used to identify relevant interactions between different combinations of alternatives chosen by respondents of the PVE choice experiment and support the specification of alternative interaction parameters in a portfolio choice model. On the other hand, RF models are used to identify the most (least) relevant attributes of the PVE choice experiment, and with this information assist the inclusion/exclusion of attribute-specific estimates of the portfolio choice model. Finally, RF models are used to predict the combinations of alternatives with the highest choice probability, and use that information to test the validity of the behavioural assumptions of several specifications of the PVE choice model.

### 3.5.1. Main findings

Firstly, we show that AR learning successfully identifies relevant interactions between chosen alternatives of a PVE choice experiment. For instance, we find that choosing to lift all restrictions for the immune people (LI) and in the Northern provinces (LN) together have the highest lift among binary association rules, despite both al-

Random Forest

| | Rk.1 | Rk.2 | Rk.3 | Rk.4 | Rk.5 | Rk.6 | Rk.7 | Rk.8 | Rk.9 | Rk.10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Comb. ID | 0 | 134 | 78 | 6 | 70 | 5 | 69 | 7 | 76 | 196 |
| NH | | | | | | X | X | X | | |
| RB | | X | X | X | X | | | X | | |
| RC | | X | X | X | X | X | X | X | X | X |
| YP | | | X | | | | | | X | |
| LI | | | | | | | | | | |
| LN | | | | | | | | | | |
| DF | | | X | | X | | X | | X | X |
| RH | | X | | | | | | | | X |
| Choice probability | 5.34% | 4.94% | 3.71% | 2.7% | 2.6% | 2.6% | 2.53% | 2.53% | 2.46% | 2.44% |
| Pressure | 0.0% | 40.41% | 38.22% | 21.5% | 31.71% | 29.54% | 39.75% | 39.49% | 28.26% | 40.66% |

Baseline portfolio choice model

| | Rk.1 | Rk.2 | Rk.3 | Rk.4 | Rk.5 | Rk.6 | Rk.7 | Rk.8 | Rk.9 | Rk.10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Comb. ID | 70 | 6 | 78 | 14 | 68 | 134 | 4 | 7 | 76 | 12 |
| NH | | | | | | | | X | | |
| RB | X | X | X | X | | X | | X | | |
| RC | X | X | X | X | X | X | X | X | X | X |
| YP | | | X | X | | | | | X | X |
| LI | | | | | | | | | | |
| LN | | | | | | | | | | |
| DF | X | | X | | X | | | | X | |
| RH | | | | | | X | | | | |
| Choice probability | 3.68% | 3.56% | 2.93% | 2.9% | 2.87% | 2.82% | 2.78% | 2.67% | 2.34% | 2.27% |
| Pressure | 31.71% | 21.5% | 38.22% | 28.01% | 21.75% | 40.41% | 11.54% | 39.49% | 28.26% | 18.05% |

MI/AR portfolio choice model

| | Rk.1 | Rk.2 | Rk.3 | Rk.4 | Rk.5 | Rk.6 | Rk.7 | Rk.8 | Rk.9 | Rk.10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Comb. ID | 78 | 134 | 132 | 196 | 6 | 70 | 4 | 7 | 204 | 14 |
| NH | | | | | | | | X | | |
| RB | X | X | | | X | X | | X | | X |
| RC | X | X | X | X | X | X | X | X | X | X |
| YP | X | | | | | | | | X | X |
| LI | | | | | | | | | | |
| LN | | | | | | | | | | |
| DF | X | | | X | | X | | | X | |
| RH | | X | X | X | | | | | X | |
| Choice probability | 3.98% | 3.79% | 3.2% | 3.18% | 2.92% | 2.64% | 2.44% | 2.41% | 2.37% | 2.31% |
| Pressure | 38.22% | 40.41% | 30.45% | 40.66% | 21.5% | 31.71% | 11.54% | 39.49% | 47.17% | 28.01% |

*Table 3.10: Top-10 ranking of combinations of alternatives with highest choice probability, RF and portfolio choice models*

ternatives are the least chosen independently. Furthermore, we find that the MI/AR approach to specify interactions in the portfolio choice model leads to model fit improvements in terms of log-likelihood and information criteria, compared with a baseline portfolio choice model. Additionally, we find that directly interpreting the sign of the interaction parameters does not indicate whether an interaction is positive or negative. Instead, a comparison of the utilities of the baseline model and the model with specified interactions can shed light on the positive or negative effect of interactions.

Secondly, our data analyses with RF models show that respondents of the PVE choice experiment mostly care about the constrained attribute (additional pressure to the healthcare system) across almost all alternatives, whereas they put considerable lower relevance to the other attributes of the PVE choice experiment. We find that the MI/RF approach leads to modest improvements of model fit compared with estimating the baseline portfolio choice model. This can be a consequence of the small relevance of the attributes, other than pressure to the healthcare system, found from the variable importances of the RF model. Despite the latter, we find additional insights from the MI/RF approach in terms of interpretation of parameters, such as preference differences for the same attribute across individual alternatives. For instance, additional deaths of people of less than 70 years old due to COVID-19 has significantly different effects (in terms of magnitude) between re-opening contact professions (RC) and allowing young people to come together in groups (YP), despite both estimates having a negative sign.

Finally, we find that RF models are able to recover the true DGP from PVE choice experiment data under different specifications of pseudo-synthetic data, and they outperform portfolio choice models in terms of model fit using empirical data, under different model specifications. We find that portfolio choice model specified with the support of AR learning leads to a predicted ranking that tends to get closer to the ranking obtained from a RF model (in terms of the Kendall's Tau), compared with the other model specifications (i.e., baseline model and the model assisted with the variable importances of a RF model). Nevertheless, all portfolio choice models underestimate the choice probability of not choosing any alternative, which ranks as the highest in the ranking obtained with a RF model. Our findings evidence that the portfolio choice models still have misalignments between their behavioural assumptions and the actual DGP embedded in the data, but the procedures we propose in this paper can help to mitigate such misalignments.

### 3.5.2. Additional uses of the outcomes of AR learning and RF models

Besides assisting the specification of portfolio choice models, the outcomes of AR learning and RF models applied in PVE choice experiments data can be directly used. With respect to AR learning, this method provides beforehand information about frequent interactions between chosen (combinations of) alternatives, without the need of specifying and estimating a choice model or compute welfare measures. Such interactions can be used in policy making to, for example, recommend in favour of conducting combinations of policies that rank high in terms of lift. In that regard, when the aim is merely identifying frequent combinations of chosen alternatives, using AR learning is advantageous because it does not rely on strict behavioural assumptions that can restrict (or privilege) certain interactions over others. Additionally, the computation runtime of AR learning is shorter than the regular estimation time of a portfolio choice model.

With respect to the variable importances of RF models, this approach can be used as an alternative to estimating the attribute-specific parameters or marginal utilities of a portfolio choice model, without the need of explicitly specifying the form of the utility function. This information can be used to prioritise (or avoid) policy options that perform high on desirable (undesirable) attributes such as, for instance, not lifting restrictions to visits in nursing homes since the relevance of the pressure to the healthcare system is high for this option. As an additional advantage, the time dedicated to train a RF model and obtain the variable importances is generally shorter than the estimation time of a portfolio choice model, and such differences are bigger as the number of individual alternatives of the PVE choice experiment increase. However, variable importances only provide information about the relevance of an alternative, and they do not inform whether the effect of an attribute is positive (negative) for choosing an alternative, unlike the attribute-specific estimates of a portfolio choice model.

Finally, when the aim is solely prediction (i.e., no focus on behavioural interpretation), the RF model is advantageous to determine the ranking of combinations of alternatives with the highest choice probability without relying on *a priori* behavioural assumptions. We find that RF models outperform several specifications of a portfolio choice model, in terms of predictive performance. We argue in favour that the probability ranking of an RF model trained with PVE choice experiment data should be a closer reflection of the true ranking that is embedded in the data.

### 3.5.3. Considerations and further research

As a consideration of this work, we advice that, while we provide potential uses and interpretations of the outcomes of AR learning and RF models, we recall that such outcomes should not be treated as equivalent to the outcomes of a choice model. For instance, finding that an association rule has a high (low) lift value does not necessarily mean that the corresponding interaction specified in a choice model will be statistically significant. As shown in this paper, we recall that interactions with the highest (lowest) lift does not necessarily lead to interaction parameters with positive (negative) sign in the choice model. We emphasise that AR learning and RF models are used as supportive tools in this paper, whereas choice models are used as confirmatory tools.

In addition, we provide suggested interpretations of the outcomes that can be obtained with currently developed data-driven methods, but a potential step beyond is to develop outcomes that are particularly tailored to the particularities of the data that is analysed, such as in the case of a PVE choice experiment. For instance, the formulas of support, confidence and lift used in AR learning are built for analysing transaction datasets, but they were not thought for the case of a PVE choice experiment, in which choices have a resource constraint, and hence the interpretation of measures can be affected. In this regard, developing expressions that consider constrained choices, such as in a PVE choice experiment, can provide more strength to the use of AR learning in these contexts. Finally, this paper opens the door to new research directions in the field of bringing data-driven methods to the choice modelling field. For instance, based in recent research (Alwosheel et al., 2021) and our experience with obtaining variable importances from RF models, we see opportunities to integrate explainable AI techniques to analyse data from PVE choice experiments.

## Acknowledgements

The authors acknowledge the Netherlands Organisation for Scientific Research (NWO) for funding part of this research (NWO Responsible Innovation grant – 313-99-333), the support of TU Delft AI Labs programme and from the ERC-BEHAVE group (ERC Consolidator grant BEHAVE—724431). We also acknowledge the valuable comments done by the two anonymous reviewers of this paper. Jose Ignacio Hernandez would like to also acknowledge the valuable feedback provided by his fellow PhD researchers to shape this work.

# 3.A. Pseudo-synthetic data generation and parametrisation

## 3.A.1. Data-generating processes

We generate four pseudo-synthetic datasets using the experimental design of the COVID-19 PVE choice experiment. The first two datasets are generated using the behavioural assumptions of the portfolio choice model proposed by Bahamonde-Birke & Mouter (2019), whereas the last two datasets are based on the MDCEV-based model proposed by Dekker et al. (2019). Table 3.11 summarises the utility and stochastic specification of each of the pseudo-synthetic datasets. For datasets 1 and 2, we use the utility specification of Bahamonde-Birke & Mouter (2019) that relies in a linear-in-parameters utiliy of each possible combination of alternatives, plus the addition of combination-specific stochastic errors (notice that errors are specified as $\varepsilon_{np}$) with a Gumbel (Extreme-Value type 1) distribution. Dataset 1 and 2 differs in the specification of explicit interactions: in dataset 1, we assume that no interactions between chosen alternatives are present (i.e., $\theta_{ij} = 0, \forall i, j$), whereas in dataset 2 we let these parameters free to be estimated. Datasets 3 and 4 are generated using the utility specification of Dekker et al. (2019), hence relying in the assumptions of the MDCEV-type choice model. Apart from differences in the specification of the utility function, the MDCEV-type datasets differ from the former approach in the specification of stochastic terms, which in this case correspond to alternative-specific terms (notice that $\varepsilon_{nj}$ are at alternative-level). For dataset 3, we assume i.i.d. Gumbel-distributed terms, whereas for dataset 4 we incorporate unobserved correlation between alternatives by using a Generalised Extreme Value (GEV) distribution.

| Dataset | Utility specification |
|---|---|
| Dataset 1 and Dataset 2 | $U_{np} = \sum_{j=1}^{J} y_{nj} \cdot (\delta_j + \beta' X_{njk}) + \delta_0 \cdot \left(B - \sum_{j=1}^{J} y_{nj} \cdot c_{nj}\right) + \sum_i \sum_j \theta_{ij} y_i y_j + \varepsilon_{np}$<br>$\varepsilon_{np} \sim \mathit{Gumbel}(0,1)$<br>$\theta_{ij} = 0, \forall i, j$ (dataset 1) |
| Dataset 3 and Dataset 4 | $U_n = \left(B - \sum_{j=1}^{J} y_{nj} \cdot c_{nj}\right) \cdot \exp(\delta_0 + \varepsilon_{n0}) + \sum_{j=1}^{J} y_{nj} \cdot \exp(\delta_j + \beta' X_{nj} + \varepsilon_{nj})$<br>$\varepsilon_{nj} \sim \mathit{Gumbel}$ (for dataset 3)<br>$\varepsilon_{nj} \sim \mathit{GEV}$ (for dataset 4) |

*Table 3.11: DGP specification of pseudo-synthetic datasets*

Table 3.12 details the values used to parametrise each of the pseudo-synthetic datasets. We define eight alternative-specific constants ranging from -0.9 to 0.5. The attribute-specific parameters are assumed equal across different alternatives and range from -0.8 to 0.03. The parameters associated to the marginal utility of non-spent resources as 0.01 for datasets 1 and 2, and -3 for datasets 3 and 4. In addition, we define positive and negative interactions between chosen alternatives for dataset 2. Specifically, we define a positive interaction when lifting all restrictions to immune people and from Northern provinces (LI and LN) are chosen together, a negative interaction when re-opening all types of businesses (RB, RC and RH) are chosen together, and a negative interaction when allowing visits in nursing homes and allowing contact between direct family members (NH and DF) are chosen together. In the same way, we explicitly define unobserved correlation between alternatives through different so-called dissimilarity parameters of the GEV distribution on dataset 4, varying across consecutive pairs of alternatives.

| Type of parameter | Description | Parameter | Value |
|---|---|---|---|
| Marginal utility of non-spent resources | Datasets 1 and 2 | $\delta_0$ | 0.01 |
| | Datasets 3 and 4 | | -3 |
| Alternative-specific constants | ASC for NH | $\delta_1$ | 0 |
| | RB | $\delta_2$ | 0.2 |
| | RC | $\delta_3$ | -0.3 |
| | YP | $\delta_4$ | 0.4 |
| | LI | $\delta_5$ | 0.5 |
| | LN | $\delta_6$ | 0.4 |
| | DF | $\delta_7$ | 0.3 |
| | RH | $\delta_8$ | -0.9 |
| Attribute-specific parameters | Additional 70+ deaths | $\beta_1$ | -0.6 |
| | Additional $<$ 70 deaths | $\beta_2$ | -0.8 |
| | Additional people w. physical injuries | $\beta_3$ | -0.1 |
| | Reduction people w. psychological injuries | $\beta_4$ | 0.03 |
| | Reduction households w. income losses | $\beta_5$ | 0.03 |
| Interaction parameters (only dataset 2) | Interaction LI & LN | $\theta_{5,6}$ | 2.5 |
| | Interaction RB, RC & RH | $\theta_{2,3,8}$ | -4.8 |
| | Interaction NH & DF | $\theta_{1,7}$ | -0.3 |
| Dissimilarity parameters (only dataset 4) | NH & RB | $\lambda_{1,2}$ | 0.03 |
| | RC & YP | $\lambda_{3,4}$ | 0.05 |
| | LI & LN | $\lambda_{5,6}$ | 0.1 |
| | DF & RH | $\lambda_{7,8}$ | 0.2 |

*Table 3.12: Parametrization of synthetic datasets*

## 3.B. Hyperparameter tuning of the RF model

We conduct a grid search process to find the best combination of hyperparameters, and we keep the combination that reports the highest test (out-of-sample) log-likelihood. Table 3.13 presents the values considered for the tuning process. We tested trees ranging from 10 to 1,000 individual decision trees, increasing this number in multiples of ten. In terms of maximum depth, we used three, five, ten and the default setting (max) of the RF model optimisation algorithm. Finally, we fixed the maximum number of variables per split in power values of four, from four to 16, plus the default setting of $\sqrt{J*(K+1)}$, named as "auto". We constructed RF models using all possible combinations of parameters of table 3.13.

| Parameter | Values |
|---|---|
| Number of trees | From 10 to 1,000, in multiples of 10. |
| Depth | 3, 5, 10, max (default). |
| Maximum variables per split | 4, 8, 16, $\sqrt{J*(K+1)}$ (auto) |

*Table 3.13: Hyperparameter values for RF model specification*

To identify the best combination of hyperparameters, we proceed in two stages. First, we train each possible RF model under differnt combinations of hyperparameters. Second, we fix either the tree depth or the maximum number of variables per split, and we plot the (out-of-sample) log-likelihood for different specifications of the other parameter as a function of the maximum number of trees. Finally, we choose the combination of hyperparameters that reports the maximum log-likelihood.

Figure 3.8 details the log-likelihood values for different number of variables per split and different number of trees, for a tree depth fixed in five layers, trained with the empirical data. We conclude that a RF model with maximum depth of five layers, 16 variables per split and 200 decision trees lead to the best log-likelihood. We conducted the same process in the pseudo-synthetic datasets, leading to the same combination of hyperparameters.

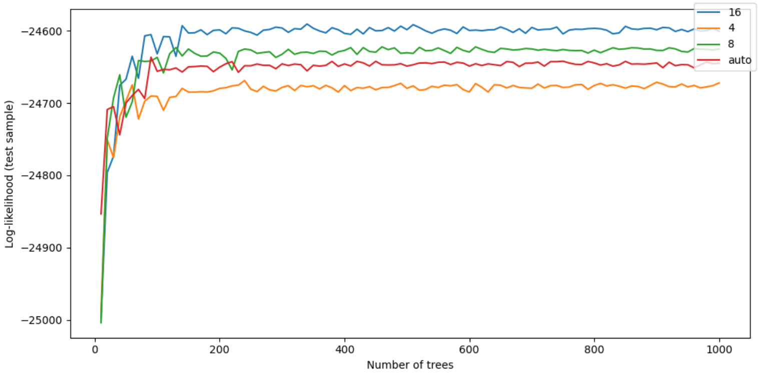


*Figure 3.8: Log-likelihood of RF models at different parameter specifications, empirical data*

## 3.C. AR learning outcomes of pseudo-synthetic data

We show that the confidence and lift values of AR learning align with the specification of interactions and unobserved heterogeneity in pseudo-synthetic data. Specifically, we gather association rules from the pseudo-synthetic datasets, and we compare the confidence and lift between datasets without and with such interactions.

Table 3.14 summarises the support, confidence and lift of a selection of rules in which we explicitly defined interactions or correlations between errors. As expected, incorporating interactions between alternatives in the portfolio choice dataset induce a change of magnitud and direction of the confidence and lift values for a same association rule. For instance, the association rule of lifting restrictions for immune people (LI) and from Northern provinces (LN) has a confidence of 24% and a lift equal to 0.74 in dataset 1 (without interactions), whereas in dataset 2 (with interactions) these values increase to 74% and 1.18, respectively, in line with the positive interaction defined for these two alternatives in dataset 2. We observe the same behaviour for association rules in which we defined negative interactions. Similar patterns are observed in the case of correlated errors in MDCEV-type datasets, in which the incoration of these correlations induce an increase on confidence and lift values, contrasted with the same rule in the dataset with i.i.d. errors. Notice that incorporating interactions or correlated errors does not necessarily mean that the lift of a rule will be above (below) one. Instead, we observe changes with respect to the lift value computed in the datasets without explicit interactions.

| Portfolio choice model | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | Dataset 1 | | | Dataset 2 | | |
| Association rules | | Without interactions | | | With interactions | | |
| Antecedents | Consequents | Support | Confidence | Lift | Support | Confidence | Lift |
| LI | LN | 0.0962 | 0.2415 | 0.7147 | 0.5018 | 0.7435 | 1.1881 |
| RC | RB | 0.1562 | 0.4451 | 0.94 | 0.0916 | 0.3519 | 0.937 |
| RH | RB | 0.0815 | 0.4097 | 0.8651 | 0.0354 | 0.3053 | 0.8128 |
| RH | RC | 0.0555 | 0.2789 | 0.7948 | 0.0219 | 0.1888 | 0.7254 |
| RH, RC | RB | 0.0199 | 0.3588 | 0.7578 | - | - | - |
| DF | NH | 0.0884 | 0.2124 | 0.8638 | 0.0383 | 0.1196 | 0.7656 |
| NH | DF | 0.0884 | 0.3597 | 0.8638 | 0.0383 | 0.2449 | 0.7656 |
| MDCEV-type model | | | | | | | |
| | | Dataset 3 | | | Dataset 4 | | |
| Association rules | | i.i.d. errors | | | GEV (correlated errors) | | |
| Antecedents | Consequents | Support | Confidence | Lift | Support | Confidence | Lift |
| RB | NH | 0.1479 | 0.248 | 0.9942 | 0.1857 | 0.2912 | 1.216 |
| RC | YP | 0.2822 | 0.6726 | 0.9867 | 0.3242 | 0.768 | 1.0677 |
| LI | LN | 0.1341 | 0.2897 | 0.7443 | 0.1542 | 0.3374 | 0.8611 |
| DF | RH | 0.0725 | 0.135 | 0.8504 | 0.0751 | 0.4992 | 0.927 |

*Table 3.14: Effect of interactions in association rules*

## 3.D. Probability rankings with pseudo-synthetic data

Table 3.15 shows the predicted log-likelihood of the RF models trained with each pseudo-synthetic datasets (RF log-likelihood), compared with their respective log-likelihood values obtained from the true DGP (True log-likelihood). The RF model is able to get close to the true DGP with a considerable precision in datasets 1 and 2, and with more distance in datasets 3 and 4. This distance between the true and predicted log-likelihood in the latter datasets can be attributed by slight differences in the choice probabilities across different combinations of alternatives, as well as the different error structures imposed in these datasets, compared with datasets 1 and 2.

| | Portfolio choice model | | MDCEV-type model | |
|---|---|---|---|---|
| | Dataset 1 | Dataset 2 | Dataset 3 | Dataset 4 |
| DGP | Without interactions | With interactions | i.i.d. errors | GEV (correlated errors) |
| True log-likelihood | -134476.81 | -110298.15 | -113452.41 | -97375.57 |
| RF log-likelihood | -135041.43 | -111779.66 | -119345.83 | -113775.94 |

Note: True log-likelihood of datasets 3 and 4 are computed using 10000 simulations.

*Table 3.15: True and predicted log-likelihood values, pseudo-synthetic datasets*

Table 3.16 summarises the Kendall's Tau values resulting from comparing the top-5 and top-10 rankings of the trained RF models with their respective rankings obtained from the true DGP, for each pseudo-synthetic dataset. We observe that the Kendall's Tau of the top-5 portfolio is close to one in three out of four datasets (Datasets 2, 3 and 4), which means that the trained RF model is able to retrieve the true ranking in this case, whereas in Dataset 1 the Kendall's Tau is of 58,3%, but still statistically different from zero, which suggests a correlation between the prediction of the RF model and the true DGP. The Kendall's Tau values of the top-10 portfolios show a decrease of predictive power on this ranking, which can be explained due to slight changes of position of some combinations. Despite the latter, all correlation values are statistically different from zero, and thus still suggesting the existence of correlation between the predicted and true probability rankings. We provide a detail of the rankings for each dataset in tables 3.17 to 3.20.

| | Portfolio choice model | | MDCEV-type model | |
|---|---|---|---|---|
| | Dataset 1 | Dataset 2 | Dataset 3 | Dataset 4 |
| DGP | Without interactions | With interactions | i.i.d. errors | GEV (correlated errors) |
| Top-5 | $0.583^{**}$ | $\sim 1.000^{**}$ | $\sim 1.000^{**}$ | $\sim 1.000^{**}$ |
| Top-10 | $0.513^{**}$ | $\sim 1.000^{**}$ | $0.800^{**}$ | $0.500^{**}$ |

P-values of Kendalls Tau: $^{**}$ : $p < 0.001$, $^{*}$ : $p < 0.01$

*Table 3.16: Kendall's Tau correlation between most likely chosen portfolios and "true" rankings. Pseudo-synthetic datasets*

| | | Rk. 1 | Rk. 2 | Rk. 3 | Rk. 4 | Rk. 5 | Rk. 6 | Rk. 7 | Rk. 8 | Rk. 9 | Rk. 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| True | NH | | | | | | | | | | |
| | RB | X | X | X | | X | X | X | | | |
| | RC | | | | | | | | | | |
| | YP | X | X | X | X | | X | | X | | X |
| | LI | X | | X | | X | | | X | X | X |
| | LN | | X | | X | | | X | X | X | |
| | DF | X | | | X | X | X | | | | X |
| | RH | | | | | | | | | | |
| | Choice prob. | 0.018 | 0.017 | 0.017 | 0.016 | 0.015 | 0.014 | 0.014 | 0.014 | 0.014 | 0.014 |
| | Pressure | 42.0% | 38.49% | 31.65% | 38.99% | 35.46% | 26.74% | 31.96% | 43.91% | 37.37% | 32.15% |
| RF | NH | | | | | | | | | | |
| | RB | X | X | X | | X | | X | | | X |
| | RC | | | | | | | | | | |
| | YP | X | X | X | X | X | X | | X | | |
| | LI | X | X | | | | X | | X | X | X |
| | LN | | | X | X | | | X | X | X | |
| | DF | | X | | X | X | X | | | | X |
| | RH | | | | | | | | | | |
| | Choice prob. | 0.018 | 0.018 | 0.017 | 0.016 | 0.015 | 0.015 | 0.015 | 0.014 | 0.014 | 0.013 |
| | Pressure | 31.65% | 42.0% | 38.49% | 38.99% | 26.74% | 32.15% | 31.96% | 43.91% | 37.37% | 35.46% |

*Table 3.17: Probability ranking, dataset 1*

| | | Rk. 1 | Rk. 2 | Rk. 3 | Rk. 4 | Rk. 5 | Rk. 6 | Rk. 7 | Rk. 8 | Rk. 9 | Rk. 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| True | NH | | | | | | | | X | | |
| | RB | | | X | | | X | | | X | |
| | RC | | | | | X | | | | | |
| | YP | | X | | | | X | X | | | |
| | LI | X | X | X | X | X | X | X | X | X | X |
| | LN | X | X | X | X | X | X | X | X | X | X |
| | DF | | | | X | | | X | | X | |
| | RH | | | | | | | | | | X |
| | Choice prob. | 0.098 | 0.081 | 0.059 | 0.05 | 0.042 | 0.034 | 0.025 | 0.02 | 0.015 | 0.012 |
| | Pressure | 37.37% | 43.91% | 47.22% | 47.72% | 48.86% | 53.76% | 54.26% | 55.33% | 57.57% | 56.44% |
| RF | NH | | | | | | | | X | | |
| | RB | | | X | | | X | | | X | |
| | RC | | | | | X | | | | | |
| | YP | | X | | | | X | X | | | |
| | LI | X | X | X | X | X | X | X | X | X | X |
| | LN | X | X | X | X | X | X | X | X | X | X |
| | DF | | | | X | | | X | | X | |
| | RH | | | | | | | | | | X |
| | Choice prob. | 0.098 | 0.082 | 0.058 | 0.05 | 0.043 | 0.035 | 0.025 | 0.02 | 0.016 | 0.012 |
| | Pressure | 37.37% | 43.91% | 47.22% | 47.72% | 48.86% | 53.76% | 54.26% | 55.33% | 57.57% | 56.44% |

*Table 3.18: Probability ranking, dataset 2*

| | | Rk. 1 | Rk. 2 | Rk. 3 | Rk. 4 | Rk. 5 | Rk. 6 | Rk. 7 | Rk. 8 | Rk. 9 | Rk. 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| True | NH | | | | | | | | | | |
| | RB | X | X | X | X | X | | X | X | | X |
| | RC | | | X | X | X | X | | | | X |
| | YP | X | X | X | X | X | X | X | X | X | X |
| | LI | X | | | X | X | X | | X | | |
| | LN | | X | | | | | X | X | X | X |
| | DF | X | X | X | X | | X | | | X | |
| | RH | | | | | | | | | | |
| | Choice prob. | 0.041 | 0.032 | 0.028 | 0.025 | 0.024 | 0.021 | 0.019 | 0.018 | 0.018 | 0.017 |
| | Pressure | 42.0% | 48.84% | 38.23% | 53.5% | 43.15% | 43.65% | 38.49% | 53.76% | 38.99% | 49.99% |
| RF | NH | | | | | | | | | | |
| | RB | X | X | X | X | X | X | | X | X | |
| | RC | | | X | X | X | | X | | X | |
| | YP | X | X | X | X | X | X | X | X | X | X |
| | LI | X | | | X | X | | X | X | | X |
| | LN | | X | | | | X | | X | X | X |
| | DF | X | X | X | X | | | X | | | |
| | RH | | | | | | | | | | |
| | Choice prob. | 0.042 | 0.032 | 0.029 | 0.025 | 0.025 | 0.02 | 0.02 | 0.018 | 0.018 | 0.017 |
| | Pressure | 42.0% | 48.84% | 38.23% | 53.5% | 43.15% | 38.49% | 43.65% | 53.76% | 49.99% | 43.91% |

*Table 3.19: Probability ranking, dataset 3*

| | | Rk. 1 | Rk. 2 | Rk. 3 | Rk. 4 | Rk. 5 | Rk. 6 | Rk. 7 | Rk. 8 | Rk. 9 | Rk. 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| True | NH | | | | | | | | X | | |
| | RB | X | X | X | X | X | | X | X | X | |
| | RC | | X | | X | X | X | | X | X | |
| | YP | X | X | X | X | X | X | X | X | X | X |
| | LI | X | | | X | X | X | X | | | X |
| | LN | | | X | | | | X | | X | X |
| | DF | X | X | X | X | | X | | X | | |
| | RH | | | | | | | | | | |
| | Choice prob. | 0.04 | 0.035 | 0.031 | 0.029 | 0.028 | 0.024 | 0.022 | 0.021 | 0.02 | 0.019 |
| | Pressure | 42.0% | 38.23% | 48.84% | 53.5% | 43.15% | 43.65% | 53.76% | 56.19% | 49.99% | 43.91% |
| RF | NH | | | | | | | X | | | |
| | RB | X | X | X | X | X | | X | X | X | |
| | RC | | X | | X | X | X | X | | X | |
| | YP | X | X | X | X | X | X | X | X | X | X |
| | LI | X | | | X | X | X | | X | | X |
| | LN | | | X | | | | | X | X | X |
| | DF | X | X | X | X | | X | X | | | |
| | RH | | | | | | | | | | |
| | Choice prob. | 0.041 | 0.037 | 0.032 | 0.03 | 0.027 | 0.023 | 0.023 | 0.022 | 0.021 | 0.019 |
| | Pressure | 42.0% | 38.23% | 48.84% | 53.5% | 43.15% | 43.65% | 56.19% | 53.76% | 49.99% | 43.91% |

*Table 3.20: Probability ranking, dataset 4*

## 3.E. Explanatory notes (only applicable for this thesis)

- In the Random Forest model (section 3.3.2) the selected criterion for doing the splits of each decision tree was the Gini impurity:

  $$G(X_i) = \sum_{j=1}^{J} P(X_i = L_j)(1 - P(X_i = L_j)), \tag{3.8}$$

  where $X_i$ is the candidate variable for making a split in the RF model, with a possible number of categories $L_1, \ldots, L_J$, and $P(X_i = L_j)$ is the predicted probability of $X_i = L_j$

  The goal of the decision trees algorithm is to find the splits such that the impurity is minimized. Likewise, the Gini impurity is also used to calculate the mean decrease of impurity as a measure of variable importance.

- It is important to state that an explicit comparison of Random Forests and XGBoost (used for Chapter 4) was not done in this chapter. Said that, one of the main advantages of Random Forests over XGBoost is that the former is conceptually easier to understand for researchers who are not familiar with the machine learning field, such as most of the readers of the Journal of Choice Modelling. The idea behind Random Forests is conceptually intuitive: train (estimate) several models simultaneously and then aggregate the results to build a model with a higher predictive power. XGBoost, on the other hand, is conceptually more difficult since models are trained sequentially, each model aims to improve the results of the previous aggregation of models, and models are trained such that a differentiable loss function is minimised. Considering the goal of this paper is to introduce a method for assisting the specification of choice models for a novel type of choice experiment (i.e., PVE experiments), I deem a more reasonable strategy to use Random Forests as they are easier to understand, and by doing so, reducing the scepticism of the reader and the referees.

- Another key difference between Random Forests and alternative modelling approaches, such as XGBoost (used for Chapter 4) is the way that probabilities are computed. In the Random Forest model, probabilities are computed by averaging the predictions across all individual decision trees. In alternative methods, including XGBoost, predicted probabilities are computed through the logistic function. An implication of this is that a Random Forest model would require more individual decision trees in order to guarantee that predicted probabilities approximate to a continuous function.

# Bibliography

Agrawal, R., T. Imieliński, A. Swami (1993) Mining association rules between sets of items in large databases, in: *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pp. 207–216.

Alwosheel, A., S. van Cranenburgh, C. G. Chorus (2021) Why did you predict that? towards explainable artificial neural networks for travel demand analysis, *Transportation Research Part C: Emerging Technologies*, 128, p. 103143.

Bahamonde-Birke, F. J., N. Mouter (2019) About positive and negative synergies of social projects: treating correlation in participatory value evaluation., in: *hEART 2019: 8th Symposium of the European Association for Research in Transportation. Budapest, Hungary.*

Breiman, L. (2001) Random forests, *Machine learning*, 45(1), pp. 5–32.

Caputo, V., J. L. Lusk (2022) The basket-based choice experiment: A method for food demand policy analysis, *Food Policy*, 109, p. 102252.

Carson, R. T., T. C. Eagle, T. Islam, J. J. Louviere (2022) Volumetric choice experiments (vces), *Journal of Choice Modelling*, p. 100343.

Cheng, L., X. Chen, J. De Vos, X. Lai, F. Witlox (2019) Applying a random forest method approach to model travel mode choice behavior, *Travel behaviour and society*, 14, pp. 1–10.

Dekker, T., P. Koster, N. Mouter (2019) The economics of participatory value evaluation.

Friedman, J., T. Hastie, R. Tibshirani, et al. (2001) *The elements of statistical learning*, vol. 1, Springer series in statistics New York.

Geurts, K., G. Wets, T. Brijs, K. Vanhoof (2003) Profiling of high-frequency accident locations by use of association rules, *Transportation research record*, 1840(1), pp. 123–130.

Hernandez, J. I., N. Mouter, A. Itten (2021) Participatory value evaluation for relaxation of covid-19 measures.

Hillel, T., M. Bierlaire, M. Elshafie, Y. Jin (2019) Weak teachers: Assisted specification of discrete choice models using ensemble learning, in: *hEART 2019: 8th*

*Symposium of the European Association for Research in Transportation. Budapest, Hungary.*

Kaur, M., S. Kang (2016) Market basket analysis: Identify the changing trends of market data using association rule mining, *Procedia computer science*, 85, pp. 78–85.

Kendall, M. G. (1945) The treatment of ties in ranking problems, *Biometrika*, 33(3), pp. 239–251.

Keuleers, B., G. Wets, T. Arentze, H. Timmermans (2001) Association rules in identification of spatial-temporal patterns in multiday activity diary data, *Transportation Research Record*, 1752(1), pp. 32–37.

Lerman, S. R. (1976) Location, housing, automobile ownership, and mode to work: a joint choice model, *Transportation Research Record*, 610, pp. 6–11.

Mouter, N., J. I. Hernandez, A. V. Itten (2021) Public participation in crisis policymaking. how 30,000 dutch citizens advised their government on relaxing covid-19 lockdown measures, *PloS one*, 16(5), p. e0250614.

Mouter, N., P. Koster, T. Dekker (2020) 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, pp. 54–73.

Mulderij, L. S., J. I. Hernandez, N. Mouter, K. T. Verkooijen, A. Wagemakers (2021) Citizen preferences regarding the public funding of projects promoting a healthy body weight among people with a low income, *Social Science & Medicine*, 280, p. 114015.

Neill, C. L., J. Lahne (2022) Matching reality: A basket and expenditure based choice experiment with sensory preferences, *Journal of Choice Modelling*, p. 100369.

Ortelli, N., T. Hillel, F. C. Pereira, M. de Lapparent, M. Bierlaire (2021) Assisted specification of discrete choice models, *Journal of Choice Modelling*, p. 100285.

Rotteveel, A., M. Lambooij, E. Over, J. Hernandez, A. Suijkerbuijk, A. de Blaeij, G. de Wit, N. Mouter (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*, pp. 1–16.

Shiftan, Y., S. Bekhor (2020) Utilizing a random forest classifier for a methodological-iterative discrete choice model specification and estimation, in: *hEART 2020: 9th Symposium of the European Association for Research in Transportation. Lyon, France.*

Sifringer, B., V. Lurkin, A. Alahi (2020) Enhancing discrete choice models with representation learning, *Transportation Research Part B: Methodological*, 140, pp. 236–261.

van Cranenburgh, S., M. Kouwenhoven (2020) An artificial neural network based method to uncover the value-of-travel-time distribution, *Transportation*, pp. 1–39.

van Cranenburgh, S., S. Wang, A. Vij, F. Pereira, J. Walker (2022) Choice modelling in the age of machine learning-discussion paper, 42, p. 100340.

Wang, S., Q. Wang, J. Zhao (2020) Deep neural networks for choice analysis: Extracting complete economic information for interpretation, *Transportation Research Part C: Emerging Technologies*, 118, p. 102701.

Wiley, J. B., H. J. Timmermans (2009) Modelling portfolio choice in transportation research, *Transport Reviews*, 29(5), pp. 569–586.

Yao, R., S. Bekhor (2020) Data-driven choice set generation and estimation of route choice models, *Transportation Research Part C: Emerging Technologies*, 121, p. 102832.

# Chapter 4

# Explainable artificial intelligence to study Participatory Value Evaluation experiments

- Hernandez, J. I., van Cranenburgh, S., de Bruin, M., Stok, M., & Mouter, N. Using XGBoost and SHAP to explain citizens' differences in policy support for reimposing COVID-19 measures in the Netherlands. *Under review.*

Several studies examined what drives citizens' support for COVID-19 measures, but no works have addressed how the effects of these drivers are distributed at the individual level. Yet, if significant differences in support are present but not accounted for, policymakers' interpretations could lead to misleading decisions. In this study, we use XGBoost, a supervised machine learning model, combined with SHAP (Shapley Additive eXplanations) to identify the factors associated with differences in policy support for COVID-19 measures and how such differences are distributed across different citizens and measures. We use secondary data from a Participatory Value Evaluation (PVE) experiment, in which 1,888 Dutch citizens answered which COVID-19 measures should be imposed under four risk scenarios. We identified considerable heterogeneity in citizens' support for different COVID-19 measures regarding different age groups, the weight given to citizens' opinions and the perceived risk of getting sick of COVID-19. Data analysis methods employed in previous studies do not reveal such heterogeneity of policy support. Policymakers can use our results to tailor measures further to increase support for specific citizens/measures.

## 4.1. Introduction

The outbreak of the COVID-19 pandemic forced governments to adopt strategies to control multiple waves of the virus. With new variants of SARS-CoV-2 appearing (e.g., Alpha, Delta, Omicron), governments faced a trade-off between different measures that could prevent new infections, avoid further deaths due to COVID-19 and reduce the risk of overloading the healthcare system. However, such measures would also increase psychological stress and impact the economy, which in turn would hinder the citizens' support and decrease adherence. By understanding what factors explain the citizens' policy support for COVID-19 measures, governments can prioritise those measures that are effective in curbing the spread of the virus and, at the same time, are widely accepted.

Previous studies shed light on the factors that explain the support for COVID-19 measures. These studies conclude, for instance, that higher policy support for COVID-19 measures is associated with the citizens' trust in institutions (Dohle et al., 2020; Gotanda et al., 2021), perceived risk, sociodemographic characteristics (Mouter et al., 2022; Sicsic et al., 2022) and geographical factors (Loria-Rebolledo et al., 2022). However, a key limitation of most of these works is that their data analysis methods, namely regressions, discrete choice models or latent class cluster analysis (LCCA), can only account for "average" effects on the policy support, either at the population level or pre-defined groups, ignoring that the support for COVID-19 measures could vary significantly across different measures and/or different citizens. For instance, regressions and discrete choice models provide outcomes that are interpretable for a representative citizen or specific measure, while LCCA identifies different groups of citizens and characterises them in terms of averages within each group. Nevertheless, none of these methods are suitable for identifying citizen- or measure-specific effects, as the number of required parameters or latent classes would lead to a computationally intractable model. For this reason, analysts rely on simpler and tractable specifications, at the expense of not being able to uncover deeper levels of heterogeneity. However, if differences in preferences across citizens or measures are substantial but not accounted for, the interpretations done by the analyst could lead to misleading decisions.

Supervised machine learning (ML) models can overcome the previously named limitations. Supervised ML models aim to predict one or more response variables as a function of a set of covariates (called features). XGBoost, a supervised ML model based on gradient-boosted trees, can learn complex interactions between covariates and individual effects without the need of being previously specified by the analyst, reaching a high prediction performance and, at the same time, overcoming the limitations of data analysis methods previously used to model policy support for COVID-19 measures. But like many other ML methods, XGBoost only provides an overall importance level of each covariate for predicting the response variable, which makes XGBoost relatively 'opaque' in terms of explainability. So-called explainable AI (XAI) methods can overcome this limitation of XGBoost. XAI methods aim to provide explanations from an otherwise 'opaque' ML model. An XAI method that gained popularity nowadays in literature is SHAP, an approach based on game theory that identifies how much each covariate of a ML model contributed to each individual prediction. Combined with the predictive capabilities of XGBoost, SHAP provides an opportunity to quantify the contribution of each of these factors to the differences in policy support at the respondent level, allowing analysts to identify how these differences are distributed across different citizens and spot nonlinear or opposite effects per measure.

This paper aims for two goals. Firstly, we use XGBoost combined with SHAP to identify what factors are associated with differences in policy support for COVID-19 measures and how such differences are distributed across different citizens and measures, thus departing from a conventional interpretation of "average" effects across (groups of) citizens. Secondly, we compare and contrast the findings of XGBoost and SHAP with alternative data analysis methods employed to analyse citizens' support for COVID-19 measures, namely a choice model and LCCA. We compare the extent to which each method allows for different interpretations and the extent that SHAP adds to the other two methods. We use data from a Participatory Value Evaluation (PVE) experiment conducted in the Netherlands to infer the Dutch citizens' preferences for reimposing a set of COVID-19 measures under different risk scenarios (Mouter et al., 2022).

## 4.2. Experiment and data

### 4.2.1. Preference elicitation method: PVE experiment

PVE experiments have been applied in diverse fields, including COVID-19 measures (Mouter et al., 2021a) , healthcare investments (Mulderij et al., 2021; Rotteveel et al., 2022) and public infrastructure projects Mouter et al. (2021b). In a PVE experiment, respondents are asked to imagine a certain scenario and then choose a combination of policy alternatives for addressing the scenario. In the PVE experiment used in this paper, four different scenarios were designed, describing different levels of COVID-19 threat and the current hospital overcrowding risk (see Table 4.1).

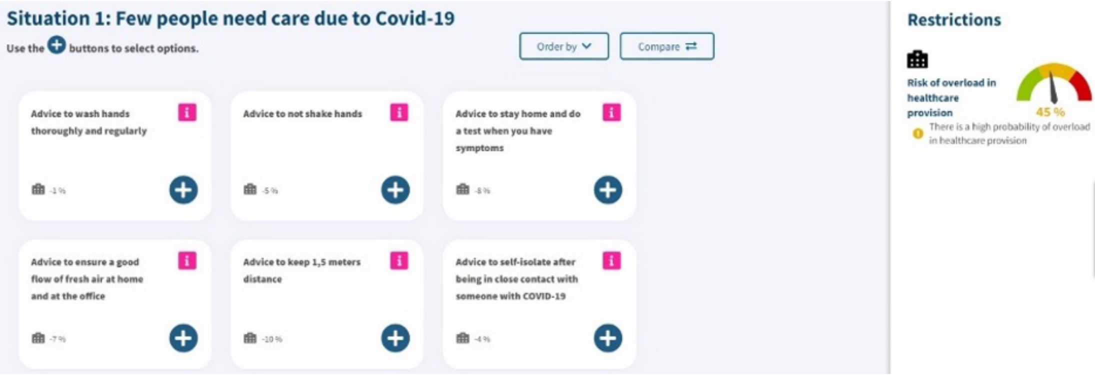
| Scenario | Description | Number of possible COVID-19 measures | Initial hospital overcrowding risk |
|---|---|---|---|
| Scenario 1 | Hospitalizations are at a low level. No operations are postponed. There is no dangerous new variant of the virus. | 9 | 45% |
| Scenario 2 | Autumn has begun, and COVID-19 spread faster. Hospitalisations of vulnerable people and non-vaccinated increase. Minor surgeries are postponed. Basic rules are imposed (i.e., wash your hands, keep 1.5 metres distance and get tested in case of symptoms). | 14 | 69% |
| Scenario 3 | A new variant that spreads faster is found in another country, but it is not clear whether this variant generates more severe symptoms. Restrictions for entering the country are imposed, as well as the basic rules. There is a risk that major surgeries in hospitals will be postponed. | 14 | 60% |
| Scenario 4 | A new variant is found in another country, which spreads faster, and it is clear that many people have severe symptoms from this variant. In addition to the basic rules and entry restrictions, more severe measures are in place (e.g., capacity limits to the hospitality industry and no massive events). The healthcare capacity is at its limits, and if no measures are taken, major surgeries will be postponed, and there is a risk that patients will no longer be able to go to a hospital. | 13 | 100% |

*Table 4.1: Description of scenarios of the PVE experiment*

Each scenario was embedded in an independent PVE experiment choice task. For every scenario, a list of possible policy alternatives was presented. By choosing a policy alternative, the hospital overcrowding risk is reduced in a specific percentage within predefined ranges (see Table 4.2). In scenarios 1, 2 and 3, respondents were allowed to choose any combination of policy alternatives, whereas in scenario 4, they must choose a combination that results in at least a 30% reduction in the hospital overcrowding risk. To avoid cognitive burden, each respondent answered three scenarios: scenarios 1 and 2 are always answered, while scenarios 3 and 4 were randomly assigned. The PVE experiment choice tasks were embedded in a web survey. After the presentation of an instruction video, respondents were presented with the PVE choice tasks (see an example in Figure 4.1). Policy alternatives with their respective reductions of the hospital overcrowding risk are presented in the left-side pane, whereas the total hospital overcrowding risk is detailed in the right-side pane as an interactive gauge. After answering the choice tasks, respondents have to fill out a questionnaire about their sociodemographic profile (e.g., gender, age, living province) and perception questions (e.g., perceived risk of being affected by a COVID-19 infection, the weight they believe governments should give to scientists or citizens' opinion, etc.)

| Measures | Risk red. range | Scenarios<br>S1 | S2 | S3 | S4 |
|---|---|---|---|---|---|
| Advice to wash hands frequently and thoroughly | 1-3 | X | | | |
| Advice to stay at home with COVID-19 symptoms and to do a test | 3-5 | X | | | |
| Advice not to shake hands | 8-14 | X | | | |
| Advice to ventilate | 3-7 | X | | | |
| Advice to keep 1.5 metres distance | 7-13 | X | | | |
| Quarantine if in intensive contact with person infected with COVID-19 | 4-8 | X | | | |
| Advice to work at home a few days a week | 2-4 | X | | | |
| Advice to work at home, unless it is absolutely necessary | 6-10 | | X | X | X |
| Mouth mask obligation in public transport/shops/hospitality industry | 2-6 | X | X | X | |
| Vaccination passport hospitality industry (2G or 3G) | 3-5 | X | X | X | |
| Vaccination passport for people working with vulnerable people | 5-8 | | X | X | X |
| Vaccination passport except in schools, work and essential shops | 4-10 | | X | X | X |
| Encourage self-testing by making it available free of charge | 6-10 | | X | X | |
| Starting a booster campaign which starts with vulnerable people | 10-15 | | X | X | X |
| Requiring shops to offer time slots for people with vulnerable health | 5-8 | | X | X | |
| Limit number of customers per square metre in non-essential shops | 1-3 | | X | | |
| Pick up orders in non-essential shops | 2-4 | | X | X | X |
| 1/3 capacity and fixed seating at events | 2-6 | | X | X | |
| Banning festivals and major sporting events | 4-8 | | X | X | |
| Strict advice not to have more than 2 visitors per day at home | 5-10 | | X | X | X |
| Advice higher education online and maximum number of students per college | 4-8 | | X | X | X |
| Lockdown after 5pm | 4-8 | | | X | |
| Lockdown after 8pm | 8-10 | | | | X |
| Closing restaurants/cafés | 10-15 | | | | X |
| Closing sports venues | 5-10 | | | | X |
| Closing cinemas, theatres, concert halls | 5-10 | | | | X |
| Closing primary/secondary schools | 15-20 | | | | X |

*Table 4.2: COVID-19 measures per scenario, adapted from (Mouter et al., 2022)*


Situation 1: Few people need care due to Covid-19

Use the ⊕ buttons to select options.

Advice to wash hands thoroughly and regularly -1 %

Advice to not shake hands -5 %

Advice to stay home and do a test when you have symptoms -8 %

Advice to ensure a good flow of fresh air at home and at the office -7 %

Advice to keep 1,5 meters distance -10 %

Advice to self-isolate after being in close contact with someone with COVID-19 -4 %

Risk of overload in healthcare provision 45 %

There is a high probability of overload in healthcare provision

*Figure 4.1: Example choice task presented in the PVE experiment for scenario 1*

## 4.2.2. Data

The data collection was conducted between 3 to 10 February 2022, when the Dutch government considered the ease of some COVID-19 measures. The sample was collected by a specialised survey company (Dynata) using a representative panel. After cleaning missing values and no responses, the final dataset comprises 1,888 responses and 15 variables (see Table 4.3).

| Type | Variable | Description | Values |
|---|---|---|---|
| Experimental feature | Overcrowding risk reduction | Reduction of risk of overloading the healthcare system (%) derived from each specific COVID-19 measure. | Numeric (see table 2) |
| Sociodemographic characteristics | Gender | Gender of the respondent. | Male<br>Female |
| | Age | Age group | 18-24 years<br>25-34 years<br>35-44 years<br>45-54 years<br>55-64 years<br>65-74 years<br>75 or more years |
| | Education | Education level. | Low<br>Medium<br>High |
| | Province | Living province. | Categorical, per province |
| | City size | Size category of the living city of respondents. | Village<br>Small city<br>Medium city<br>Big city |
| | Work status | Type of work/work status. | Full-time<br>Part-time<br>Retired<br>Incapacitated<br>Student<br>Unemployed |
| Vaccination status | Vaccinated | Whether the respondent has at least one COVID-19 vaccine | No<br>Yes |
| | Boosted | Whether the respondent has a booster shot | No<br>Yes |
| Perception indicators | Risk (infected) | Perceived risk of getting infected by COVID-19 | No risk<br>Small risk<br>Moderate risk<br>High risk<br>Extreme risk |
| | Risk (getting sick) | Perceived risk of getting very sick by COVID-19 | Same as risk (infected) |
| | Risk (hospitalised) | Perceived risk of getting hospitalised by COVID-19 | Same as risk (infected) |
| | Risk (death) | Perceived risk of dying by COVID-19 | Same as risk (infected) |
| | Weight citizens' opinion compared to scientists' opinion | The weight that a respondent believes the government should put on the opinion of citizens in this survey with respect to the opinion of scientists and experts. Higher values indicate a higher value to the citizens' opinion. | Only to citizens<br>Citizens more than scientists<br>Citizens equally to scientists<br>Scientists more than citizens<br>Only scientists |
| Response variables | Choice | Response variables per measure. They indicate if a specific COVID-19 measure was chosen in the scenario. | No<br>Yes |

*Table 4.3: Variables used in this study*

We considered 14 covariates (independent variables) in this study, selected based on previous studies, including the first analysis of this PVE experiment (Mouter et al., 2022). We distinguish between four covariates types: experimental features, sociodemographic characteristics, vaccination status, and perception indicators. Regarding the experimental features, we include the overcrowding risk reduction of each COVID-19 measure. The sociodemographic characteristics considered in this study are the respondents' gender, age group, education level, living province, city size and work status. The vaccination status is divided into two covariates: whether the respondent is vaccinated at least once, and whether they received a booster shot. The first set of perception indicators considered in this study is the respondent's perceived risk that their health would be affected by COVID-19 in four levels: getting infected by the virus, getting very sick, being hospitalised and dying due to the disease. The final perception indicator is the respondent's weight they think the government should give to the citizens' opinion, relative to the scientists' opinion. Finally, the response variables (Choice) are binary variables equal to one if a COVID-19 measure was chosen by the respondent and zero otherwise. Each measure is associated with an independent response variable, and the response variables for the same measure are independent across scenarios.

## 4.3. Methods

Data is analysed using XGBoost (Chen & Guestrin, 2016), a specific ML model of the family of tree-boosting models. XGBoost was chosen among alternative ML models (i.e. neural networks and Random Forests) since tree-boosting models have been proven to be robust to overfitting and, furthermore, reaching higher predictive performance in choice data (Wang et al., 2021)[1] . Then, SHAP is applied in the trained (estimated) XGBoost model to uncover what relations XGBoost has learned from the data and explain the policy support for COVID-19 measures. Finally, the outcomes of SHAP are visualised and interpreted. The following subsections describe XGBoost, SHAP and the use of their outcomes.

### 4.3.1. XGBoost

XGBoost is a ML system for tree-boosting. Tree-boosting is an algorithm that combines the outcomes of a set of decision tree (DT) models to form a model with higher predictive performance. A DT is a ML model that predicts a response variable

[1] Furthermore, no considerable model fit improvements were found with alternative ML models in preliminary tests.

$Y$ as a set of conditions that the set of covariates $X$ must hold, forming a tree structure. Given a response variable $Y$ and a set of covariates $X$, the tree-boosting algorithm aims to predict $Y$ as detailed in equation (1):

$$\hat{Y} = \hat{f}_T(X) = \sum_{t=1}^{T} \hat{f}_t(X), \tag{4.1}$$

where $T$ is the number of DT models, $\hat{f}_T$ is the tree-boosted model and $\hat{f}_t$ is the t-th DT model. In the tree-boosting algorithm, each DT model is added sequentially. On each iteration, the new DT model corrects the mispredictions of the tree-boosted model that is formed thus far. Mathematically, the tree-boosting algorithm optimises a loss function $l(\cdot)$ that depends on the response variable $Y_i$ at a step $t$ and the predictions $\hat{Y}_i^{(t-1)}$ of the previous $t-1$ models, plus a regularisation term $\Omega(\cdot)$. On each step $t$ of the tree-boosting algorithm, the overall loss function can be written as in equation (2):

$$L^{(t)} = \sum_{i=1}^{N} l\left(\hat{Y}_t^{(t-1)}, Y_i\right) + \sum_{t=1}^{T} \Omega(\hat{f}_t). \tag{4.2}$$

The main challenge of tree-boosting is to optimise this loss function. XGBoost implements a specific form of boosting named gradient-boosting (Friedman, 2001). In gradient-boosting, given an expression for $l(\cdot)$ that depends on the type of response variable (e.g., for binary responses, $l(\cdot)$ is a log-loss function), the loss $L^{(t)}$ is expressed in terms of its gradient and hessian. By doing so, the loss function becomes tractable for optimisation.

Our implementation of XGBoost employs three fixed hyperparameters and two hyperparameters that are selected such that the loss function is minimised (table 4.4 details the employed hyperparameters). The latter type of hyperparameters are selected using a grid search process, in which each possible combination of hyperparameters are used to train the XGBoost model using a 10-fold cross validation. The average loss is computed and the final model is the one for which the average loss is minimum. For all scenarios, the optimal hyperparameters are a Gamma value equal to 2, a maximum tree depth equal to 3 and a minimum child weight equal to 5.

After selecting the optimal hyperparameters, the training process was done using a combination of 10-fold cross validation and a split sample. On each scenario, a random split of the data is done: 80% of the sample is used for training, and the remaining 20% is left as a holdout (test) sample. The training process is performed using 10-fold cross validation using the training sample only. After the model is trained, predictions and SHAP values are computed with the holdout sample.

| Hyperparameter | Description | Candidate Values |
|---|---|---|
| Loss function | The objective loss function to optimise. | binary:logistic (fixed) |
| Evaluation metric | Metric evaluated on each iteration as stopping criterion of the optimisation algorithm | logloss (fixed) |
| Gamma | Minimum loss reduction for making a partition of the tree models. A higher value implies a higher regularisation at the risk of underfitting. | 0, 1, 2 |
| Maximum tree depth | Maximum number of levels of the tree models. Higher depth can capture more interactions, at the risk of overfitting the final model. | 3, 5, 7 |
| Minimum child weight | Minimum sum of weights of a child node. Higher values prevent the algorithm to make too much splits on the trees, at the risk of underfitting the final model | 1, 2, 3, 5 |

*Table 4.4: Hyperparameters used in XGBoost*

### 4.3.2. SHAP

SHAP (Lundberg et al., 2017) is a technique to provide explanations for an otherwise “opaque” ML model. SHAP calculates how much each covariate contributes to the prediction of each respondent of the sample with respect to the average prediction in terms of Shapley values. Shapley values are a concept of coalitional game theory that describes the distribution of payments across coalitions of players in a cooperative game.

While SHAP has gained increasing popularity in the ML field, its use for choice problems has been rather minor and recent. A brief literature review shows that the use of SHAP to address choice problems has been scoped mostly in the transportation field (e.g., Dong et al., 2022; Ji et al., 2022; Jin et al., 2022; Lee, 2022). For instance, Dong et al. (2022) use SHAP in an artificial neural network to explain individual and general route choice behaviour from GPS data in South Korea; Ji et al. (2022) applies SHAP in an XGBoost model to uncover interactions between covariates that explain Cyclists’ behaviour in China; Jin et al. (2022) compares the explanations from gradient-boosting methods and SHAP with the interpretations of a multinomial logit model to explain vehicle transactions in the United States; and Lee (2022) uses SHAP and XGBoost to explain the decision of giving up the use of public transport during the COVID-19 pandemic in South Korea. To the authors’ knowledge, the only applications of SHAP outside the transportation field are Wang et al. (2022), who use SHAP and a series of ML models (e.g., Random Forests, neural networks, XGBoost) to explain the decision of getting online healthcare in China, and this work.

SHAP relates ML with game theory by assuming that a set of covariates $X_n = \{x_{n1}, x_{n2}, \dots\}$ for a specific respondent n are players in a game that consists of predicting the response variable $Y_n$. The game is the ML model, and the payoffs are the predictions $\hat{f}(X_n)$. Each covariate can contribute to the prediction standalone or forming a coalition with one or more other covariates. The Shapley value $\phi_{nk}$ of a covariate value $x_{nk}$ for a respondent n is the averaged marginal contribution of $x_{nk}$ to predict $Y_n$, across all possible coalitions (Molnar, 2020), given by equation (3):

$$\phi_{nk} = \sum_{S \subset \{1,\ldots,K\} \setminus \{k\}} \frac{|S|(K-|S|-1)!}{K!} \left(\hat{f}_x(S \cup k) - \hat{f}_x(S)\right). \tag{4.3}$$

where $S$ is a subset of the covariates of the model, $K$ is the number of covariates, and $\hat{f}_x(S)$ is the prediction for the covariates in set $S$ marginalised over the covariates that are not included in $S$.

The outcome of SHAP is a matrix $N \times K$ of SHAP values, computed per response variable. In other words, SHAP values are computed at each respondent's level, per covariate and per response variable (i.e., per COVID-19 measure).

SHAP values satisfy the properties of local accuracy, missingness and consistency (Lundberg et al., 2017). Local accuracy guarantees that the sum of SHAP values for a respondent n is equal to the difference between the prediction for n and the average prediction across all respondents. Missingness guarantees that if a covariate value $x_{nk}$ is missing, then its SHAP value is zero, thus not affecting the local accuracy property. Consistency guarantees that if the contribution of $x_{nk}$ increases, then its SHAP value also increases.

SHAP presents three key advantages over alternative XAI methods, such as the Local Interpretable Model-Agnostic Explanations (LIME) proposed by Ribeiro et al. (2016) and Layer-Wise Relevance Propagation (LRP) proposed by Bach et al. (2015). Firstly, SHAP bases its explanations on computing Shapley values, which makes this method theoretically robust and stable compared to LIME and LRP, which base their explanations on random perturbations over the dataset. Secondly, SHAP is model agnostic, similar to LIME, but different from LRP, which is specific to neural networks. Therefore, SHAP can be used on any supervised ML model. Thirdly, SHAP allows for both local and global explanations, since the computed Shapley values can be aggregated (i.e., averaged) to explain the mean contribution of each covariate.

### 4.3.3. Using the outcomes of SHAP: SHAP importances and visualising SHAP values

SHAP values are used in two stages (see Table 4.5). In the first stage, we compute so-called SHAP importances. SHAP importances quantify the importance of each covariate, and they are computed as the absolute value of the SHAP values, averaged across respondents. Higher (lower) SHAP importances indicate that, on average, a covariate has a greater (smaller) effect on the policy support for COVID-19 measures. Thus, the analyst should prioritise interpreting covariates with high SHAP importances. It is important to notice that a low SHAP importance does not necessarily mean that a covariate has a negligible effect, but it means that such effect is smaller than the effect of other covariates. Thus, we use SHAP importances to identify, visualise and interpret the three most relevant covariates in all three risk scenarios, while a detailed visualisation of all covariates per scenario is presented in Appendix 4.A. In the second stage, the SHAP values identified as important are visualised per COVID-19 measure and risk scenario. Visualising SHAP values consists of describing the respondent-specific SHAP values in plots to facilitate their interpretation. Through visualisations of SHAP values, the analyst can identify, for instance, how the effects of specific covariates on the policy support for COVID-19 measures are distributed across respondents, identify observed heterogeneity or nonlinear effects that are difficult to elucidate from a direct (i.e., non-visual) inspection of the SHAP values.

| | Stage 1: SHAP importances | Stage 2: Visualisation of SHAP values |
|---|---|---|
| Definition | The absolute SHAP values averaged across respondents. | The SHAP values associated with a covariate, presented in plots |
| Information that provides | The average effect of a specific covariate on the policy support for COVID-19 measures. | The distribution (sparsity) of the effects of specific effects and nonlinear effects |
| Type of interpretation | Numerical [0,1] | Visual (plots) |
| Meaning | <u>If low:</u><br>The covariate has a small effect on the policy support for COVID-19 measures, on average. | <u>Summary plot:</u><br>It plots the SHAP values per covariate, sorted by their importance. Each plot shows the distribution of the effects associated with a covariate for the policy support for a specific COVID-19 measure. Each point is coloured according to its associated covariate value. |
| | <u>If high:</u><br>The covariate has a great effect on the policy support for COVID-19 measures, on average. | <u>Scatter plot:</u><br>It plots the SHAP values for a specific covariate. It allows for identifying nonlinear effects in the policy support for a specific COVID-19 measure. |

*Table 4.5: Summary of approaches to interpret SHAP values used in this paper*

We consider two visualisations: SHAP summary plots and SHAP scatter plots. Summary plots detail the distribution of SHAP values associated with a specific covariate. Summary plots allow identifying the magnitude, direction and distribution of the effects on the policy support for COVID-19 measures. Each point of the summary plot is the SHAP value of a specific respondent associated with a specific covariate. The horizontal axis details the magnitude of the SHAP value. If two SHAP values are of similar value, they are stacked vertically. SHAP values are coloured according to the covariate values to detail the direction of the effects of each covariate. SHAP scatter plots detail the relationship between a specific covariate with its associated SHAP values. The vertical axis of the scatter plot details the magnitude of the SHAP values associated with a specific covariate, whereas the horizontal axis details the values of such covariate. SHAP scatter plots allow analysts to identify how the effects on the policy support for a COVID-19 measure are distributed across the values of a specific covariate. From a scatter plot, the analyst can identify nonlinear effects or specific effects per groups of respondents.

## 4.4. Results

### 4.4.1. SHAP importances

We compute the SHAP importances per risk scenario, averaged across respondents and COVID-19 measures (see Table 4.6). In addition, the average SHAP importance across risk scenarios is calculated (last column) to identify which covariates are the most (least) important across scenarios, on average.

| | Scenario 1 | Scenario 2 | Scenario 3 | Scenario 4 | Average |
|---|---|---|---|---|---|
| Gender | 0.013 | 0.011 | 0.019 | 0.011 | 0.014 |
| Age | **0.032** | **0.025** | 0.023 | **0.029** | **0.027** |
| Education | 0.017 | 0.015 | 0.018 | 0.009 | 0.015 |
| Province | 0.022 | 0.019 | 0.021 | **0.021** | 0.021 |
| City size | 0.008 | 0.011 | 0.014 | 0.015 | 0.012 |
| Work status | 0.021 | 0.018 | **0.026** | 0.019 | 0.021 |
| Vaccinated | 0.021 | 0.016 | 0.014 | 0.011 | 0.016 |
| Boosted | 0.013 | 0.024 | 0.023 | 0.013 | 0.018 |
| Risk (infected) | 0.015 | 0.010 | 0.012 | 0.012 | 0.012 |
| Risk (getting sick) | **0.029** | **0.026** | **0.027** | 0.013 | **0.024** |
| Risk (hospitalised) | 0.011 | 0.013 | 0.016 | 0.020 | 0.015 |
| Risk (death) | 0.019 | 0.015 | 0.019 | 0.017 | 0.017 |
| Weight citizens'/scientists' opinion | **0.027** | **0.032** | **0.039** | **0.021** | **0.030** |
| Overcrowding risk reduction | 0.011 | 0.008 | 0.012 | 0.012 | 0.011 |

*Table 4.6: SHAP importances per risk scenario. The filling intensity details a higher importance per scenario*

On average, the most important covariates are, in descending order, the weight of citizens'/scientists' opinion, age and the perceived risk of getting sick of COVID-19.

These three covariates are also the most important in all scenarios, except in scenario 3, where work status becomes the third-most important covariate. On the other hand, the overcrowding risk reduction generated by the measures is consistently ranked as one the least important covariates. These results indicate that sociodemographic characteristics and perception indicators explain better the differences in the policy support for COVID-19 measures than the resulting reductions in the risk of overloading the healthcare system. In the following subsections, we focus on the visualisation of SHAP values of age, the perceived risk of getting sick of COVID-19 and the weight of citizens'/scientists' opinion.

### 4.4.2. Visualising SHAP values

In this subsection, we present visualisations of the SHAP values for the three most important covariates, namely the age group, the weight of citizens'/scientists' opinion and the perceived risk of getting sick of COVID-19. A complete set of summary plots per covariate, measure and scenario is provided in Appendix 4.A.

#### Age group

We generate summary plots of the SHAP values associated with age per COVID-19 measure and risk scenario (see Figure 4.2). As a first observation, the overall effects tend to be smaller for scenario 1 (less severe) compared to the other risk scenarios. Aside from the findings in line with previous studies, i.e., older age is associated with higher policy support, visual inspection of the summary plots confirms heterogeneous distributions of the effects, potential nonlinear effects and effects with an opposite direction for specific measures.

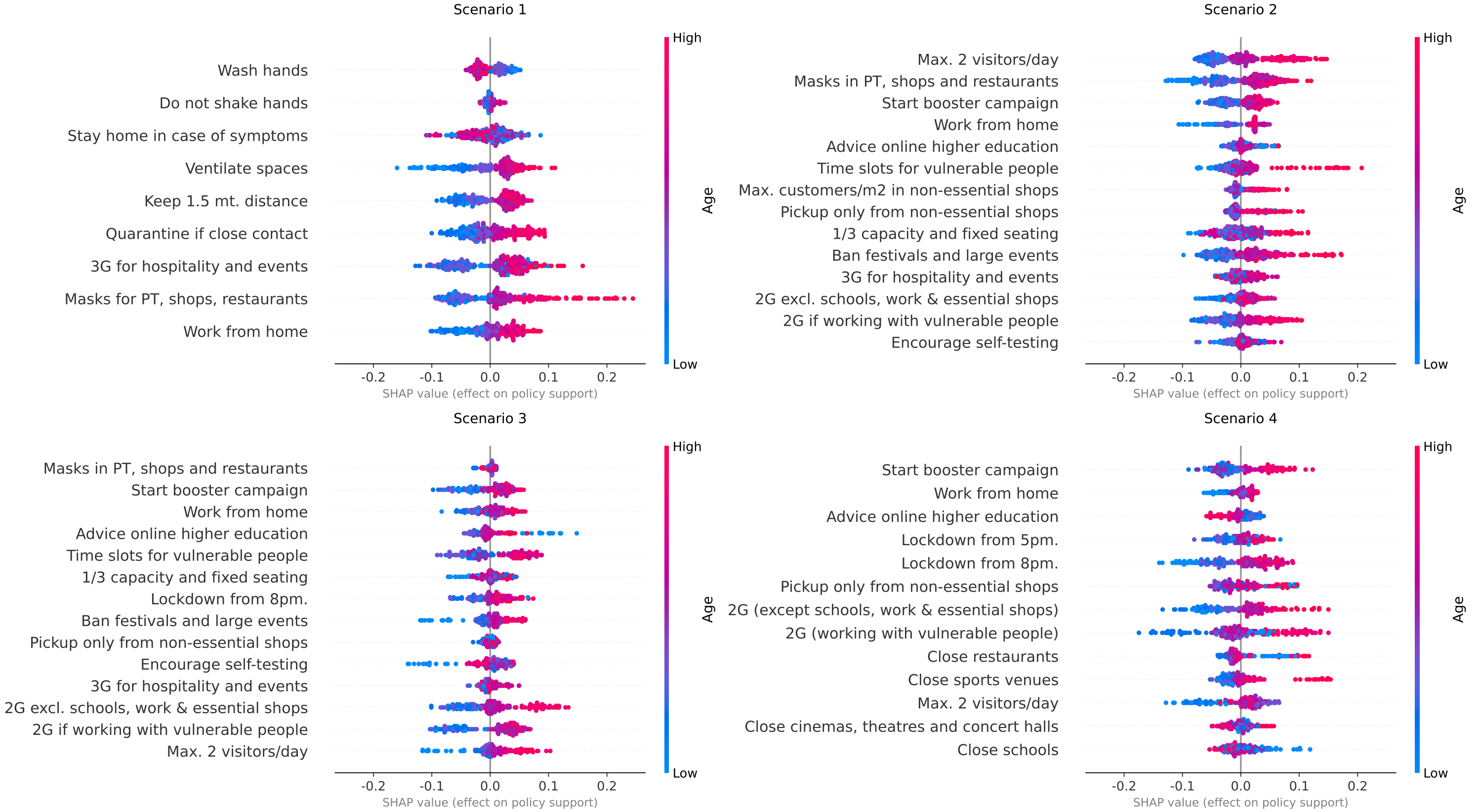


*Figure 4.2: SHAP summary plots of age group, per measure and risk scenario*

Heterogeneous distributions of the effects are shown in summary plots either as clusters of SHAP values agglomerated in one or more locations or as a line of SHAP values sparsely distributed in a plot. Clusters are associated with groups of respondents with a similar effect on policy support. In contrast, sparsely distributed effects indicate differences in policy support for respondents that belong to an age group. For instance, the SHAP values associated with an advice to work from home in scenario 1 and to receive a maximum of two visitors per day in scenario 2 present three clusters of effects, with a first cluster associated with a lower effect on policy support and low age, a second cluster associated with close-to-null effects and middle age, and a third cluster associated with higher effect and older age. Sparse distributions are observed, for instance, for the advice of having maximum 2 visitors per day at home in scenario 3 or a 2G COVID-19 certificate for those who work with vulnerable people in scenario 4, where the sparse effects are associated with the extreme age groups, indicating clear differences on the policy support for such measures across respondents of the extreme age groups.

Nonlinear effects are shown in summary plots as SHAP values with similar effects (i.e., close together) but associated with different age groups. An example of nonlinear effects is with the imposition of a 3G COVID-19 certificate for public transport, shows and restaurants in scenario 1. While visual inspection confirms that older age is associated with higher policy support for the measure, there is a group of points associated with middle age (coloured in purple) located in the lower tail of the plot, indicating that such respondents have low policy support comparable with respondents of the lowest age group. A scatter plot (see figure 4.3) confirms that the effect of age for implementing this measure resembles a piecewise-linear function. Age groups between 25 and 44 years old are associated with negative SHAP values, while from 45 years and older, the SHAP values are positive. The effect does not seem to be increasing or decreasing within each of the two groups but remains constant, with a jump at 45-54 years old and then remaining constant.

As another example, the SHAP values associated with imposing a COVID-19 certificate (2G) for those who work with vulnerable people in scenario 4 present a region of points around zero (no effect) and positive values associated with the lowest age. Further inspection with a scatter plot (figure 4.4) show clear differences in the policy support for such measure per specific age group. The group of 18-24 years old has dispersed effects around zero and above. The age groups between 25 and 64 years old are associated with negative support, being respondents of 25-34 years old the group with the lowest support. The group of 65 years old or more are the respondents with positive support for this measure.

Finally, the effect of age on the policy support of certain measures goes in the

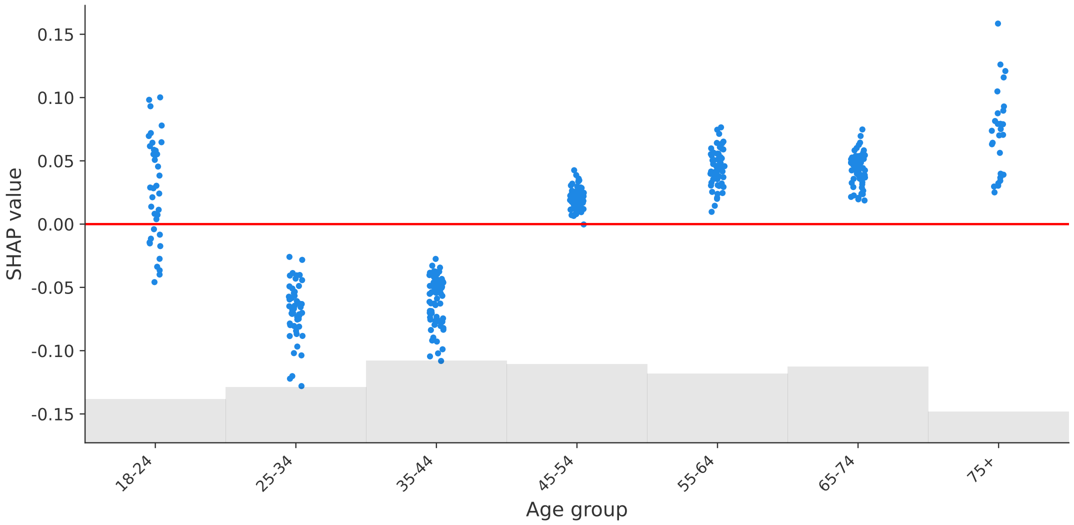


*Figure 4.3: SHAP values of age of implementing a 3G COVID-19 certificate for scenario 1*

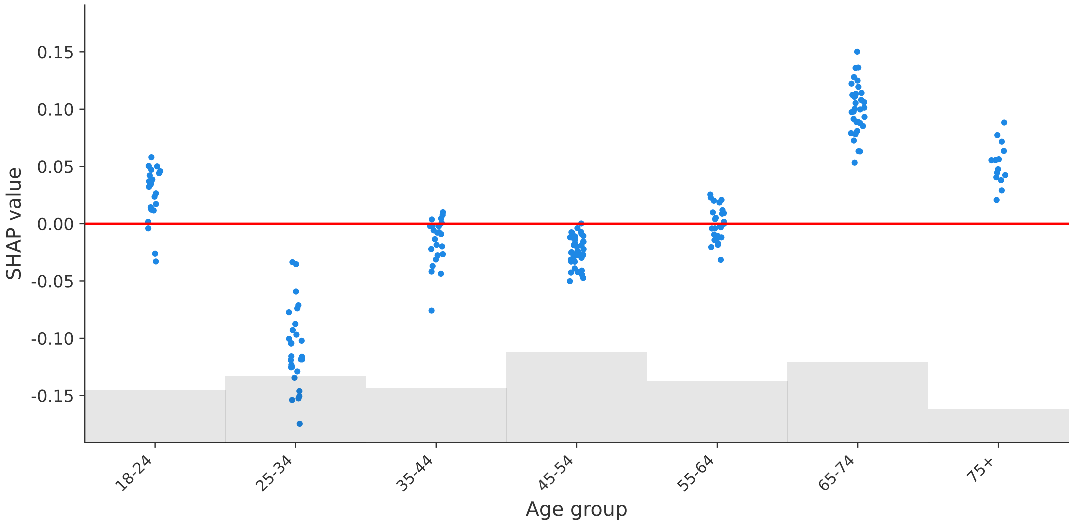


*Figure 4.4: SHAP values of age of implementing a 2G COVID-19 certificate for those who work with vulnerable people in scenario 4*

opposite direction than expected for specific measures (recall Figure 4.2). For instance, we observe that some people (points) of the lowest age groups are associated with higher policy support for advising online higher education in scenarios 3 and 4, and for closing schools in scenario 4, since these measures are likely not to affect them directly as they have lower chances of having children, compared to middle and older age groups.

#### Weight citizens' opinion compared to scientists' opinion

We generate the SHAP summary plots for the Weight citizens' opinion compared to scientists' opinion per COVID-19 measure and risk scenario (see Figure 4.5). As a first result, we observe that respondents who believe the government should weigh the citizens' opinion more than the opinion of scientists are associated with lower policy support for COVID-19 measures, and vice versa for respondents who give more weight to scientists' opinion. This result was not explored further in the previous analysis of this PVE experiment, despite this covariate being important for explaining the differences in policy support. Furthermore, SHAP summary plots evidence heterogeneous effects, either in clusters (agglomerations) of effects and sparse distributions or a combination of both. A combination of clusters of effects and sparse distribution is observed in a summary plot as one or more groups of SHAP values associated with a specific group of covariate values (i.e., the values of the weight of citizens'/scientists' opinion), followed by a line of points associated with the rest of respondents, or vice versa. For example, consider the SHAP summary plot for the advice of working from home in scenario 2. On the one hand, respondents who believe the government should only consider citizens' opinion are associated with lower policy support for this measure, and such effect widely differs across respondents, illustrated by the blue line of points. This result indicates strong differences in the support for this measure for respondents with the same perception about the weight the government should give to citizens' opinion. On the other hand, for the same measure, respondents who believe the government should give more opinion to scientists' opinion are associated with higher policy support, and they are concentrated in a single cluster, and hence they have a similar effect on the support for this measure.

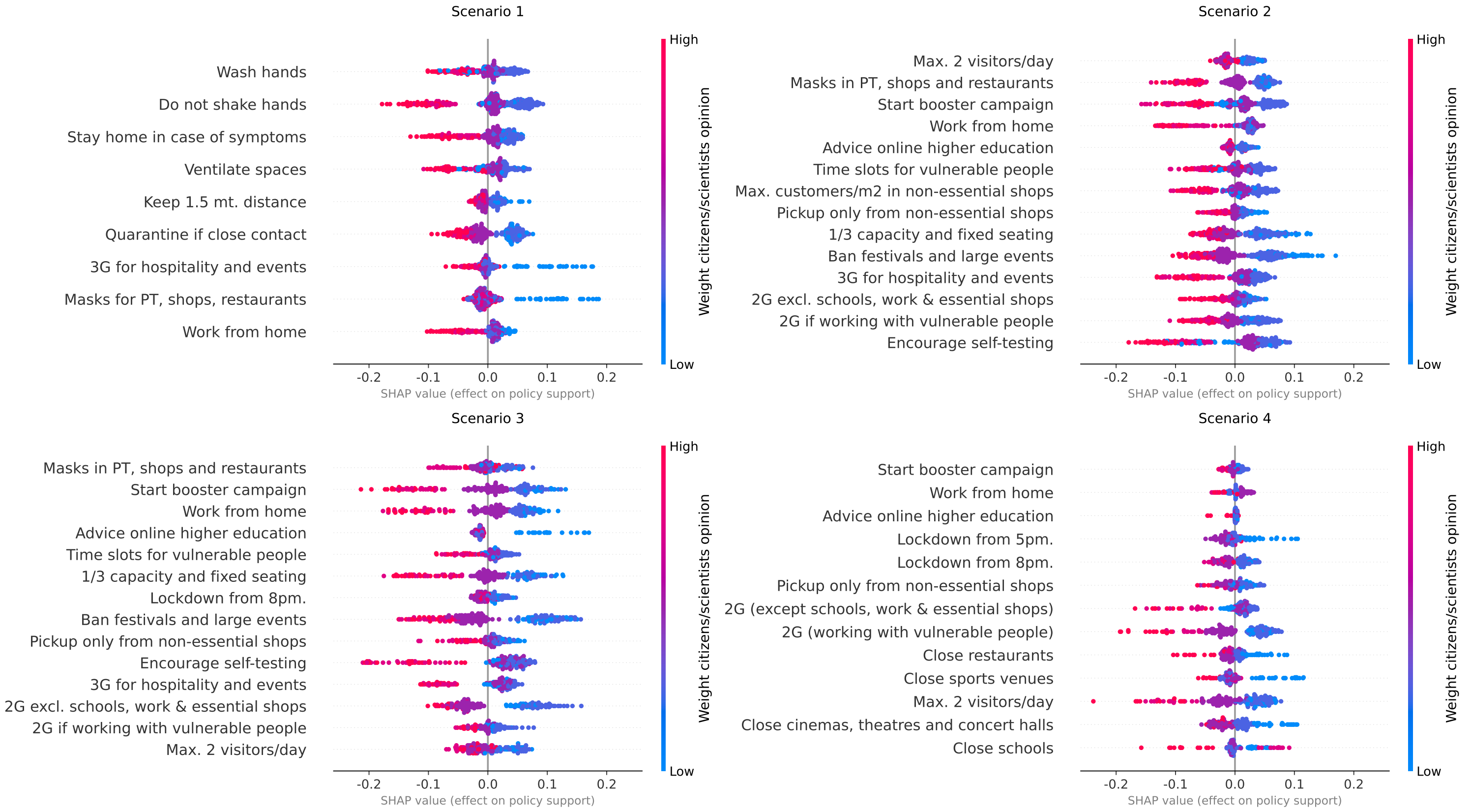


*Figure 4.5: SHAP summary plots of the Weight citizens' opinion compared to scientists' opinion, per measure and risk scenario*

### Perceived risk of becoming sick of COVID-19

We generate the summary plots of the perceived risk of becoming sick of COVID-19, per measure and risk scenario (see Figure 4.6). As a difference with the previous covariates, sparsity of effects is more observed for respondents with a stronger opinion, i.e., with the highest and lowest perceived risk of becoming sick of COVID-19. In contrast, respondents with a moderate opinion are concentrated in a cluster close to the origin. As expected, the range of SHAP values is higher in scenarios 1, 2 and 3 since this covariate was one of the most important, whereas for scenario 4, the range of SHAP values is considerably shorter. Nevertheless, further inspection of SHAP values per scenario confirms differences in the importance of this covariate between specific measures in the same scenario. For instance, in scenario 2, for imposing mandatory masks, starting a booster campaign, working from home and encouraging self-testing, the range of SHAP values is considerably higher than for the rest of the measures in the same scenario. This is a sign that, for these measures, the perceived risk of getting sick of COVID-19 is of considerably higher importance than for the other measures in this scenario.

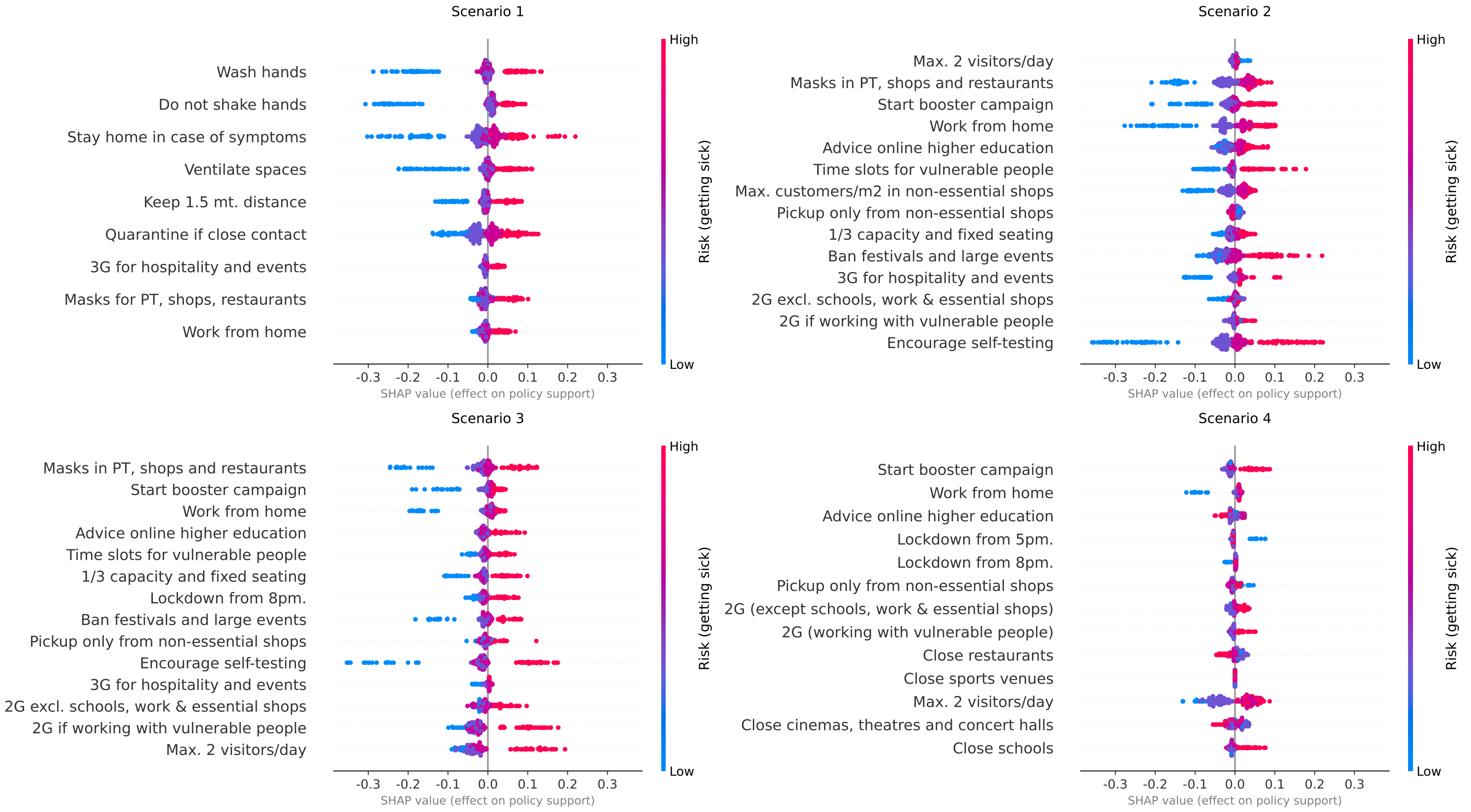


*Figure 4.6: SHAP summary plots of the perceived risk of getting sick of COVID-19, per measure and risk scenario*

### 4.4.3. Contrasting SHAP with choice modelling analysis and LCCA

The findings obtained with SHAP are contrasted with the results obtained from a choice model and LCCA for scenario 1 (see Table 4.7). The choice model and LCCA correspond to the models used by Mouter et al. (2022). We estimated a new version of the choice model, in which the same covariates of this study are included per COVID-19 measure, as a difference from the original study, in which only a set of constants and a single parameter for the overcrowding risk reduction were estimated. The results and the choice model are detailed in Appendix 4.B. The results of the LCCA presented in this section are from Mouter et al. (2022). In this paper, only the results of scenario 1 are compared and contrasted since it is the only scenario in which the choice model converged in a reasonable amount of time (i.e., less than six hours).

We find that SHAP reaches the same interpretations of the choice model while adding new insights in terms of heterogeneity of effects across respondents. Compared with LCCA, SHAP identifies a more detailed level of heterogeneity as the effects are computed per respondent instead of effects per pre-defined groups. For instance, we find in all models that people of the oldest age are associated with higher policy support. In SHAP, we also identify clusters of respondents with similar effects, sparse distributions of effects for respondents of a similar age and nonlinear effects that the other models do not identify. The results for the other covariates follow the same pattern: SHAP provides equivalent results to choice models and LCCA, with the addition of heterogeneity at the respondent level.

Regarding statistical significance and importance of covariates, we find that the covariates identified as the most important in SHAP coincide with the covariates identified as statistically significant in the choice model per specific COVID-19 measures. On the one hand, age group, the weight of citizens'/scientists' opinion and the perceived risk of getting very sick of COVID-19 are identified as the most important covariates on average by SHAP (see Table 4.6), and for each specific measure, these covariates rank on the higher part of the most important covariates per specific COVID-19 measures and at the same time they are statistically significant in the choice model (see Appendix 4.A and 4.B). On the other hand, the overcrowding risk reduction is ranked as the least-important covariate on average, and it ranks in the lowest positions per COVID-19 measure, coinciding with the fact that this covariate is not statistically significant in the choice model. Neither the weight of citizens'/scientists' opinion, the perceived risk of getting sick of COVID-19 nor the overcrowding risk reduction is considered in the LCCA analysis of Mouter et al. (2022). Based on the analyses made in this paper, we compare and contrast SHAP with choice models and LCCA in four dimensions (see Table 4.8).

| Covariate | Choice model | LCCA (Mouter et al., 2022) | SHAP (this work) |
|---|---|---|---|
| Age group | Interpretation: older age group is associated with higher policy support (see Appendix 4.B). | Interpretation: People of 65 years or more are over-represented in the cluster that recommends all measures. Not conclusive for other age groups. Thus, heterogeneity across predefined groups is identified. | Interpretation: Same as in choice model, with the addition of heterogeneity of effects across respondents (see figure 2) in the form of clusters and sparse distributions. Nonlinear effects observed (see figure 3). |
| | Significance/importance: All estimates are statistically significant, except for those associated with an advice to wash hands, not shaking hands and to stay home in case of symptoms. | Significance/importance: Age is statistically significant. No information for specific measures. | Significance/importance: Age is the most important covariate, on average. Age is not among the most important covariates for advising to wash hands, not shaking hands and to stay home in case of symptoms. |
| Weight citizens'/scientists' opinion | Interpretation: More weight to citizens' opinion is associated with lower policy support (see Appendix 4.B). | Not included in the analysis. | Interpretation: Same as in the choice model, with the addition of heterogeneity of effects as clusters, sparse distributions, or a combination (see figure 5). |
| | Significance/importance: All estimates are statistically significant. | | Significance/importance: The second-most important covariate, on average. It appears among the most important covariates across measures (see Appendix 4.A). |
| Risk (getting sick) | Interpretation: A higher perceived risk of getting sick of COVID-19 is associated with a higher policy support (see Appendix 4.B). | Not included in the analysis. | Interpretation: Same as in the choice model, with the addition of heterogeneity of effects in the form of sparse distributions (see figure 6). |
| | Significance/importance: All estimates are statistically significant, except for imposing a COVID-19 certificate (3G) in the catering industry and advising working from home. | | Significance/importance: The third-most important, on average. This covariate is not among the most important for imposing a COVID-19 certificate (3G) in the catering industry and advising working from home. |
| Overcrowding risk reduction | Not statistically significant in all measures, except for an advice of washing their hands. | Not included in the analysis. | Consistently among the least important covariates per measure, except in an advice of washing their hands. |

*Table 4.7: Contrast of interpretations between models and SHAP, scenario 1*

| Dimension | Choice model | LCCA | SHAP |
|---|---|---|---|
| Interpretation of covariates | Sign and magnitude of the estimated parameters indicate positive (negative) effect of the covariate on policy support. | Probabilities per latent class characterise each predefined group in terms of sociodemographic characteristics. | Sign and magnitude of SHAP values indicate the positive (negative) effect of the covariate value on the policy support per respondent. |
| Importance of covariates | Statistical significance of parameters. | Statistical significance of parameters. | Magnitude of SHAP importances, compared with the other covariates. |
| Heterogeneity of effects | Yes (observed and unobserved). Limited by the model specification. | Yes (observed). Limited by the number of latent classes. | Yes (observed). SHAP values are at the respondent level. |
| Estimation time | From one hour to more than six hours. | Two to three minutes | Two to three minutes |

*Table 4.8: Contrast of models and SHAP*

In terms of interpretation of results, we find that SHAP allows identifying the effect of covariates in the policy support in a similar way as in a choice model, with the addition of providing information at the respondent level. A similar analysis can be done with LCCA, in which the interpretation of results is made per predefined groups in terms of the probability of belonging to each of such groups. Regarding identifying the importance of covariates, both choice models and LCCA rely on identifying the statistical significance of a set of estimated parameters. In contrast, SHAP identifies the importance order of each covariate in terms of the SHAP importances.

In terms of heterogeneity, all models can capture observed (differences on effects of covariates) heterogeneity, whereas a choice model can also capture unobserved (stochastic) heterogeneity. On the one hand, SHAP is able to identify observed heterogeneity at the respondent level, thus identifying how the effects of each covariate are distributed across covariates and measures. On the other hand, choice models and LCCA can capture observed heterogeneity, but such ability is limited by the a priori model specification provided in the former, and the a priori definition of the number of latent classes in the latter. However, evaluating all possible model specifications in a choice model is time-unfeasible, whereas specifying a too high number of latent classes in LCCA can lead to a non-informative model (i.e., nonparsimonious, with few or no statistically significant parameters).

A final and practical difference between all models is the estimation time, which is critical in crises when results are needed in shorter time spans for decision-making. On the one hand, choice models are the least convenient approach, with an estimation time of around one hour for scenario 1. Furthermore, after six hours, we could not obtain convergence of the choice model for scenarios 2, 3 and 4. On the other hand, LCCA and SHAP estimation times are around three minutes for all scenarios. Considering that we show SHAP provides similar results as a choice model in the same scenario, with the addition of identifying heterogeneity of effects per covariate and measure, SHAP can be used instead of the choice model for this application.

## 4.5. Discussion

In this paper, we study the factors (covariates), i.e., sociodemographic characteristics, perception indicators and experimental variables, that lead to differences in the policy support for COVID-19 measures under different risk scenarios, with a focus on how such differences are distributed across citizens. We use data from a PVE experiment to determine the citizens' preferences for COVID-19 measures in the Netherlands (Mouter et al., 2022). We model the data with XGBoost, a ML model, and compute the SHAP values to identify the effect of each used covariate on the policy support for COVID-19 measures for each respondent of the PVE experiment. Our results show that the heterogeneity of effects on the policy support for measures can lead to considerable differences between respondents of similar profiles (e.g., age, perception) or nonlinear effects that, if neglected by only considering average effects, could lead to misinterpretation of results. Furthermore, we show that SHAP analysis provides similar results as conventional approaches (i.e., choice models), but with the addition of providing effects at the respondent level and in a considerably minor estimation time.

### 4.5.1. Main findings

First, we show how the policy support for COVID-19 measures is distributed across respondents in terms of the age group of respondents, the weight they believe the government should give to the opinion of citizens compared to the opinion of scientists, and the perceived risk of becoming sick of COVID-19, which are the covariates identified as with the highest importance by SHAP importances (see Table 4.6). Aside from confirming the findings of previous studies, including the first analysis of the PVE experiment (Mouter et al., 2022), we identify clusters of different types of respondents but with similar policy support, sparse distributions of effects for respondents with similar characteristics, effects in the opposite direction for specific measures and nonlinear effects for specific groups of respondents. For instance, we find that for closing schools in a high-risk scenario (scenario 4), respondents of the lowest age group are associated with higher policy support for the measure than respondents of other age groups, going in an opposite direction to the "average" interpretation for the rest of measures (see Figure 4.2). As another example, we find that the policy support for implementing a COVID-19 certificate in scenario 1 across different age groups is a piecewise-linear function, with a negative effect for groups less than 45 years old and a positive effect for older groups (see Figure 4.3). Similar findings are made for the weight of citizens'/scientists' opinion and perceived risk of getting sick of COVID-19, where combinations of clusters and sparse distributions of effects are found for specific measures and scenarios (see figures 5 and 6).

Second, we show that SHAP analysis delivers the same interpretation results and identification of important covariates as a conventional choice model, with the addition of providing how the effects are distributed at the respondent level, whereas contrasted with an LCCA, SHAP provides a deeper level of heterogeneity as there is no need of pre-defining a number of latent classes. The visualisation of SHAP values allows determining that older age, a higher weight to the opinion of scientists and a higher perceived risk of getting sick of COVID-19 are associated with higher policy support for COVID-19 measures, with a similar conclusion obtained from interpreting the estimated parameters of the choice model (see Table 4.7). Furthermore, SHAP values also provide information about how the effects are distributed across respondents, allowing for a more nuanced analysis per covariate, measure and risk scenario. Finally, we argue in favour of using SHAP for interpreting results and identifying importance, as this method provides the same results as a choice model in a considerably shorter time: two to three minutes contrasted with one to more than six hours (see Table 4.8).

### 4.5.2. Policy implications

SHAP analysis can help policymakers understand which types of citizens are the most (least) reluctant to specific measures in greater detail than previous methods (i.e., choice models and LCCA) and tailor measures to increase policy support. For instance, as we found that negative support for a COVID-19 certificate in a low risk scenario (scenario 1) is concentrated in citizens 45 years old or less (see Figure 4.3), policymakers can build information campaigns focused on such age groups to increase support for this measure. As another example, since we found that respondents of the middle and high age groups are associated with lower policy support for closing schools in a high-risk scenario (scenario 4, Figure 4.2), policymakers can focus on such age groups to prepare compensation packages, since at the same time these groups are more likely to have children in school age than citizens of the lowest age group (i.e., below 25 years old).

### 4.5.3. Considerations and research directions

We identify a number of considerations in this paper. First, our findings are bounded by the population context, the moment the sample was collected and the use of PVE as an elicitation framework. Therefore, the findings of this paper should not be extrapolated for other countries or other moments of the pandemic, even though our findings align with previous studies regarding policy support for COVID-19 measures (Sicsic et al., 2022). Second, our methodological approach (i.e., XGBoost and SHAP) only establishes associations between covariates and the policy support for COVID-19 measures, e.g., older age is associated with higher policy support. Still, we do not establish causality, e.g., if age is higher, then policy support is higher (van Cranenburgh et al., 2022). Researchers and policymakers should keep this distinction clear at the moment of drawing conclusions from this work, as the issue of causality is an ongoing debate in the field of ML. Finally, SHAP has a longer computation time than alternative explanation methods (e.g., LIME, LRP), often in the order of minutes at the minimum. Hence, researchers and policymakers should carefully assess the advantages of SHAP (i.e., built in solid theory, global and local explanations) in light of its computational demands, particularly when the urgency of obtaining results is a priority.

Finally, as a further research direction, we envision using SHAP to explain the policy support for measures for specific profiles of respondents. In other words, policymakers could use SHAP to construct a citizen profile of interest (e.g., middle-aged, from the countryside, with a high perceived risk, etc.) and determine its policy support for specific measures and scenarios. This paper did not cover this direction since the range of possible profiles to explore is unfeasible to cover in a manuscript. To overcome this, developing a consultation (web-based) platform to build specific queries is possible. The interested analyst can construct specific profiles of citizens from a previously-trained ML model and obtain their specific set of SHAP values as a result. Policymakers could benefit from such a web-based platform by counting with information about the policy support for COVID-19 measures for different individuals, different measures and scenarios in a fine-grained level of detail.

## Acknowledgements

The project was funded by the National Institute of Public Health and the Environment (RIVM). The corresponding author would like to acknowledge the Netherlands Organisation for Scientific Research (NWO) for funding his PhD (NWO Responsible Innovation grant – 313-99-333) and the support of the TU Delft AI Labs Programme.

# 4.A. SHAP summary plots per risk scenario

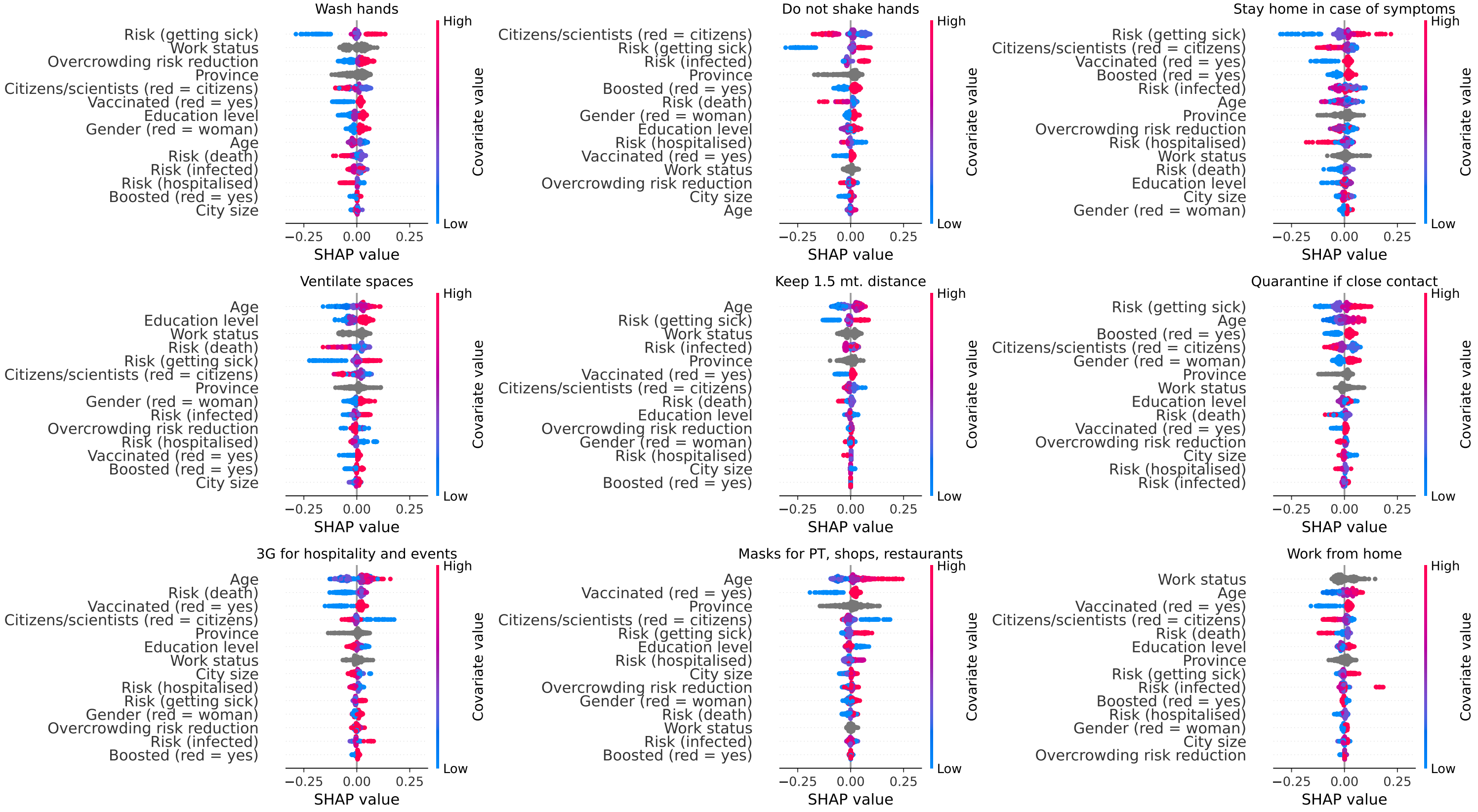


*Figure 4.7: SHAP summary plots per COVID-19 measure, scenario 1*

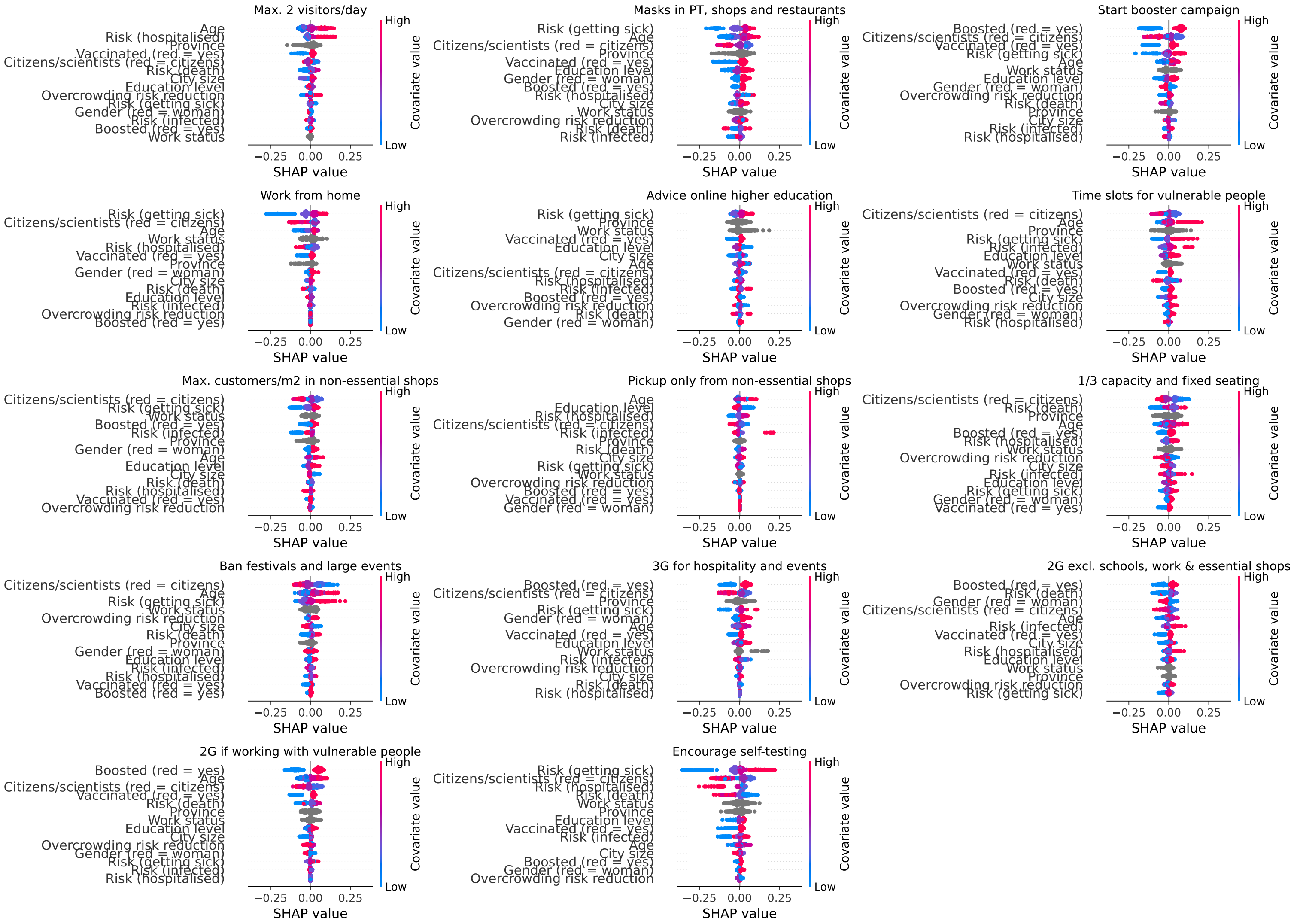


*Figure 4.8: SHAP summary plots per COVID-19 measure, scenario 2*

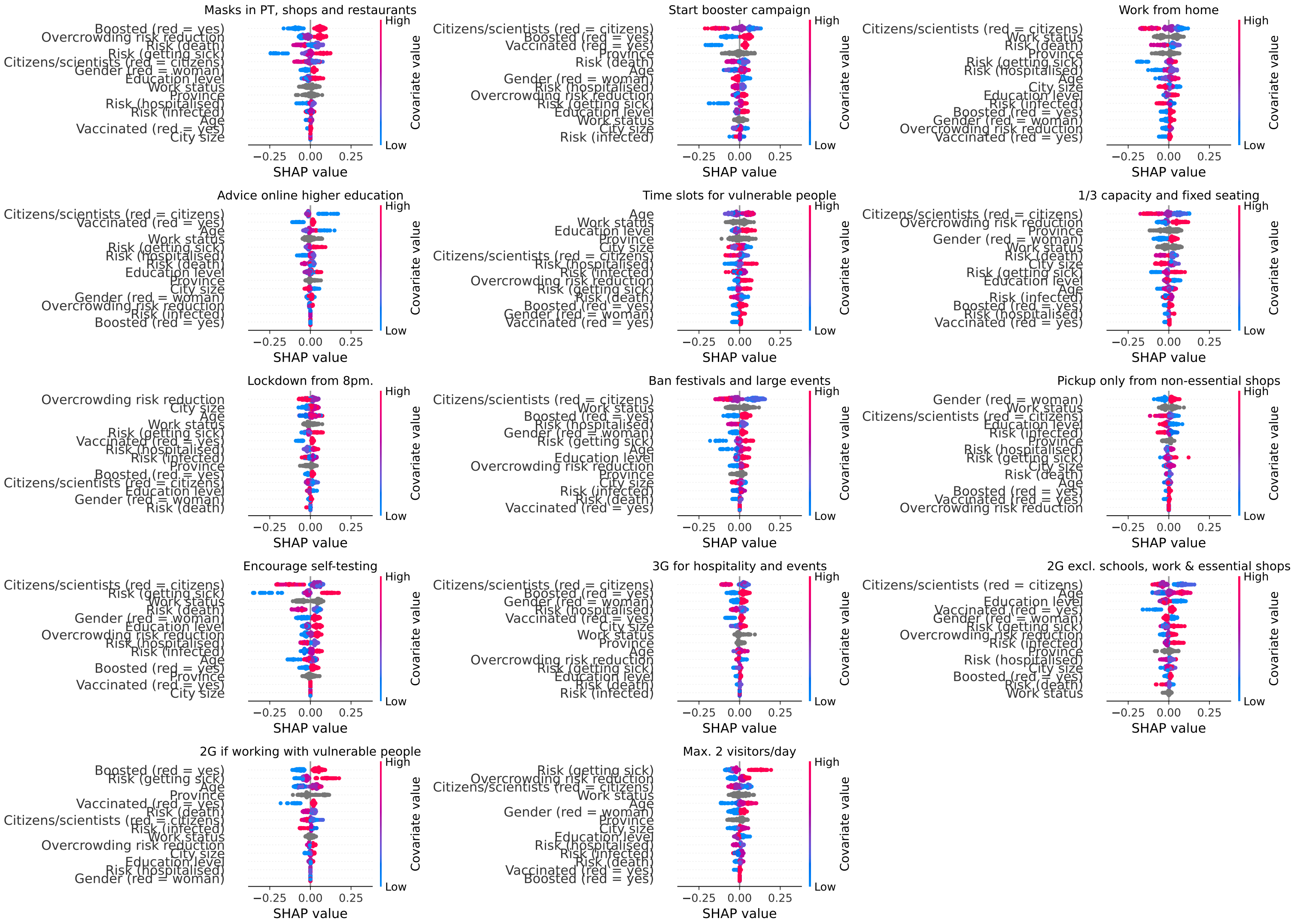


*Figure 4.9: SHAP summary plots per COVID-19 measure, scenario 3*

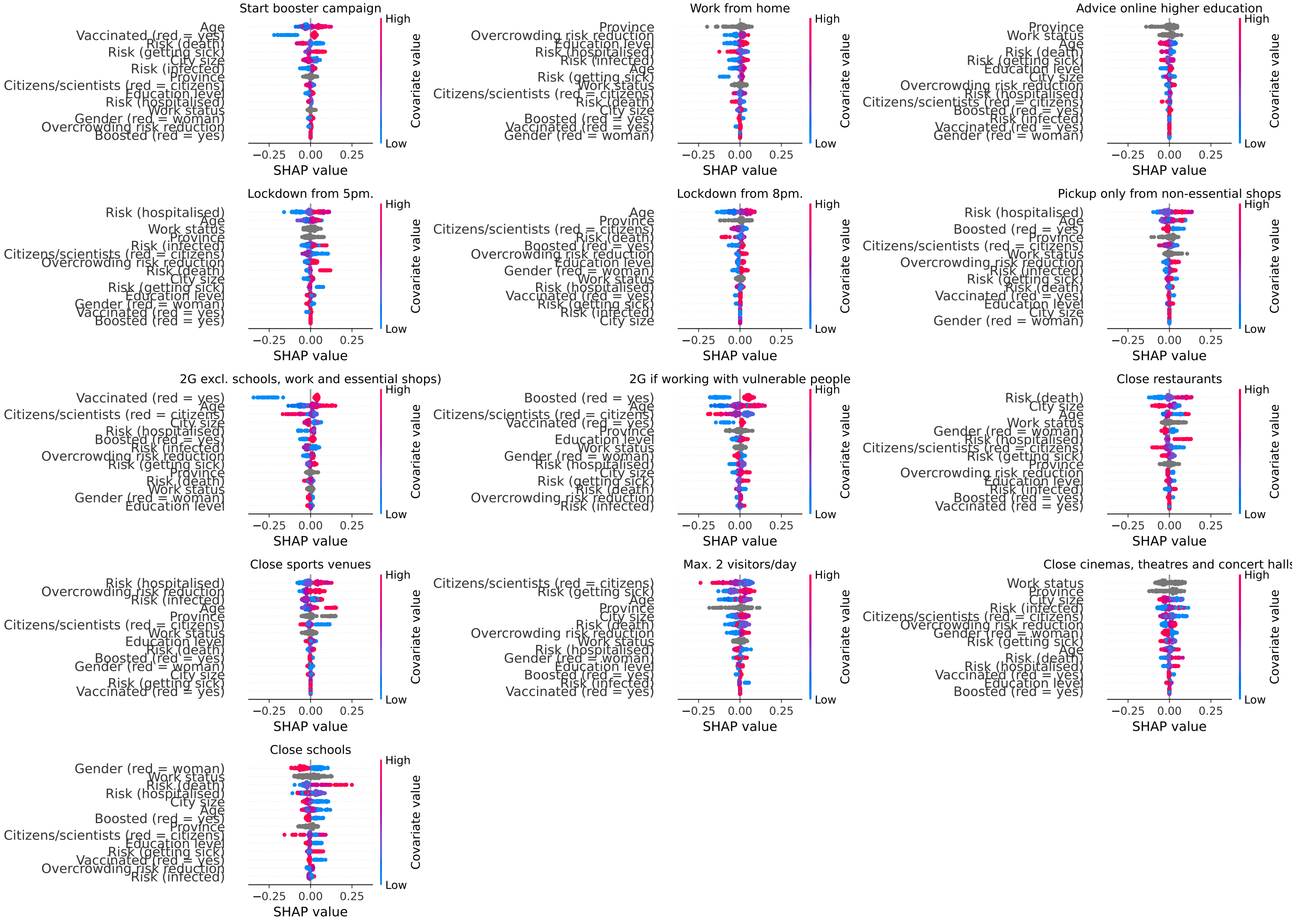


*Figure 4.10: SHAP summary plots per COVID-19 measure, scenario 4*

## 4.B. Description and estimation results of the choice model

The choice model employed to contrast the findings of SHAP is an extended version of the portfolio choice model employed by Mouter et al. (2022). The version employed in our work includes the same set of covariates employed in the SHAP analysis as explanatory variables per COVID-19 measure, in order to allow a contrast between both analysis methods. The model used by Mouter et al. (2022) only considered the sum of overcrowding risk reductions of the selected package of measures by the respondents.

The portfolio choice model was initially proposed by Bahamonde-Birke & Mouter (2019). This model is based on the Random Utility Model (RUM) employed to analyse discrete choice data, extended to consider that respondents can choose packages of alternatives instead of a single (discrete) alternative.

In the portfolio choice model, respondents seek to maximise their utility derived from their chosen combination of alternatives, and hence, higher utility is associated with higher policy support. In turn, the utility of each alternative is a function of their experimental features (i.e., attributes) and individual-specific covariates (e.g., sociodemographic characteristics, perception indicators). Specifically, the utility of respondent $n$ for a combination of alternatives $p$ is given by equation (A1):

$$U_{np} = V_{np} + \varepsilon_{np} = \sum_{j=1}^{J} y_{nj} \left( \delta_j + \beta' X_{nj} + \theta_j' Z_n \right) + \varepsilon_{np}, \tag{4.4}$$

Where $y_{nj}$ is a binary variable equal to one if the respondent choose alternative $j$, $X_{nj}$ is a vector of characteristics of the alternative $j$, (e.g., overcrowding risk reductions in the PVE experiment), $Z_n$ is a vector of individual-specific covariates of respondent $n$, $\delta_j$, $\beta$ and $\theta_j$ are parameters to be estimated, and $\varepsilon_{np}$ is a stochastic error with a Gumbel distribution. Under these assumptions, the probability of choosing a combination of alternatives $p$ take the form of a multinomial logit (MNL) model, as described by equation (A2):

$$p \left( U_{np} \geq U_{nq}, \forall q \neq p \right) = \frac{\exp(V_n p)}{\sum_q \exp(V_n q)}, \tag{4.5}$$

The estimated parameters $\delta_j$, $\beta$ and $\theta_j$ have an economic interpretation. Firstly, $\delta_j$ are alternative-specific constants interpreted as the utility increase when their associated alternative is chosen. Secondly, the sign of $\beta$ is interpreted as the contribution of an attribute increase to the respondents utility. If $\beta$ is positive, then increases on its associated attribute generate increases on the respondent's utility, and if $\beta$ is negative, then the increase of its associated attribute generates a decrease of the respondent's utility. Lastly, the sign of $\theta_j$ is interpreted as the effect of the individual-specific covariates on the respondent's utility. If $\theta_j$ is positive, the associated covariate $Z$ induces increases in the utility, while if If $\theta_j$ is negative, the covariate is associated with decreases in the utility.

| | Advice to wash hands | Advice to not shake hands | Advice to stay home in case of symptoms | Advice to ventilate spaces | Advice to keep 1.5 mt. distance | Advice of quarantine if close contact | COVID-19 certificate (3G) for hospitality industry | Wear masks in PT, shops & restaurants | Advice to work from home |
|---|---|---|---|---|---|---|---|---|---|
| Measure-specific constant | -0.780** | -1.062*** | -0.676* | -0.372 | -0.937** | -1.078*** | -2.161*** | -1.701*** | -0.852** |
| | (0.246) | (0.320) | (0.304) | (0.257) | (0.288) | (0.278) | (0.373) | (0.273) | (0.276) |
| Overload risk reduction | 0.140* | 0.064 | -0.004 | -0.019 | 0.015 | -0.020 | 0.029 | 0.011 | 0.024 |
| | (0.061) | (0.059) | (0.020) | (0.029) | (0.019) | (0.029) | (0.066) | (0.031) | (0.058) |
| Is a woman | 0.196 | 0.262* | 0.171 | 0.198 | 0.102 | 0.324** | 0.217 | 0.158 | 0.158 |
| | (0.108) | (0.105) | (0.106) | (0.104) | (0.103) | (0.104) | (0.118) | (0.110) | (0.103) |
| Middle age | 0.075 | 0.178 | -0.040 | 0.165 | 0.267* | 0.291* | 0.202 | -0.036 | 0.224 |
| | (0.126) | (0.122) | (0.123) | (0.121) | (0.118) | (0.119) | (0.133) | (0.125) | (0.118) |
| Higher age | -0.041 | -0.171 | 0.171 | 0.040 | 0.122 | 0.295 | 0.366 | 0.646** | 0.196 |
| | (0.239) | (0.229) | (0.236) | (0.230) | (0.224) | (0.226) | (0.240) | (0.231) | (0.222) |
| Middle education | 0.191 | 0.073 | 0.190 | 0.087 | -0.107 | -0.034 | -0.124 | -0.288* | 0.004 |
| | (0.131) | (0.128) | (0.129) | (0.127) | (0.125) | (0.127) | (0.141) | (0.132) | (0.126) |
| Higher education | 0.333* | 0.216 | 0.163 | 0.353** | -0.185 | -0.020 | -0.267 | -0.407** | 0.125 |
| | (0.139) | (0.134) | (0.135) | (0.134) | (0.131) | (0.132) | (0.149) | (0.138) | (0.131) |
| Friesland | 0.197 | -0.002 | 0.127 | 0.139 | 0.139 | 0.204 | -0.223 | -0.003 | -0.192 |
| | (0.184) | (0.177) | (0.182) | (0.180) | (0.175) | (0.176) | (0.203) | (0.191) | (0.176) |
| Gelderland | 0.165 | 0.088 | -0.324 | -0.119 | 0.096 | -0.037 | -0.155 | 0.211 | 0.259 |
| | (0.228) | (0.221) | (0.221) | (0.220) | (0.219) | (0.223) | (0.253) | (0.233) | (0.218) |
| Groningen | -0.158 | 0.083 | -0.354 | -0.447 | 0.412 | 0.239 | 0.166 | -0.077 | -0.570 |
| | (0.331) | (0.331) | (0.327) | (0.325) | (0.327) | (0.328) | (0.353) | (0.359) | (0.340) |
| Limburg | 0.234 | 0.255 | 0.256 | 0.116 | 0.402 | 0.474* | 0.007 | 0.130 | 0.137 |
| | (0.221) | (0.214) | (0.219) | (0.213) | (0.209) | (0.211) | (0.234) | (0.223) | (0.208) |
| North Brabant | 0.293 | 0.151 | 0.034 | 0.056 | 0.294 | 0.189 | -0.126 | 0.299 | 0.117 |
| | (0.162) | (0.156) | (0.158) | (0.157) | (0.154) | (0.156) | (0.176) | (0.165) | (0.153) |
| North Holland | 0.360 | 0.368 | 0.050 | 0.066 | 0.346 | 0.275 | 0.044 | 0.614* | 0.080 |
| | (0.262) | (0.253) | (0.251) | (0.248) | (0.243) | (0.245) | (0.275) | (0.252) | (0.243) |
| Utrecht | 0.482 | 0.118 | 0.297 | -0.063 | 0.466 | 0.486 | 0.396 | 0.164 | 0.365 |
| | (0.299) | (0.276) | (0.286) | (0.275) | (0.274) | (0.275) | (0.293) | (0.293) | (0.273) |
| Overijssel | 0.386* | 0.321 | 0.178 | -0.046 | 0.247 | 0.357* | -0.249 | 0.031 | -0.110 |
| | (0.181) | (0.174) | (0.175) | (0.172) | (0.169) | (0.170) | (0.195) | (0.182) | (0.169) |
| Zeeland | -0.056 | -0.415* | 0.214 | -0.034 | 0.221 | 0.209 | 0.112 | 0.102 | -0.160 |
| | (0.211) | (0.208) | (0.214) | (0.208) | (0.206) | (0.208) | (0.230) | (0.223) | (0.207) |
| South Holland | -0.136 | 0.107 | -0.310 | 0.336 | -0.034 | 0.149 | -0.282 | -0.491 | 0.352 |
| | (0.277) | (0.278) | (0.276) | (0.284) | (0.276) | (0.277) | (0.332) | (0.325) | (0.274) |
| Observations | 1,888 | | | | | | | | |
| Log-likelihood | -10,803.28 | | | | | | | | |

*Table 4.9: Estimation results of the choice model, scenario 1 (continue in the next page)*

| | Advice to wash hands | Advice to not shake hands | Advice to stay home in case of symptoms | Advice to ventilate spaces | Advice to keep 1.5 mt. distance | Advice of quarantine if close contact | COVID-19 certificate (3G) for hospitality industry | Wear masks in PT, shops & restaurants | Advice to work from home |
|---|---|---|---|---|---|---|---|---|---|
| Medium city | 0.003 | -0.125 | 0.079 | **-0.259*** | -0.239 | -0.090 | 0.089 | -0.146 | -0.159 |
| | (0.138) | (0.133) | (0.136) | **(0.132)** | (0.131) | (0.132) | (0.150) | (0.142) | (0.131) |
| Big city | -0.073 | -0.090 | -0.160 | **-0.244*** | -0.070 | -0.086 | 0.117 | -0.013 | -0.146 |
| | (0.114) | (0.110) | (0.111) | **(0.109)** | (0.107) | (0.109) | (0.123) | (0.115) | (0.108) |
| Incapacitated | 0.217 | 0.110 | 0.231 | 0.403** | 0.099 | 0.079 | -0.037 | -0.108 | -0.114 |
| | (0.147) | (0.142) | (0.144) | (0.142) | (0.139) | (0.140) | (0.163) | (0.153) | (0.139) |
| Retired | 0.295 | 0.090 | 0.249 | 0.381 | -0.005 | -0.057 | -0.124 | -0.207 | -0.241 |
| | (0.232) | (0.222) | (0.226) | (0.222) | (0.217) | (0.219) | (0.253) | (0.239) | (0.220) |
| Housewife -husband | 0.396 | 0.303 | 0.136 | 0.456 | 0.141 | 0.064 | -0.224 | -0.209 | 0.157 |
| | (0.254) | (0.243) | (0.250) | (0.244) | (0.237) | (0.239) | (0.257) | (0.245) | (0.236) |
| Not working | 0.209 | 0.171 | -0.133 | 0.359 | 0.181 | 0.023 | -0.025 | 0.171 | 0.082 |
| | (0.191) | (0.185) | (0.184) | (0.183) | (0.180) | (0.182) | (0.206) | (0.190) | (0.180) |
| Student | 0.434 | -0.146 | **0.602*** | -0.052 | 0.245 | 0.380 | 0.219 | 0.163 | -0.049 |
| | (0.279) | (0.253) | **(0.272)** | (0.250) | (0.250) | (0.251) | (0.290) | (0.271) | (0.252) |
| Vaccinated | **0.446**** | 0.407* | **0.483**** | 0.317 | **0.482**** | 0.210 | **0.658**** | **0.762***** | **0.347*** |
| | **(0.169)** | (0.167) | **(0.166)** | (0.166) | **(0.169)** | (0.171) | **(0.226)** | **(0.203)** | **(0.170)** |
| Boosted | 0.110 | 0.222 | 0.318* | 0.036 | -0.003 | 0.195 | 0.248 | 0.118 | 0.122 |
| | (0.137) | (0.131) | (0.132) | (0.132) | (0.130) | (0.131) | (0.153) | (0.140) | (0.130) |
| High risk (infected) | 0.032 | 0.098 | -0.111 | 0.042 | **-0.291*** | -0.102 | -0.026 | -0.058 | -0.061 |
| | (0.122) | (0.116) | (0.119) | (0.117) | **(0.115)** | (0.116) | (0.131) | (0.122) | (0.114) |
| High risk (getting sick) | 0.078 | **0.282*** | **0.407**** | 0.148 | **0.285*** | **0.325*** | 0.047 | 0.027 | 0.197 |
| | (0.138) | **(0.132)** | **(0.135)** | (0.132) | **(0.129)** | **(0.130)** | (0.148) | (0.139) | (0.129) |
| High risk (hospitalised) | -0.029 | -0.106 | -0.064 | -0.110 | -0.114 | 0.052 | -0.239 | **0.356*** | -0.126 |
| | (0.179) | (0.172) | (0.175) | (0.171) | (0.166) | (0.168) | (0.192) | **(0.173)** | (0.166) |
| High risk (death) | -0.253 | **-0.382*** | -0.182 | **-0.424*** | -0.021 | -0.164 | 0.270 | -0.118 | -0.175 |
| | (0.180) | **(0.173)** | (0.176) | **(0.172)** | (0.168) | (0.170) | (0.193) | (0.175) | (0.169) |
| Higher weight to scientists opinion | **0.255*** | **0.367***** | **0.295**** | **0.253*** | **0.233*** | **0.382***** | **0.290**** | **0.249*** | 0.106 |
| | **(0.107)** | **(0.102)** | **(0.104)** | **(0.102)** | **(0.099)** | **(0.100)** | **(0.112)** | **(0.105)** | (0.099) |
| Observations | 1,888 | | | | | | | | |
| Log-likelihood | -10,803.28 | | | | | | | | |

*Table 4.10: Continuation of 4.9*

# Bibliography

Bach, S., A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek (2015) On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, *PloS one*, 10(7), p. e0130140.

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

Chen, T., C. Guestrin (2016) XGBoost: A Scalable Tree Boosting System, in: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, Association for Computing Machinery, New York, NY, USA, pp. 785–794.

Dohle, S., T. Wingen, M. Schreiber (2020) Acceptance and Adoption of Protective Measures During the COVID-19 Pandemic: The Role of Trust in Politics and Trust in Science, *Social Psychological Bulletin*, 15(4), pp. 1–23.

Dong, G., Y. Kweon, B. B. Park, M. Boukhechba (2022) Utility-based route choice behavior modeling using deep sequential models, *Journal of big data analytics in transportation*, 4(2-3), pp. 119–133.

Friedman, J. H. (2001) Greedy function approximation: A gradient boosting machine, *Annals of Statistics*, 29(5), pp. 1189–1232.

Gotanda, H., A. Miyawaki, T. Tabuchi, Y. Tsugawa (2021) Association between trust in government and practice of preventive measures during the COVID-19 pandemic in Japan, *Journal of general internal medicine*, 36(11), pp. 3471–3477.

Ji, S., X. Wang, T. Lyu, X. Liu, Y. Wang, E. Heinen, Z. Sun (2022) Understanding cycling distance according to the prediction of the xgboost and the interpretation of shap: a non-linear and interaction effect analysis, *Journal of Transport Geography*, 103, p. 103414.

Jin, L., A. Lazar, C. Brown, B. Sun, V. Garikapati, S. Ravulaparthy, Q. Chen, A. Sim, K. Wu, T. Ho (2022) What Makes You Hold on to That Old Car? Joint Insights from Machine Learning and Multinomial Logit on Vehicle-level Transaction Decisions, *arXiv preprint arXiv:2205.06622*.

Lee, E. H. (2022) Exploring transit use during covid-19 based on xgb and shap using smart card data, *Journal of Advanced Transportation*, 2022.

Loria-Rebolledo, L. E., M. Ryan, V. Watson, M. G. Genie, R. A. Sakowsky, D. Powell, S. Paranjothy (2022) Public acceptability of non-pharmaceutical interventions to control a pandemic in the UK: A discrete choice experiment, *BMJ open*, 12(3), p. e054155.

Lundberg, S. M., P. G. Allen, S.-I. Lee (2017) A Unified Approach to Interpreting Model Predictions, *Advances in Neural Information Processing Systems*, 30.

Mouter, N., J. I. Hernandez, A. V. Itten (2021a) Public participation in crisis policy-making. How 30,000 Dutch citizens advised their government on relaxing COVID-19 lockdown measures, *PLOS ONE*, 16(5), p. e0250614.

Mouter, N., K. T. Jara, J. I. Hernandez, M. Kroesen, M. de Vries, T. Geijsen, F. Kroese, E. Uiters, M. de Bruin (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, p. 115430.

Mouter, N., P. Koster, T. Dekker (2021b) Participatory value evaluation for the evaluation of flood protection schemes, *Water Resources and Economics*, 36, p. 100188.

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

Ribeiro, M. T., S. Singh, C. Guestrin (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.

Rotteveel, A. H., M. S. Lambooij, E. a. B. Over, J. I. Hernández, A. W. M. Suijkerbuijk, A. T. de Blaeij, G. A. de Wit, N. Mouter (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), pp. 428–443.

Sicsic, J., S. Blondel, S. Chyderiotis, F. Langot, J. E. Mueller (2022) Preferences for COVID-19 epidemic control measures among French adults: A discrete choice experiment, *The European Journal of Health Economics*, pp. 1–18.

van Cranenburgh, S., S. Wang, A. Vij, F. Pereira, J. Walker (2022) Choice modelling in the age of machine learning - Discussion paper, *Journal of Choice Modelling*, 42, p. 100340.

Wang, S., B. Mo, S. Hess, J. Zhao (2021) Comparing hundreds of machine learning classifiers and discrete choice models in predicting travel behavior: An empirical benchmark, *arXiv preprint arXiv:2102.01130*.

Wang, Y., Y. Zhao, J. Song, H. Liu (2022) What drives patients to choose a physician online? a study based on tree models and shap values, in: *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*, IEEE, pp. 1676–1683.

# Chapter 5

# An economically-consistent discrete choice model based on artificial neural networks

- Hernandez, J.I., Mouter, N. & van Cranenburgh, S. An economically-consistent discrete choice model with flexible utility specification based on artificial neural networks. *Under review.*

Random utility maximisation (RUM) models are one of the cornerstones of discrete choice modelling. However, specifying the utility function of RUM models is not straightforward and has a considerable impact on the resulting interpretable outcomes and welfare measures. In this paper, we propose a new discrete choice model based on artificial neural networks (ANNs) named “Alternative-Specific and Shared weights Neural Network (ASS-NN), which provides a further balance between flexible utility approximation from the data and consistency with two assumptions: RUM theory and fungibility of money (i.e., “one euro is one euro”). Therefore, the ASS-NN can be used to derive economically-consistent outcomes, such as marginal utilities or willingness to pay, without explicitly specifying the utility functional form. Using a Monte Carlo experiment and empirical data from the Swissmetro dataset, we show that ASS-NN outperforms (in terms of goodness of fit) conventional multinomial logit (MNL) models under different utility specifications. Furthermore, we show how the ASS-NN is used to derive marginal utilities and willingness to pay measures.

## 5.1. Introduction

Random utility maximisation (RUM) models (McFadden, 1974) are one of the cornerstones of discrete choice modelling. RUM models provide a framework to analyse travel demand (Ben-Akiva et al., 1985; Ben-Akiva & Bierlaire, 2003). The strength of RUM models relies on their interpretability and connection with economic theory (Small & Rosen, 1981). Notably, the estimates of RUM models can inform the analyst about individual preferences for attribute changes, substitution rates and willingness to pay. This property of RUM models makes them a particularly insightful approach for transport appraisal.

However, a key challenge of RUM models is the specification of the utility function. Conventionally, specifying the utility of a RUM model concerns a trial-and-error process, where the analyst estimates several competing models (with different functional specifications), based on prior knowledge (e.g., findings from previous studies, economic theory). The analyst selects the final specification based on, on the one hand, behavioural intuition (i.e., the size and magnitude of the estimated parameters make sense) and, on the other hand, goodness-of-fit or information criteria. Nevertheless, the true utility functional form is not known but assumed a priori by the analyst. Furthermore, the selected utility specification has a considerable impact on the derived interpretable measures, such as the estimated willingness to pay (Torres et al., 2011; van der Pol et al., 2014). In consequence, the selected utility specification is not a trivial choice, considering the relevance of such measures for policymaking and appraisal.

An alternative that can help to circumvent this challenge of RUM models is using machine learning (ML) models. ML models are methods aimed to learn patterns and/or approximate mathematical functions directly from the data. In the last years, ML models have been increasingly adopted in the discrete choice modelling field (van Cranenburgh et al., 2022). ML models differ from discrete choice models (DCMs) in their modelling paradigms. DCMs are theory-driven, in the sense that the analyst assumes the underlying data-generating process (DGP), and the goal is to find the model parameters that better describe such model, given data. Unlike DCMs, ML models adopt a data-driven paradigm under the guiding principle that the true DGP is unknown and complex, but it can be uncovered from the data. This difference in paradigms allows ML models to reach higher performance than their theory-driven counterparts for predictive tasks (Wang et al., 2021a).

Among specific ML models, Artificial Neural Networks (ANNs) have gained considerable ground in the discrete choice modelling field (e.g., Alwosheel et al., 2018; Sifringer et al., 2020; van Cranenburgh & Alwosheel, 2019; Wang et al., 2020a, 2021b).

ANNs are ML models loosely based on the structure of brains, aimed to approximate mathematical functions from data. In an ANN, the underlying DGP is modelled as a set of layers and nodes interconnected to each other. This allows ANNs to model complex interactions between input variables (covariates) without the need to be specified a priori by the analyst. Notably, ANNs can be structured to build discrete choice models with a flexible utility function (Bentz & Merunka, 2000), which allows accounting for interactions and nonlinear effects that the analyst could overlook and, therefore, overcoming the limitations of manually specifying a specific utility function.

Despite their strengths, ANNs provide limited information of behavioural and economic interest without further intervention. This is because the parameters of ANNs lack interpretation, as a difference from conventional discrete choice models. To overcome this limitation, several works have proposed alternative structures that restrict part of the ANN to increase its interpretability (Han et al., 2022; Sifringer et al., 2020; Wong & Farooq, 2021). This strategy, however, involves a trade-off between having a flexible utility function, i.e., that can capture interactions and non-linearities without being specified a priori, and consistency with RUM and economic theory to derive measures that can be used for welfare analysis. On the one hand, an ANN with the highest flexibility to approximate utility functions (e.g., an ANN with no intervention) provides outcomes that may violate consistency with RUM theory and, therefore, the connection between their derived welfare measures and economic theory cannot be guaranteed. On the other hand, an ANN with a high level of intervention would provide interpretable outcomes that satisfy RUM and economic assumptions but at the expense of having a utility specification that is not flexible enough to identify interactions or nonlinear effects from the data. To balance these trade-offs, the analyst must provide a structure that provides enough flexibility to the ANN to approximate utility functions and, at the same time, satisfies consistency with RUM and economic assumptions that guarantee to derive meaningful interpretable outcomes and welfare measures.

In this paper, we propose the “Alternative-Specific and Shared weights Neural Network” (ASS-NN), a discrete choice model based on ANNs that incorporates domain knowledge to guarantee consistency with RUM and economic theory. The ASS-NN is built upon the “Alternative-Specific Utility Deep Neural Network” (ASU-DNN), proposed by Wang et al. (2020b). Both models feature alternative-specific utility functions that are approximated from the data. Our proposed model, in addition, postulates “fungibility of money”, also known as “one euro is one euro”. Fungibility of money refers to the notion that money can be spent in different goods (alternatives) interchangeably. As a result of this assumption, the marginal utility of costs for the same individual must be equal across alternatives of the same cost level. In addition, we discuss that the alternative-specific utility structure of both the ASU-DNN and the ASS-NN are consistent with RUM (Hess et al., 2018) and, in consequence, the outcomes and welfare measures obtained from the ASS-NN can be connected with economic theory. To incorporate fungibility of money, the cost-dependent utility of the ASS-NN is modelled with separate sets of hidden layers with shared weights, which forces equal marginal utility of costs across alternatives for a same individual and same cost level. The trained ASS-NN can be used to estimate the marginal utility of attribute increases, marginal rates of substitution, and willingness to pay for attribute changes, e.g., the value of travel time (VTT).

We show the use of the ASS-NN using a Monte Carlo experiment and empirical data from the Swissmetro dataset (Bierlaire et al., 2001). The Swissmetro dataset is a mode choice experiment that is widely known by the transportation research community, and it has been previously used as benchmark data in other ANN-based discrete choice models proposed in the literature (e.g., Sifringer et al., 2020). We first conduct a Monte Carlo experiment with pseudo-synthetic choices generated from the Swissmetro dataset to show that the ASS-NN succeeds in approximating the true utility function from the data under different utility specifications, as well as in recovering the marginal utility of attribute increases and willingness to pay. Then, we train the ASS-NN with empirical data to approximate the marginal utilities and willingness to pay, and we compare these outcomes with those from the ASU-DNN and conventional multinomial logit (MNL) models under different utility specifications. To allow researchers to replicate our work and encourage open science, the code and data used in this paper are published in a Git repository: https://github.com/ighdez/ass_nn_paper.

The remainder of this paper is as follows. Section 5.2 describes the methodology and how the ASS-NN is implemented. Section 5.3 presents the setting and results of the Monte Carlo experiment. Section 5.4 presents the empirical data and results. Finally, section 5.5 provides a discussion and conclusion.

## 5.2. Methodology

### 5.2.1. Theoretical models

The RUM model is a theoretical framework to describe individual choice behaviour based on the notion that decision-makers seek to maximise their utility from a set of discrete goods. In this section, proceed to formalise a general RUM model that can be approximated with the ASS-NN.

Let $n$ be a decision-maker that perceives utility from the consumption of $J$ mutually-exclusive goods. Each alternative $j$ is characterised by observable attributes (Lancaster, 1966). For a given alternative $j$ faced by decision-maker $n$, let $X_{nj} = \{x_{n11}, \ldots, x_{n1K}\}$ be the vector of non-cost attributes of such alternative, and $c_{nj}$ be the cost value. Then, the utility of each good $U_{nj}$ perceived by decision-maker $n$ for alternative $j$ is a function of such observable characteristics plus a stochastic error term, as defined by equation (5.1):

$$U_{nj} = V_n(X_{n1}, X_{n2}, \ldots, X_{nJ}, c_{n1}, c_{n2}, \ldots, c_{nJ}, w) + \varepsilon_{nj}, \tag{5.1}$$

where $V_n$ is a function that depends on observed attributes and costs, $w$ is a vector of weights to be estimated and $\varepsilon_{nj}$ is a stochastic error term. The model described in equation (1) is a general discrete choice model without a specific utility functional form that can be approximated with a fully-connected ANN (Bentz & Merunka, 2000).

However, a limitation of this model is that is not consistent with RUM theory. As discussed by Hess et al. (2018), a RUM-consistent model must satisfy two conditions: regularity and transitivity. Regularity refers that adding a new alternative to the choice set should not increase the choice probability of the other alternatives. Transitivity states that if an alternative A is preferred over B, and B over C, then alternative A is preferred over C. The model of equation 5.1 does not exhibit regularity, since the attributes of one alternative can affect the utility of other alternative(s). Thus, if an alternative is added to the choice set, and the attributes of such alternative affects the utility of other alternatives in such a way that their choice probabilities increase, then regularity is violated, and the model is not consistent with RUM.

The ASU-DNN model (Wang et al., 2020b) overcomes this limitation. The ASU-DNN is an ANN-based model that features alternative-specific utility functions that are approximated from the data. Formally, the ASU-DNN specifies the utility as in equation 5.2:

$$U_{nj} = V_{nj}(X_{nj}, c_{nj}, w_j) + \varepsilon_{nj}, \tag{5.2}$$

where $V_{nj}$ is the observed utility of alternative $j$ and $w_j$ is an alternative-specific vector of weights to be estimated for alternative $j$. As a difference with the model of equation 5.1, the ASU-DNN model is consistent with RUM since $V_{nj}$ only depends of its corresponding attributes and costs.

However, the ASU-DNN does not restrict the cost-dependent utility function to have the same form across different alternatives, which implies that the fungibility assumption does not hold. Behaviourally speaking, in the ASU-DNN, a decision-maker could value one euro spent on a specific alternative in an intrinsically different way than the same euro spent on another alternative. However, a key requirement for deriving meaningful welfare measures is that money is a perfect substitute of itself.

Thus, under the ASU-DNN, we cannot ensure that euros are interchangeable across different alternatives of the choice set, which makes welfare measures unfeasible to compare.

To remedy this issue, we propose a to modify the utility function of equation 5.2 to incorporate the fungibility of money assumption, such as in equation 5.3:

$$U_{nj} = f_j(X_{nj}; w_j) + g_j(c_{nj}; \bar{w}_c) + \varepsilon_{nj}, \tag{5.3}$$

where $f_j(\cdot)$ is an alternative-specific utility function of alternative $j$ that depends only on the non-cost attributes of the same alternative and $g_j(\cdot)$ is the utility function of alternative $j$ that depends only on the costs of the same alternative. The vectors $w_j$ and $\bar{w}_c$ are weights (parameters) to be estimated, which define the shape of $f_j$ and $g_j$, respectively. The vectors $w_j$ are alternative-specific, meaning they are independent across alternatives, whereas the vector $\bar{w}_c$ is shared across alternatives.

This model is consistent with RUM since the utility functions are alternative-specific and only depend on their corresponding attributes (in the same way as the ASU-DNN). In addition, this model is also consistent with the fungibility of money assumption since the cost-dependent utility functions only depend on their own attributes, but the weights are shared across alternatives. Therefore, the cost-dependent utility functions share the same shape, leading to the same marginal utility of costs across different alternatives for a given cost value.

### 5.2.2. The alternative-specific and shared weights neural network (ASS-NN)

The ASS-NN is an ANN structure aimed to approximate the utility function of equation (5.3). Figure 5.1 illustrates an example of the ASS-NN structure for a choice situation composed of three alternatives and three attributes per alternative, namely travel cost, travel time and an additional attribute that is present only in the first two alternatives. The travel costs of each alternative are modelled with an alternative-specific hidden layer with shared weights, named as “shared layer”. The shared layer only receives the information from the costs of its correspondent alternative and translates it into a single utility value. The non-cost attributes, i.e. the travel time and the additional attribute for this example, are modelled with regular alternative-specific hidden layers (i.e., with independent weights), named as “alternative-specific-utility (ASU) layers”. For each alternative, the utility values coming from the shared ASU layers are summed to form a single utility associated to the alternative. Optionally, the utility of each alternative can incorporate bias nodes that mimic the alternative-specific constants of a discrete choice model. Finally, the utility values are transformed to choice probabilities using a Softmax function that guarantees that the probabilities sum up to one.

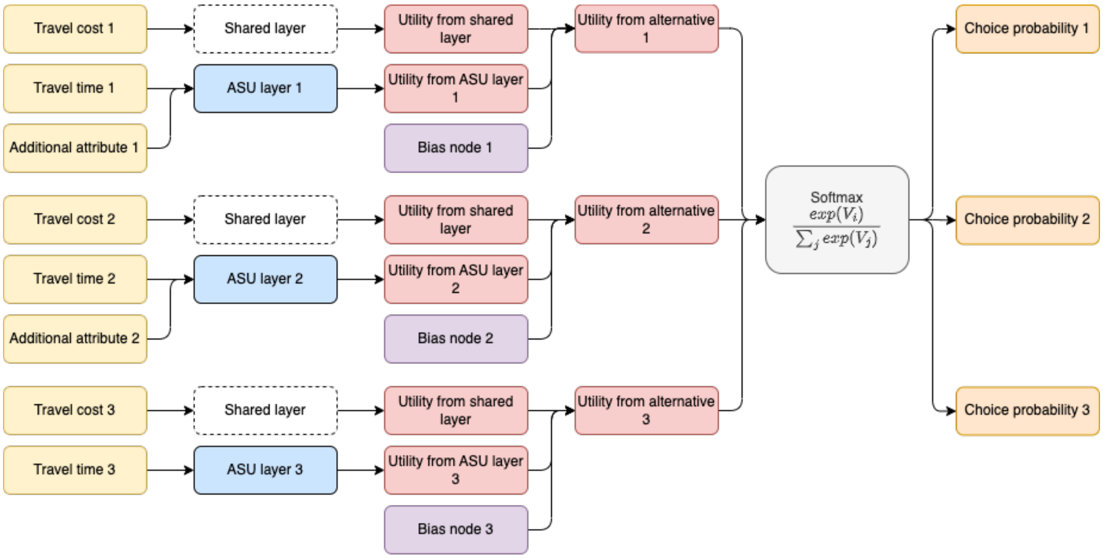


*Figure 5.1: Illustration of the ASS-NN for a 3-alternatives choice situation*

Given data with $N$ observations, the ASS-NN is trained by finding the weights that minimise the categorical cross-entropy (CE) function described by equation (5.4):

$$CE(y, p; w) = (1/N) \cdot \sum_{n=1}^{N} \sum_{j=1}^{J} \ln \left(p_{nj}\right) \cdot y_{nj}) \tag{5.4}$$

where $p_{nj}$ is the probability of choosing alternative $j$ by decision-maker $n$, and $y_{nj}$ is a binary variable that is equal to one when alternative $j$ is chosen by the decision-maker. The CE function is an averaged version of the log-likelihood function of discrete choice models.

## 5.2.3. Implementation and training of the ASS-NN

The ASS-NN is implemented in three steps. In the first step, we prepare the data in a compatible format, and we split it into a training (estimation) and out-of-sample testing (prediction) dataset. The second step consists of finding the optimal hyperparameters. The last step consists of training the ASS-NN and deriving outcomes from it using simulation, namely the marginal utility of attribute increases and the VTT (Small, 2012), a specific form of the willingness to pay for attribute changes based on marginal rates of substitution. Below, we detail each of these three steps.

### Step 1: Data preparation

Table 5.1 describes the basic dataset structure for the ASS-NN. The data is arranged in "wide" format, the standard format in choice modelling statistical packages (e.g., Biogeme, Apollo). Each row represents a single choice situation, while each column represents the variables of such choice situations. The minimum variable requirements are 1) an integer variable that identifies the selected alternative (Choice), 2) the cost attributes that are modelled with shared layers (TC1 and TC2), and 3) the non-cost attributes that are modelled with ASU layers, such as travel time (TT1 and TT2). The cost variables that are modelled with shared weights must be present in all alternatives by construction, while this is not required for the non-cost attributes that are modelled with ASU layers.

| Respondent ID | Choice | TT1 | TC1 | TT2 | TC2 |
|---|---|---|---|---|---|
| 1 | 2 | 50 | 15 | 45 | 18 |
| 2 | 1 | 65 | 10 | 70 | 8 |
| 3 | 1 | 50 | 10 | 58 | 8 |
| ... | ... | ... | ... | ... | ... |

*Table 5.1: Data format for the ASS-NN*

Before training the ASS-NN, data is normalised to avoid numerical overflow issues during the optimisation of the CE function. We use so-called Min-Max normalisation, in which each variable is scaled between zero and one using the minimum and maximum values of the variable as bounds for the normalisation. Specifically, the Min-Max normalisation applies the transformation detailed in equation 5.5:

$$x_{scaled} = \frac{x - min(x)}{max(x) - min(x)}, \tag{5.5}$$

where $x_{scaled}$ is the scaled value of the variable $x$. For the cost variables, the normalisation is done considering the minimum and maximum values of all cost variables as bounds, since such variables are modelled with hidden layers with shared weights.

After normalising, data is randomly split in a training sample (80% of the data) used to train the ASS-NN, and a test sample (the remaining 20% of the data) used for out-of-sample prediction. This is done because ANNs, in general, are prone to overfit, which hinders the ability of the model to provide generalisable measures for data points that were not used for training. In addition, using split samples prevents data-leakage issues, i.e., when a model learns from the test sample, which may provide overly optimistic predictions (Hillel et al., 2021). We use a stratified random sampling of the data, using the choice variable for stratification to ensure that both training and test samples keep the same observed market shares.

### Step 2: Optimal hyperparameters search

Before training the ASS-NN, it is necessary to determine the optimal hyperparameters. Specifically, we focus on the network topology and the activation functions. On the one hand, the network topology is the number of hidden layers and nodes per layer and determines the ability of the ASS-ANN to learn from the data. A too-simple hyperparameters (i.e., too few hidden layers/nodes) will underfit the training sample, and the ASS-NN will not be able to learn relevant interactions from the data. A too-complex hyperparameters will overfit the training sample, hindering their generalisation outside the domain of the training data. By finding the optimal hyperparameters, we aim for a model that maximises out-of-sample predictive performance. On the other hand, the activation functions are transformations located on each hidden node and their role is to convert the information from each preceding hidden layer and pass the transformed information to the subsequent layers. Non-linear activation functions are defined to allow the ASS-NN to identify nonlinear effects from the data.

We use a grid search procedure for finding the optimal hyperparameters. Table 5.2 summarises the candidate hyperparameters for the hidden layers, nodes per layer and activation functions. A set of ASS-NNs for each possible combination of hyperparameters is trained 100 times each. Repeated training is performed because ANNs are overspecified models which are not necessarily globally concave/convex, making them prone to get stuck in local minima. Performing repeated training mitigates that risk, as the prediction performance of networks with poor solutions can be compensated by networks with a better performance. On each training repetition, 20% of the last observations of the training sample are taken apart, and the ASS-NN is trained using the remaining 80%. This excluded sample, known as validation data, is used to calculate a validation cross-entropy on each training epoch (iteration). The training process stops when the validation cross entropy does not improve for 6 consecutive epochs.

| Parameter | Values |
|---|---|
| Hidden layers | 1, 2 |
| Hidden nodes per layer | 1 layer: 5, 6, 7, 8, 9, 10, 15, 20, 30 |
| | 2 layers: 5, 10, 20, 30 |
| Activation functions | ReLU, Hyperbolic tangent (tanh) |

*Table 5.2: Hyperparameter specifications*

For each training repetition, we use the test sample to compute the log-likelihood (i.e., the unaveraged version of the CE function) and the Rho-squared. The Rho-squared compares the predictive ability of a discrete choice model with a random sampling of the choice probabilities Mokhtarian (2016), as defined in equation (5.6):

$$\rho^2 = 1 - \frac{LL_{test}}{N_{test} \cdot \ln(1/J)}, \tag{5.6}$$

where $LL_{test}$ is the log-likelihood value obtained from the test sample, $N_{test}$ is the test sample size, and $J$ is the number of alternatives. Higher values indicate that the model reaches a better prediction performance than mere random chance. Thus, the optimal hyperparameters correspond to those that result in the model that minimises the test log-likelihood and Rho-squared.

### Step 3: Training and simulation of outcomes

After the optimal hyperparameters are identified, the ASS-NN is trained 100 times using the training dataset to mitigate the possibility of predicting on a single, poor-performing network. The predictions of each network are averaged across the 100 training repetitions. Same as in step 2, on each training repetition, the last 20% of the training data is taken apart to calculate the validation CE on each epoch and end the training process if no further improvements of this metric are found for six consecutive epochs.

The trained ASS-NN is used to derive the marginal utility of attribute increases and the VTT using simulation, following the approach of Wang et al. (2020b). Firstly, for a given decision-maker n and alternative j, the marginal utility (MU) of such alternative with respect to increments of the attribute k is given by equation (5.7):

$$MU_{njk} = \partial \hat{V}_j / \partial x_{njk} = \partial \hat{f}_j / \partial x_{njk} + \partial \hat{g}_j / \partial x_{njk}, \tag{5.7}$$

where $\hat{f}_j$ and $\hat{g}_j$ are the approximated utility functions that depend on the non-cost and cost attributes, respectively. When $x_{njk}$ is non-cost attribute, the $\partial \hat{g}_j / \partial x_{njk} = 0$. In contrast, if $x_{njk}$ is the cost, $\partial \hat{f}_j / \partial x_{njk} = 0$. The MU for attribute increases is computed per decision-maker since $\hat{f}_j$ and $\hat{g}_j$ are functions of the corresponding attribute and, therefore, their associated MUs depend on the attribute values.

The MU for attribute increases provide behavioural information about the decision-makers' preferences for increases in specific attributes. For a decision-maker $n$, if $MU_{njk} > 0$, then increases of the attribute $k$ in alternative $j$ are preferred. Conversely $MU_{njk} < 0$, then increases of the attribute $k$ in alternative $j$ are not preferred by decision-maker $n$. If $MU_{njk} = 0$, decision-maker $n$ utility for alternative $j$ is not affected for changes in attribute $k$. This interpretation is similar as the estimated taste parameters of a linear-in-parameters MNL model, but in an individual-specific way. The VTT is constructed as the marginal rate of substitution (MRS) between two attributes. Mathematically, the MRS between attributes $k$ and $l$ for decision-maker $n$ is defined by the ratio of marginal utilities, as shown in equation (5.8):

$$MRS_n^{kl} = MU_{njk}/MU_{njl} \tag{5.8}$$

The MRS between two attributes provides information about the extent that a given decision-maker is willing to substitute attributes $k$ and $l$ to keep their utility for alternative $j$ without changes.

When the denominator of equation (5.8) is the MU of cost, the MRS becomes the VTT, which is the willingness-to-pay for reductions in travel time, in terms of travel costs (Small, 2012). Similar willingness-to-pay measures can be derived from the VTT expression. For instance, the ratio between the MU of travel headway (i.e., the time distance between train services) and the MU of costs is known as the Value of Waiting Time (VoWT), which is the willingness to pay for increasing the frequency of public transport services, in terms of travel costs. Both the VTT and VoWT are defined by equations (5.9) and (5.10):

$$VTT_n = MU_n^{TT}/MU_n^{TC} \tag{5.9}$$

$$VoWT_n = MU_n^{FREQ}/MU_n^{TC} \tag{5.10}$$

## 5.3. Monte Carlo analysis

To show the extent that the ASS-NN learns the utility from the data, we conducted a Monte Carlo analysis. Specifically, we generate pseudo-synthetic data based on the stated preference (SP) part of the Swissmetro dataset (Bierlaire et al., 2001), hereafter the Swissmetro data. The Swissmetro data is a mode choice experiment widely known in the transportation research community and has been used before as a benchmark dataset for ANN-based discrete choice models (e.g., Sifringer et al., 2020). The Swissmetro dataset was carried out in 1998 in Switzerland to elicit travellers' preferences for the Swissmetro, an innovative rail-based rapid transport mode. Respondents were presented with hypothetical mode choice situations based on their current trip offering three alternatives: train, Swissmetro or car. Each alternative varied in terms of their travel time in minutes, travel cost in Swiss Francs (CHF) and headway (for train and Swissmetro) in minutes between each service. After pre-processing the data and cleaning choice situations with less than 3 alternatives, the experimental design to generate pseudo-synthetic data comprises 9,036 combinations of attributes from 1,858 individuals.

### 5.3.1. Pseudo-synthetic data generation

Pseudo-synthetic data is generated by using the experimental design of the Swissmetro dataset (i.e., the attributes of each alternative) to generate synthetic choices, based in RUM models under different utility function specifications and "true" parameters (Garrow et al., 2010). As we know the "true" DGP a priori, we can contrast the outcomes of the ASS-NN, namely goodness-of-fit measures, marginal utilities and willingness to pay measures, with the "true" outcomes.

We generate two pseudo-synthetic datasets, using the travel time and travel cost of each alternative from the Swissmetro dataset and two different model specifications of the utility function that are commonly observed on empirical applications of discrete choice models. Table 5.3 summarises the specification of both datasets. The first dataset is generated using a linear-in-parameters utility function. This specification leads to marginal utilities equal to the corresponding parameters associated to travel time and travel cost, respectively. Furthermore, the marginal utilities are constant across different modes, which in turn determines that the VTT is the same for all modes. The second dataset follows a log-linear utility function. To avoid numerical overflow of the natural logarithm when the travel cost is zero (i.e., when a respondent holds an annual discount card), we added a constant of 0.1 to all attributes. Under the log-linear utility specification, the marginal utilities and VTT depend on the current travel time and cost of each respondent, which implies that the VTT could differ across different travel modes and respondents.

| Name | Utility function | True parameters | Marginal utility | VTT |
|---|---|---|---|---|
| Dataset 1 | $V_j = \beta_{TC} \cdot TC_j + \beta_{TT} \cdot TT_j$ | $\beta_{TC} = -2$ | $MU_{TC} = -2$ | $VTT = 3/2$ |
| (Linear) | | $\beta_{TT} = -3$ | $MU_{TT} = -3$ | (for all alternatives) |
| Dataset 2 | $V_j = \beta_{TC} \cdot \ln(TC_j + 0.1) + \beta_{TT} \cdot \ln(TT_j + 0.1)$ | $\beta_{TC} = -3$ | $MU_{TC} = -3/(TC_j + 0.1)$ | $VTT = (3/5) \cdot \frac{TC_j+0.1}{TT_j+0.1}$ |
| (Log-linear) | | $\beta_{TT} = -5$ | $MU_{TT} = -5/(TT_j + 0.1)$ | |

*Table 5.3: Model specification of pseudo-synthetic datasets*

We expect the ASS-NN to recover the marginal utility of travel time and cost increases, as well as the VTT values. Furthermore, we expect that the average values of the marginal utility and VTT are close to the corresponding true values. In addition, we expect the ASS-NN outperforms a MNL model with a linear-in-parameters utility function as the DGP departs from a linear model.

## 5.3.2. Results of the Monte Carlo Analysis

Table 5.4 summarises the training results across the 100 repetitions of the ASS-NN, trained on each pseudo-synthetic dataset. We present the log-likelihood evaluated in the full, train and test samples and the Rho-squared evaluated in the test sample. These values are contrasted with the true goodness-of-fit measures and those obtained from a MNL model with a linear utility function. In addition, we present the optimal hyperparameters of the ASS-NN for each dataset. The measures of the linear MNL model are included to show the extent that a misspecified model leads to poor predictive performance when the model assumptions are not aligned with the true DGP.

| | Dataset 1 (linear) | | | Dataset 2 (log-linear) | | |
|---|---|---|---|---|---|---|
| | True value | Linear MNL | ASS-NN | True value | Linear MNL | ASS-NN |
| Log-likelihood (full sample) | -5,807.57 | -5,807.07 | **-5,809.23** | -4,535.97 | -5,056.09 | **-4,562.64** |
| Log-likelihood (training sample) | -4,621.36 | -4,620.88 | **-4,623.59** | -3,618.93 | -4,041.19 | **-3,641.88** |
| Log-likelihood (test sample) | -1,186.20 | -1,186.19 | **-1,186.00** | -917.04 | -1,014.90 | **-920.76** |
| Rho-squared (test sample) | 0.40 | 0.40 | **0.40** | 0.54 | 0.49 | **0.54** |
| Estimation/training time (secs.) | - | <1 s. (total) | **6.73s./train** | - | <1 s. (total) | **13.04/train** |
| Optimal number of hidden layers | - | - | **1 layer** | - | - | **2 layers** |
| Optimal number of nodes per layer | - | - | **15 nodes** | - | - | **10 nodes** |
| Activation function | - | - | **tanh** | - | - | **tanh** |

Note: The goodness-of-fit metrics of the ASU-NN are the averaged values across the 100 repetitions, per dataset.

*Table 5.4: Training results of the ASS-NN against true values and from a linear MNL model*

We observe that the ASS-NN reaches a goodness-of-fit close to the true values, in all samples and datasets, suggesting this model succeeds in approximating the utility function from the data. Compared with a linear MNL model, the ASS-NN reaches a negligibly lower predictive performance than the choice model, i.e., lower log-likelihood and Rho-squared, when the data is generated with a linear-in-parameters utility function (dataset 1). This result is expected, as the linear MNL is correctly specified for this pseudo-synthetic dataset. However, when the data is generated with a log-linear utility function (dataset 2), the ASS-NN outperforms the linear MNL model in terms of goodness-of-fit measures, e.g., a Rho-squared in test sample of 0.54 for the ASS-NN, against 0.49 for the linear MNL model.

Table 5.5 summarises the mean, bias and Root Mean Squared Error (RMSE) of the marginal utilities obtained with the ASS-NN, contrasted with the true values and with the values of a MNL model with a linear utility function. The mean of the marginal utility is presented since the ASS-NN predicts these values at the choice situation level. In contrast, the bias and RMSE are calculated to quantify the extent that the marginal utilities deviate from the true values.

| | | Mean | | | Bias | | RMSE | |
|---|---|---|---|---|---|---|---|---|
| | | True value | Linear MNL | **ASS-NN** | Linear MNL | **ASS-NN** | Linear MNL | **ASS-NN** |
| Dataset 1 (Linear) | Train cost | -2 | -2.01 | **-2.00** | -0.01 | **<-0.01** | 0.01 | **0.07** |
| | Train time | -3 | -3.05 | **-2.86** | -0.05 | **0.13** | 0.05 | **0.24** |
| | SM cost | -2 | -2.01 | **-1.98** | -0.01 | **0.02** | 0.01 | **0.11** |
| | SM time | -3 | -3.05 | **-2.90** | -0.05 | **0.09** | 0.05 | **0.19** |
| | Car cost | -2 | -2.01 | **-2.01** | -0.01 | **-0.01** | 0.01 | **0.05** |
| | Car time | -3 | -3.05 | **-2.92** | -0.05 | **0.08** | 0.05 | **0.14** |
| Dataset 2 (Log-linear) | Train cost | -5.53 | -2.14 | **-4.06** | 3.39 | **1.68** | 7.41 | **5.28** |
| | Train time | -3.23 | -3.67 | **-3.03** | -0.44 | **0.20** | 1.54 | **0.59** |
| | SM cost | -5.00 | -2.14 | **-3.60** | 2.87 | **1.60** | 7.21 | **5.26** |
| | SM time | -6.04 | -3.68 | **-5.31** | 2.36 | **0.73** | 3.66 | **1.91** |
| | Car cost | -3.59 | -2.14 | **-3.39** | 1.45 | **0.19** | 2.36 | **0.37** |
| | Car time | -3.90 | -3.68 | **-3.57** | 0.22 | **0.32** | 1.90 | **0.97** |

*Table 5.5: Marginal utilities of the ASS-NN against true values and from a linear MNL model*

For the linear-in-parameters utility DGP (dataset 1), the ASS-NN succeeds on recovering the true marginal utility travel cost on the average, with small numerical differences between the mean estimate and the true values. In contrast, for the marginal utility of travel time, these differences are higher. This can be explained by the selected structure of the ASS-NN. On the one hand, the cost-specific utility is modelled as a set of layers with shared weights across alternatives, which are trained with the attribute levels of all alternatives together. Thus, the cost-specific utility is modelled with a lower number of weights and a bigger amount of data than the other attributes. On the other hand, the travel time is modelled with independent sets of layers per alternative, with their own sets of independent weights, which are trained only with the data for its specific alternative.

For the log-linear utility DGP (dataset 2), the ASS-NN outperforms the MNL model with linear function on recovering all the predicted marginal utilities on average, compared with the true values. Furthermore, the ASS-NN consistently reaches lower average bias and RMSE than the linear MNL model. We also observe that the ASS-NN has a greater ability of recovering the marginal utility of travel time than the marginal utility of travel cost, as the bias and RMSE of the former are lower than the latter, in contrast to the findings of dataset 1, where the ASS-NN gets closer-to-truth marginal utilities of cost than marginal utilities of time. Table 5.6 summarises the mean, bias and RMSE of the VTT for each mode, contrasted with the true values from the DGP and a linear MNL model. On average, the ASS-NN successfully recovers the true VTT per mode, with different degrees of precision per pseudo-synthetic dataset. On the one hand, when the DGP is a linear utility function (dataset 1), the mean bias of the VTTs obtained from the ASS-NN is between -0.15 and -0.07, while the RMSE values lie between 0.05 and 0.03. Both values are higher than those from a linear MNL model. On the other hand, when the true DGP is a log-linear function (dataset 2), the mean VTT values of the ASS-NN are closer to the true values compared with the VTTs obtained from the linear MNL model, which is also reflected in the differences of the bias and RMSE between each model. Furthermore, we observe higher precision of the ASS-NN on predicting VTTs than marginal utilities, as the bias and RMSE of the former are considerably smaller than those from the latter.

| | | Mean | | | Bias | | RMSE | |
|---|---|---|---|---|---|---|---|---|
| | | True value | Linear MNL | **ASS-NN** | Linear MNL | **ASS-NN** | Linear MNL | **ASS-NN** |
| Dataset 1 | Train | 1.50 | 1.52 | **1.42** | 0.02 | **-0.08** | 0.02 | **0.01** |
| | Swissmetro | 1.50 | 1.52 | **1.45** | 0.02 | **-0.05** | 0.02 | **0.01** |
| | Car | 1.50 | 1.52 | **1.44** | 0.02 | **-0.05** | 0.02 | <**0.01** |
| Dataset 2 | Train | 0.98 | 1.72 | **0.85** | 0.73 | **0.01** | 1.14 | **0.02** |
| | Swissmetro | 2.24 | 1.72 | **1.87** | -0.52 | **-0.06** | 2.11 | **0.19** |
| | Car | 1.17 | 1.72 | **1.07** | 0.55 | **-0.03** | 0.72 | **0.04** |

*Table 5.6: VTT of the ASS-NN, compared with true values and a linear MNL model*

## 5.4. Application with empirical data

The results of these Monte Carlo analyses show that the ASS-NN can learn the utility function from the data and provides interpretable outcomes close to the true values. In this section, we apply the ASS-NN to empirical choices from the Swissmetro data.

### 5.4.1. Data description

The data used in this section is the same as in the Monte Carlo analysis, i.e., the Swissmetro dataset, but includes the respondents' actual choices, instead of using pseudo-synthetic choices. The empirical dataset comprises 9,036 choice situations from 1,858 individuals, with their respective attribute levels and responses. Table 5.7 summarises the observed market shares, which evidence unbalance of the chosen transport modes. The most chosen mode is Swissmetro, with 5,177 observations (57.3% of the sample), followed by car with 3,080 observations (34.1% of the sample), and train with 779 times (8.6% of the sample).

| | Frequency | % |
|---|---|---|
| Train | 779 | 8.6% |
| Car | 5,177 | 57.3% |
| Swissmetro | 3,080 | 34.1% |

*Table 5.7: Observed market shares per travel mode*

We use the travel cost and travel time of each travel mode and the travel headway of train and Swissmetro as inputs of the ASS-NN. The travel cost variables of each mode are modelled with hidden layers with shared weights, while the travel time and headway are modelled with alternative-specific layers with independent weights. All attributes are scaled to hundreds (i.e., divided by 100) in order to avoid numerical overflow issues in the MNL models.

Table 5.8 presents the summary statistics of the (unscaled) attributes used for the analysis. On average and median, Swissmetro is the most expensive mode, but at the same time is the fastest in terms of travel time. The minimum travel cost of train and Swissmetro equals zero, as a difference from the minimum travel cost for car. This is because some respondents stated they own an annual public transport card that allows them to travel for free. The cost of such a card can be assumed as sunk costs (since travellers already paid for it), and therefore, the travel cost for such respondents is zero regardless of the trip they select as long as it is by public transport. In terms of travel time, Swissmetro is the fastest mode on average and in terms of mean values, followed by car and train.

Figure 5.2 shows the distribution of the (unscaled) travel time and travel costs per mode, which were designed following a pivoted design, i.e., based on the respondents' current travel time and travel cost. The travel headway, on the other hand, was designed based on three possible levels per mode: the trains' headway could be either 30, 60 or 120 minutes per train, whereas the headway of Swissmetro could be either 10, 20 or 30 minutes per service. The distributions of travel cost and travel time per mode are skewed toward the left. Most respondents faced travel costs between zero and 400 CHF, while higher values can be considered outliers. A considerable part of

| | | Mean | Median | Std. Dev. | Minimum | Maximum |
|---|---|---|---|---|---|---|
| Cost | Train | 91.27 | 81 | 65.51 | 0 | 576 |
| (CHF) | Swissmetro | 110.19 | 97 | 80.47 | 0 | 768 |
| | Car | 93.43 | 84 | 47.48 | 8 | 520 |
| Time | Train | 173.69 | 167 | 78.58 | 31 | 1,049 |
| (minutes) | Swissmetro | 91.38 | 81 | 55.76 | 8 | 796 |
| | Car | 146.88 | 136 | 77.17 | 32 | 1,560 |
| Headway | Train | 70.04 | 60 | 37.42 | 30 | 120 |
| (mins/service) | Swissmetro | 20.05 | 20 | 8.16 | 10 | 30 |

*Table 5.8: Summary statistics of the attributes per alternative*

respondents has travel costs equal to zero for train and Swissmetro, which are those who own the annual public transport card. Regarding travel time, most respondents faced between zero and 4 hours (approximately 200 minutes), with slightly longer travel times for train and car trips, compared with Swissmetro trips.

### 5.4.2. MNL models for contrast

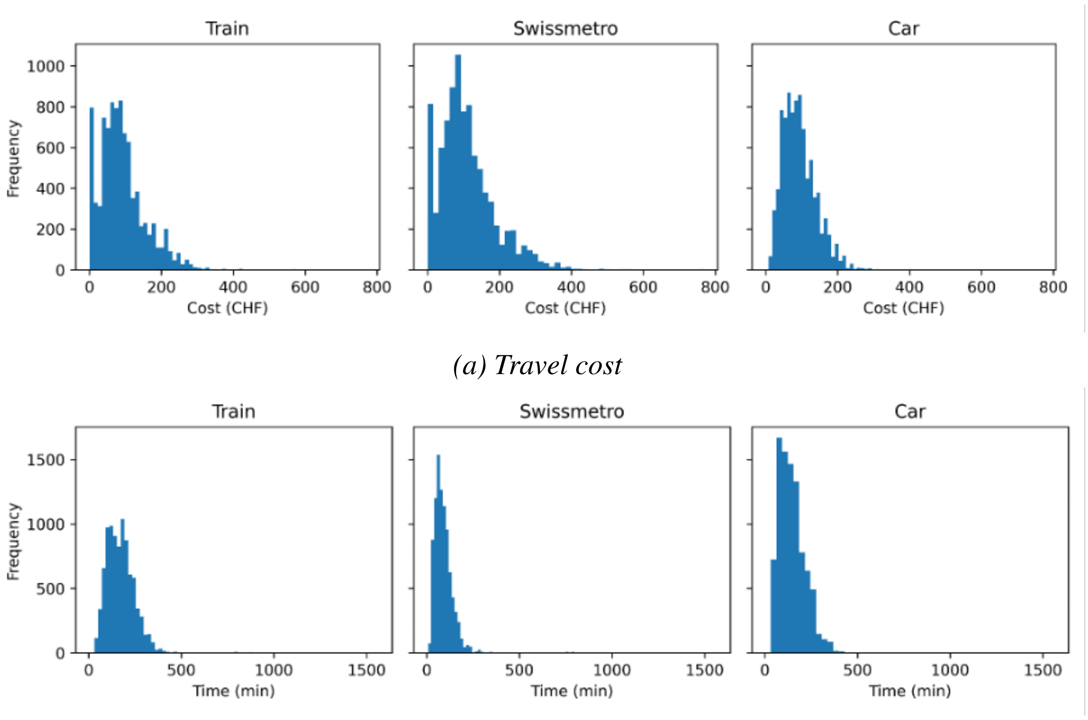
In addition to comparing with the results of the ASU-DNN, we contrast the results of the ASS-NN model with two MNL models with different utility function specifications. The first model is specified with a linear-in-parameters utility function (henceforth a linear MNL model), in which the cost-specific parameter is equal across alternatives, whereas the parameters of travel time and headway are alternative-specific. Additionally, we include alternative-specific constants to reflect the labelled nature of the choice experiment. Thus, the (observed) utility of this model is defined as in equations (5.11) to (5.13):

$$V_{TRAIN} = \beta_{TC} \cdot TC_{TRAIN} + \beta_{(TT,TRAIN)} \cdot TT_{TRAIN} + \beta_{(HE,TRAIN)} \cdot HE_{TRAIN} \tag{5.11}$$

$$V_{SM} = \alpha_{SM} + \beta_{TC} \cdot TC_{SM} + \beta_{(TT,SM)} \cdot TT_{SM} + \beta_{(HE,SM)} \cdot HE_{SM} \tag{5.12}$$

$$V_{CAR} = \alpha_{CAR} + \beta_{TC} \cdot TC_{CAR} + \beta_{(TT,CAR)} \cdot TT_{CAR} \tag{5.13}$$

where $\alpha_{SM}$ and $\alpha_{CAR}$ are alternative-specific constants, $\beta_{TC}$ is the alternative-shared parameter associated with travel cost, $\beta_{(TT,j)}$ are the alternative-specific parameters associated with travel time, and $\beta_{(HE,j)}$ are the alternative-specific parameters associated with headway.



*(a) Travel cost*

*(b) Travel time*

*Figure 5.2: Distribution of travel cost and travel time per mode*

The second model is a MNL model with a log-linear utility function (henceforth the log-linear MNL model), which is formulated in a similar way as in the linear MNL model, but with attributes in logarithms. To address numerical overflow when the travel cost is equal to zero, we add a constant to all attributes equal to 0.1. The utility functions are specified as in equations (5.14) to (5.16):

$$\begin{aligned} V_{TRAIN} &= \beta_{TC} \cdot \ln(TC_{TRAIN} + 0.1) \\ &\quad + \beta_{(TT,TRAIN)} \cdot \ln(TT_{TRAIN} + 0.1) + \beta_{(HE,TRAIN)} \cdot \ln(HE_{TRAIN} + 0.1) \qquad (5.14) \\ V_{SM} &= \alpha_{SM} + \beta_{TC} \cdot \ln(TC_{SM} + 0.1) \\ &\quad + \beta_{(TT,SM)} \cdot \ln(TT_{SM} + 0.1) + \beta_{(HE,SM)} \cdot \ln(HE_{SM} + 0.1) \qquad (5.15) \\ V_{CAR} &= \alpha_{CAR} + \beta_{TC} \cdot \ln(TC_{CAR} + 0.1) \\ &\quad + \beta_{(TT,CAR)} \cdot \ln(TT_{CAR} + 0.1) \qquad (5.16) \end{aligned}$$

### 5.4.3. Results with empirical data

**Goodness of fit**

Table 5.9 summarises the training results of the ASS-NN, contrasted with the ASU-DNN and the MNL models. The ASU-DNN reaches the highest log-likelihood in all samples and Rho-squared in the test sample, followed by the ASS-NN. This result is expected since the ASS-NN is a restricted version of the ASU-DNN. Compared with the MNL models, the ASS-NN outperforms both linear and log-linear MNL models in terms of Rho-squared (test sample): the difference between the ASS-NN and the linear MNL model is of 0.3 points and 0.1 points for the log-linear MNL model. These results sign potential non-linear effects in the underlying DGP that a linear MNL models does not account for, whereas they are captured to a greater extent by the log-linear MNL, the ASU-DNN and the ASS-NN. In terms of estimation/training time, the MNL models are considerably faster, with less than one second of estimation time, while the ASS-NN and the ASU-DNN take between 10 and 16 seconds per training repetition, on average.

**Marginal utility of attribute increases**

Table 5.10 presents the average marginal utilities obtained with the ASS-NN, compared with the predictions of the ASU-DNN and MNL models. The ASS-NN predicts an average marginal utility of cost that slightly varies across modes (between -1.51 and -1.34), which is explained by the differences in travel costs presented in the choice

| | **ASS-NN** | ASU-DNN | Linear MNL | Log-linear MNL |
|---|---:|---:|---:|---:|
| Log-likelihood (full) | **-6,940.79** | -6,747.95 | -7,214.94 | -6,994.60 |
| Log-likelihood (train) | **-5,548.44** | -5,388.77 | -5,767.31 | -5,591.37 |
| Log-likelihood (test) | **-1,392.35** | -1,359.18 | -1,447.63 | -1,403.24 |
| Rho-squared (test) | **0.30** | 0.32 | 0.27 | 0.29 |
| Estimation/train time | **9.98s./training** | 16.24s./training | <1sec. (total) | <1sec. (total) |
| Number of hidden layers | **2 layers** | 2 layers | - | - |
| Number of nodes per layer | **10 nodes** | 10 nodes | - | - |
| Activation function | **tanh** | **tanh** | - | - |

*Table 5.9: Training results of the ASS-NN compared with the ASU-DNN and MNL models, empirical data.*

experiment (see summary statistics of Table 9). In contrast, the ASU-DNN predicts considerably different marginal utilities of cost per mode (between -2.56 and -0.86). This is expected, as the ASU-DNN does not restrict the cost-dependent utilities to have the same form across different alternatives. Compared to the MNL models, the ASS-NN predicts a higher marginal utility of costs than the log-linear MNL model (between -2.03 and -1.28). Finally, the linear MNL model predicts a constant marginal utility of cost of -0.81.

| Mode | Attribute | **ASS-NN** | ASU-DNN | Linear MNL | Log-linear MNL |
|---|---|---:|---:|---:|---:|
| Train | Cost (x100) | **-1.51** | -2.56 | -0.81 | -2.03 |
| | Time (x100) | **-2.06** | -1.80 | -2.04 | -2.28 |
| | Headway (x100) | **-0.88** | -0.91 | -0.78 | -0.92 |
| Swissmetro | Cost (x100) | **-1.34** | -1.43 | -0.81 | -1.85 |
| | Time (x100) | **-2.06** | -2.06 | -1.51 | -2.34 |
| | Headway (x100) | **-1.12** | -1.19 | -0.73 | -0.71 |
| Car | Cost (x100) | **-1.48** | -0.86 | -0.81 | -1.28 |
| | Time (x100) | **-1.10** | -1.46 | -1.00 | -1.30 |

*Table 5.10: Average predicted marginal utilities ASS-NN, ASU-DNN and MNL models. Empirical data.*

In terms of travel time, the ASS-NN predicts a higher average marginal utility for train and Swissmetro trips (-2.06) than for car trips (-1.10), which signs that individuals have a higher sensitivity for saving travel time for average train and Swissmetro trip than for the average car trip. The ASU-DNN follows a similar pattern, with higher marginal utility of time for train and Swissmetro trips (-1.80 and -2.06, respectively), compared with car trips (-1.46). Compared with the MNL models, we observe that the ASS-NN is more conservative than a log-linear MNL model regarding the average marginal utility of travel time of train and Swissmetro trips, while for car trips, the marginal utility of time is the same. In contrast, compared with the linear MNL model, the ASS-NN predicts a higher average marginal utility of travel time for Swissmetro and car trips. Finally, the ASS-NN predicts a higher predicted marginal utility of headway for Swissmetro trips (-1.12) than train trips (-0.88). In contrast, the MNL models predict higher values for train than for Swissmetro trips.

To explore differences in the marginal utility for different attribute levels, we plot the predicted marginal utilities of the ASS-NN for each mode against their respective attribute values. These plots are illustrated in Figure 5.3. As expected, the predicted marginal utility of costs is equal for the same cost levels across different travel modes due to the consistency of the ASS-NN with the fungibility of money assumption. Conversely, the ASS-NN predicts different shapes of the marginal utility of travel time and headway per travel mode at different values of the associated levels, respectively. For travel time, we observe that the marginal utility of car trips is more inelastic than train and Swissmetro trips. Furthermore, the predicted marginal utilities of travel time suggest that, for trips up to 180 minutes (3 hours) approximately, individuals are more sensitive to changes in travel time for train and Swissmetro trips than for car trips, ceteris paribus. For trips between 180 and 300 minutes (3 to 5 hours) approximately, respondents are more sensitive to changes in travel time for train and car trips than for Swissmetro trips, ceteris paribus. For trips of 300 minutes (5 hours) approximately or more, individuals are more sensitive to travel time changes of car trips than the other modes. Finally, we observe that the slope of the marginal utility of travel headway for Swissmetro trips is steeper than for train trips as the waiting time per service increases. This suggests that, in terms of travel headway, the marginal utility for Swissmetro trips is more elastic than for train trips.

### VTT and VoWT

Table 5.11 summarises the average predicted VTT and VoWT by the ASS-NN per travel mode, contrasted with the predictions of the ASU-DNN and MNL models. The VTT and VoWT values presented in this table are computed after dropping outliers (upper 5% of the sample), negative VTT values (up to 16 observations) and negative VoWT values (up to 105 observations). The same procedure was conducted for the VTT and VoWT of the MNL models.

The ASS-NN predicts that Swissmetro trips have the highest average VTT (2.11

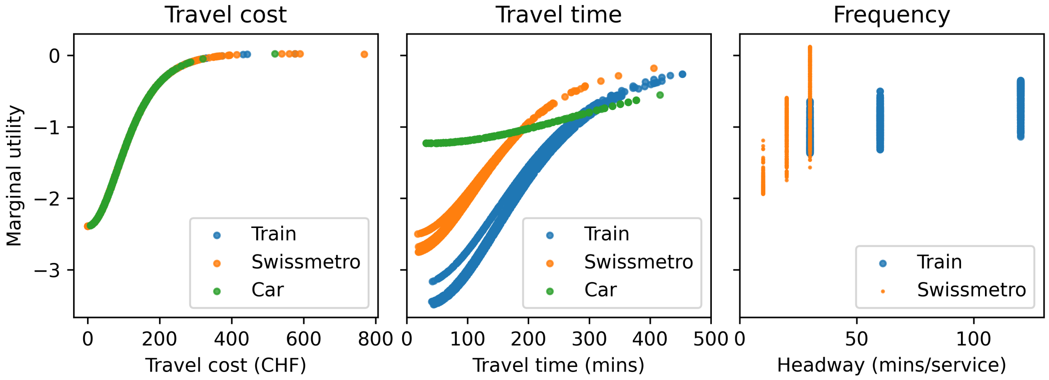


*Figure 5.3: Predicted marginal utility with respect to travel time, cost and headway, empirical data.*

| | VTT (CHF/min) | | | | VoWT (CHF/min) | | | |
|---|---|---|---|---|---|---|---|---|
| | ANN-based | | MNL | | ANN-based | | MNL | |
| | **ASS-NN** | ASU-DNN | Linear | Log-linear | **ASS-NN** | ASU-DNN | Linear | Log-linear |
| Train | **1.52** | 0.69 | 2.52 | 1.73 | **0.68** | 0.34 | 0.96 | 0.78 |
| Swissmetro | **2.11** | 1.51 | 1.86 | 2.13 | **1.14** | 0.90 | 0.90 | 0.72 |
| Car | **0.81** | 1.70 | 1.24 | 1.03 | **-** | | - | - |

*Table 5.11: Average predicted VTT and VoWT compared with the ASU-DNN and MNL models, empirical data.*

CHF/min), followed by train trips (1.52 CHF/min) and car trips (0.81 CHF/min). This order pattern is similar to the log-linear MNL model, whereas the linear MNL model predicts the highest average VTT for train trips, followed by Swissmetro and car trips. In contrast, the ASU-DNN predicts a higher average VTT for car trips, followed by Swissmetro and train trips. In magnitude, the ASS-NN consistently predicts more conservative average VTT values than the MNL models, except in the case of Swissmetro trips, where the lowest predicted VTT is in the linear MNL model. Compared to the ASU-DNN, the ASS-NN predicts a higher VTT except for car trips. In terms of the VoWT, the ASS-NN predicts the highest value for Swissmetro trips (1.14 CHF/min), followed by train trips (0.68 CHF/min). The same pattern in followed in the ASU-DNN, where Swissmetro trips have the highest VoWT (0.9 CHF/min), followed by train trips (0.34 CHF/min). In contrast, in the MNL models, the highest predicted VoWT is associated with train trips. However, the differences across different modes for the same model are not substantial in magnitude in the MNL models. Furthermore, it is more reasonable that the VoWT of Swissmetro trips would be higher than for train trips, as the former trips have a higher frequency and are more expensive than the latter trips.

Finally, Figure 5.4 compares the average VTT of the ASS-NN per travel mode for different trip times. We observe that for short trips (less than 60 minutes), the average VTT of train and Swissmetro trips are similar and close to 2 CHF/min, whereas the average VTT for car trips is considerably lower (0.5 CHF/min approx.). For trips between 60 and 89 minutes (1 to 1.5 hours), the average VTT of Swissmetro trips rises considerably to reach almost 2.5 CHF/min, while it decreases for train trips to 1.5 CHF/min, which evidences a potential mode switch from train to Swissmetro for trips of this time range. As the travel time increases, the average VTT of Swissmetro trips decreases, to the extent of intersecting the average VTT of train trips in the group of 180 to 239 minutes (3 to 4 hours), while the average VTT of car trips steadily increases. We can expect that, for longer trips, the average VTT of a car would surpass the average VTT of train and Swissmetro, if these results follow the same trajectory, which can be evidence that, for long trips, individuals would switch to a car for such travel.

## 5.5. Discussion and conclusion

In this paper, we propose a new discrete choice model based on ANNs, called ASS-NN. The ASS-NN is based in the ASU-DNN, which balances a flexible utility approximation from the data and satisfies consistency with RUM theory. In addi-

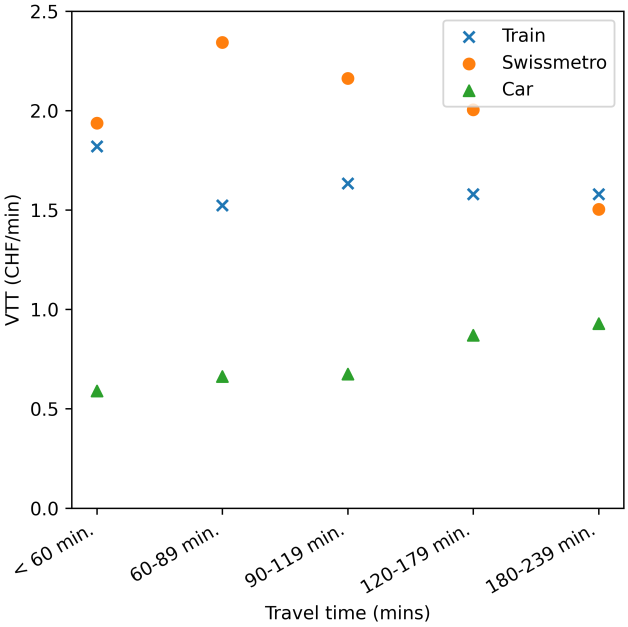


*Figure 5.4: Average VTT per mode for different travel time ranges.*

tion, the ASS-NN is consistent with the assumption of fungibility of money (i.e., one euro is one euro). By accommodating for this assumptions, the ASS-NN considers that money can be spent in different goods interchangeably, which is key for deriving economically-sound welfare measures.

### 5.5.1. Main findings

The Monte Carlo experiment shows that the ASS-NN succeeds on approximating the utility from the data under different model specifications, reaching goodness of fit (i.e., log-likelihood and Rho-squared) close to the ground truth. Furthermore, we show that the ASS-NN has a higher accuracy on recovering the marginal utility of attribute increases and VTT than a misspecified MNL model. The differences between the recovered marginal utilities and VTT across different utility specifications are in line with previous findings in the literature (Torres et al., 2011; van der Pol et al., 2014). Our findings in the Monte Carlo experiment support the use of the ASS-NN for recovering interpretable outcomes without the need of explicitly defining the utility's functional form.

Our empirical results show that, without specifying the utility functional form, the

ASS-NN outperforms (in terms of goodness of fit) MNL models under two different utility specifications widely used in other empirical studies, namely linear and log-linear utility functions. Furthermore, the ASS-NN predicts a marginal utility of costs consistent with the fungibility of money assumption (see Figure 5.3), as a difference with the ASU-DNN. Regarding the marginal utility of travel time, the ASS-NN predicts differences across different modes for the same travel time value, with public transport modes being more attractive for short-distance trips. At the same time, this trend reverts for long-distance trips.

Furthermore, we found that respondents assign a higher average value of travel time (VTT) to public transport trips, particularly train and Swissmetro trips, than car trips. As the trip length increases, the VTT for train and Swissmetro trips decreases in favour of car trips. The study also shows a higher average value of waiting time (VoWT) for Swissmetro trips compared to train trips, which contradicts the predictions of the MNL models.

### 5.5.2. Limitations and further research directions

While we identify promising implications and uses of the ASS-NN, we also acknowledge three limitations of our work. Firstly, ANNs are known for requiring a higher amount of data than DCMs. As shown by (Alwosheel et al., 2018), ANNs require around 50 times the amount of data per estimated weight, much higher than in most conventional choice models, including the MNL model. Such a criterion applied in the context of our paper implies that the number of weights of each ASS-NN should be around 180. A second limitation of our work is the treatment of unreasonably high or below zero VTT/VoWT values. The former case is a consequence of marginal utility values of cost that lie close to zero (Sillano & de Dios Ortúzar, 2005), while the latter is not theoretically possible from an economic perspective (Hess et al., 2005). We relied on dropping such problematic VTT/VoWT values from the sample before presenting the results. However, as such values still may provide relevant behavioural information, further research should be done to treat these cases more elegantly and properly. Finally, the Swissmetro dataset could be rather small for the standard practice of ANN-based models, which can explain the small goodness-of-fit improvements in our applications.

We envision three further research direction from this work. First, we foresee the possibility of exploring methodologies to incorporate panel choices and random parameters to the ASS-NN, equivalent to Mixed Logit models. Panel structures can be incorporated by slight modifications in the network architecture of the ASS-NN. At the same time, random parameters may require alternative network structures that account for probability distributions, such as probabilistic neural networks (Mao et al., 2000). Secondly, we suggest testing the role of sample size on the results of the ASS-NN, for instance, with bigger datasets such as the London Passenger Mode Choice data (Hillel et al., 2018). Thirdly, the envision an extension of the ASS-NN to incorporate sociodemographic characteristics, in a similar way of the work of Sifringer et al. (2020). Overall, we observe that the flexibility of ANNs provides clear opportunities further to incorporate machine learning methodologies in our choice modelling toolbox.

## Acknowledgements

The authors acknowledge the support of the TU Delft AI Labs programme. José Ignacio Hernández would like to acknowledge the Netherlands Organisation for Scientific Research (NWO) for funding his PhD (NWO Responsible Innovation grant – 313-99-333) and the valuable feedback and suggestions from Francisco Garrido Valenzuela during the conception and development of this work. Also, we are thankful for the insightful feedback provided by three anonymous reviewers during the revision stage of this paper.

# Bibliography

Alwosheel, A., S. van Cranenburgh, C. G. Chorus (2018) Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis, *Journal of Choice Modelling*, 28, pp. 167–182.

Ben-Akiva, M., M. Bierlaire (2003) Discrete Choice Models with Applications to Departure Time and Route Choice, in: Hall, R., ed., *Handbook of Transportation Science*, International Series in Operations Research & Management Science, Springer US, Boston, MA, pp. 7–37.

Ben-Akiva, M., S. Lerman, S. Lerman (1985) *Discrete Choice Analysis: Theory and Application to Travel Demand*, MIT Press.

Bentz, Y., D. Merunka (2000) Neural networks and the multinomial logit for brand choice modelling: A hybrid approach, *Journal of Forecasting*, 19(3), pp. 177–200.

Bierlaire, M., K. Axhausen, G. Abay (2001) The acceptance of modal innovation: The case of Swissmetro, in: *Swiss Transport Research Conference*, CONF.

Garrow, L. A., T. D. Bodea, M. Lee (2010) Generation of synthetic datasets for discrete choice analysis, *Transportation*, 37(2), pp. 183–202.

Han, Y., F. C. Pereira, M. Ben-Akiva, C. Zegras (2022) A Neural-embedded Choice Model: TasteNet-MNL Modeling Taste Heterogeneity with Flexibility and Interpretability.

Hess, S., M. Bierlaire, J. W. Polak (2005) Estimation of value of travel-time savings using mixed logit models, *Transportation Research Part A: Policy and Practice*, 39(2), pp. 221–236.

Hess, S., A. Daly, R. Batley (2018) Revisiting consistency with random utility maximisation: Theory and implications for practical work, *Theory and Decision*, 84(2), pp. 181–204.

Hillel, T., M. Bierlaire, M. Z. E. B. Elshafie, Y. Jin (2021) A systematic review of machine learning classification methodologies for modelling passenger mode choice, *Journal of Choice Modelling*, 38, p. 100221.

Hillel, T., M. Z. Elshafie, Y. Jin (2018) Recreating passenger mode choice-sets for transport simulation: A case study of london, uk, *Proceedings of the Institution of Civil Engineers-Smart Infrastructure and Construction*, 171(1), pp. 29–42.

Lancaster, K. J. (1966) A New Approach to Consumer Theory, *Journal of Political Economy*, 74(2), pp. 132–157.

Mao, K., K.-C. Tan, W. Ser (2000) Probabilistic neural-network structure determination for pattern classification, *IEEE Transactions on Neural Networks*, 11(4), pp. 1009–1016.

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

Mokhtarian, P. L. (2016) Discrete choice models' $\rho 2$: A reintroduction to an old friend, *Journal of choice modelling*, 21, pp. 60–65.

Sifringer, B., V. Lurkin, A. Alahi (2020) Enhancing discrete choice models with representation learning, *Transportation Research Part B: Methodological*, 140, pp. 236–261.

Sillano, M., J. de Dios Ortúzar (2005) Willingness-to-Pay Estimation with Mixed Logit Models: Some New Evidence, *Environment and Planning A: Economy and Space*, 37(3), pp. 525–550.

Small, K. A. (2012) Valuation of travel time, *Economics of Transportation*, 1(1), pp. 2–14.

Small, K. A., H. S. Rosen (1981) Applied Welfare Economics with Discrete Choice Models, *Econometrica*, 49(1), pp. 105–130.

Torres, C., N. Hanley, A. Riera (2011) How wrong can you be? Implications of incorrect utility function specification for welfare measurement in choice experiments, *Journal of Environmental Economics and Management*, 62(1), pp. 111–121.

van Cranenburgh, S., A. Alwosheel (2019) An artificial neural network based approach to investigate travellers' decision rules, *Transportation Research Part C: Emerging Technologies*, 98, pp. 152–166.

van Cranenburgh, S., S. Wang, A. Vij, F. Pereira, J. Walker (2022) Choice modelling in the age of machine learning - Discussion paper, *Journal of Choice Modelling*, 42, p. 100340.

van der Pol, M., G. Currie, S. Kromm, M. Ryan (2014) Specification of the Utility Function in Discrete Choice Experiments, *Value in Health*, 17(2), pp. 297–301.

Wang, S., B. Mo, S. Hess, J. Zhao (2021a) Comparing hundreds of machine learning classifiers and discrete choice models in predicting travel behavior: An empirical benchmark, *arXiv preprint arXiv:2102.01130*.

Wang, S., B. Mo, J. Zhao (2020a) Deep neural networks for choice analysis: Architecture design with alternative-specific utility functions, *Transportation Research Part C: Emerging Technologies*, 112, pp. 234–251.

Wang, S., Q. Wang, N. Bailey, J. Zhao (2021b) Deep neural networks for choice analysis: A statistical learning theory perspective, *Transportation Research Part B: Methodological*, 148, pp. 60–81.

Wang, S., Q. Wang, J. Zhao (2020b) Deep neural networks for choice analysis: Extracting complete economic information for interpretation, *Transportation Research Part C: Emerging Technologies*, 118, p. 102701.

Wong, M., B. Farooq (2021) ResLogit: A residual neural network logit model for data-driven choice modelling, *Transportation Research Part C: Emerging Technologies*, 126, p. 103050.

# Chapter 6

# A new software package to estimate non-parametric models to compute the value of travel time distribution from binary choice experiments

- Hernández, J. I., & van Cranenburgh, S. (2023). NP4VTT: A new software for estimating the value of travel time with nonparametric models. *Journal of Choice Modelling*, 48, 100427.

Two-attribute-two-alternative stated choice experiments are widely used to infer the Value-of-Travel-Time (VTT) distribution. Two-attribute-two-alternative stated choice experiments have the advantage that their data can be analysed using nonparametric models, which allow for the inference of the VTT distribution without having to impose assumptions on its shape. However, a software package that enables researchers to estimate nonparametric models promptly is currently lacking. As a result, nonparametric models are underused. This paper aims to fill this software void. It presents NP4VTT, a Python package that enables researchers to estimate and compare nonparametric models in a fast and convenient way. It comprises five nonparametric models for estimating the VTT distribution from data coming from two-attribute-two-alternative stated choice experiments. We illustrate the use of NP4VTT by applying it to the Norwegian 2009 VTT data. We hope this software package will help researchers studying the VTT make more informed decisions concerning the shape of the VTT distribu-

tion and encourages the use and development of nonparametric models for choice behaviour analyses.

## 6.1. Introduction

Accurate inference of the Value-of-Travel-Time (VTT) is of great importance for policy appraisal (Small, 2012). In current practice, inference of the VTT and its distribution is commonly based on data from Stated Choice (SC) experiments. In the VTT literature, there are two streams of designing VTT SC experiments. The first stream propagates the use of choice tasks with three (or more) attributes. Thus, besides travel cost and travel time, the choice task comprises a third (and possibly a fourth) attribute, such as congestion level or reliability. Proponents of this stream argue that choice tasks with three or more attributes contribute lead behaviourally realistic choice tasks (Hess et al., 2020). The second stream propagates the use of choice tasks with only two attributes: travel cost and travel time. Proponents of this stream argue that two-attribute-two-alternative designs enable respondents to trade off travel cost and travel time in a clean and crisp way, thereby yielding more reliable VTT estimates than one would obtain from an experiment that is 'polluted' by context effects or interactions arising from the use of more attributes. This stream has a strong base in Western and Northern European countries, including the United Kingdom, the Netherlands, Denmark, and Sweden (Fosgerau et al., 2007; Kouwenhoven et al., 2014).

From a modelling perspective, two-attribute-two-alternative experimental designs can be analysed with nonparametric and semi-nonparametric models. In contrast to their parametric counterpart — which is dominant in the analysis of data from three or more attribute SC experiments — (semi-) nonparametric models do not impose prior assumptions on the shape of the VTT distribution. Hence, in nonparametric models, the analyst does not need to predefine the distribution of the VTT (e.g., lognormal, normal, log uniform, etc.) before estimation. As a result of their agnosticism to the shape of the VTT distribution, nonparametric models can more accurately recover the VTT distribution. This desirable feature of nonparametric models explains their increasing use in recent VTT studies (e.g., Börjesson & Eliasson, 2014). In particular, nonparametric models are used to help the analyst decide which parametric distribution to use for the estimation of a (parametric) Random Utility Maximisation (RUM) model, which will be used for policy appraisal.

However, although the merits of nonparametric models are increasingly acknowledged in VTT studies, the burden of using them is high. Whereas for their parametric counterparts — i.e. discrete choice models — numerous software packages and libraries are nowadays available, such as PandasBiogeme (Bierlaire, 2018), Apollo (Hess & Palma, 2019), and Stata (StataCorp, 2005), for nonparametric models this is not the case. Analysts that wish to estimate nonparametric models often need to code their models from scratch. This hampers the more widespread use of nonparametric models for VTT analysis.

This paper fills this software void. It presents a new software package called NP4VTT. NP4VTT is a Python package that enables analysts to estimate a range of (semi-)nonparametric models to recover the VTT distribution. More specifically, NP4VTT comprises five nonparametric models that have been used in recent VTT studies, namely: Local constant model[1] (Fosgerau, 2006, 2007), Local logit (Fosgerau, 2007), Rouwendal model (Rouwendal et al., 2010), Artificial Neural Network (ANN) based VTT model (van Cranenburgh & Kouwenhoven, 2021), and a Logistic Regression based VTT model (van Cranenburgh & Kouwenhoven, 2021). For completeness and serving as a benchmark, we also added a parametric Random Valuation (RV) model to this package.

While the models included in NP4VTT vary significantly in terms of how they operate, they have in common that they all work on two-attribute-two-alternative choice data. The strength of NP4VTT is that it enables analysts to apply all these nonparametric models in an easy and convenient way and enables comparison of their results, considering their respective strengths and weaknesses. Furthermore, NP4VTT is not confined to VTT applications, it can more generally be used to uncover the distribution of the substitution rate between any two attributes of a two-attribute-two-alternative choice task. Furthermore, to illustrate the use of NP4VTT inside and outside VTT applications, we provide two examples in a Git repository[2]. The first example illustrates how each nonparametric model included in NP4VTT is used to uncover the VTT distribution from the Norwegian 2009 VTT study (Ramjerdi et al., 2010). The second example shows how NP4VTT can be used to calculate the distribution of the "willingness to pay for reducing environmental damage" in a hypothetical case study: we designed a hypothetical two-attribute-two-alternative choice experiment based on the Contingent Valuation (CV) survey to assess the damages of the Exxon Valdez oil spill in 1989 (Carson et al., 2003)[3].

The remaining part of this paper is organised as follows. Section 6.2 introduces the data format and describes the five nonparametric models implemented in this software package. Descriptions of the nonparametric models are kept short. They are meant

[1] Regarding, terminology: nonparametric models are also referred to as "estimators". However, throughout this paper, we use the word "models" for reasons of coherence instead of mixing "estimator", "model" and "method". Likewise, for reasons of coherence, we stick with the word "estimation", despite that for some "models", "training" is more appropriate. Finally, we refer to all models as "nonparametric", despite that a further distinction can be made into "semi" and "full" nonparametric models.

[2] The URL of the Git repository is https://github.com/ighdez/py-np4vtt

[3] We appreciate the suggestion of one anonymous reviewer to elaborate more on applications of NP4VTT outside the transportation field.

to convey the general idea of a model, not to provide in-depth expositions. Readers interested in the technical details are referred to the original works where the nonparametric models are introduced. Section 6.3 presents the NP4VTT software. Section 6.4 illustrates the use of the package by applying it to the data from the Norwegian 2009 VTT study. Section 6.5 provides a brief discussion of future developments and the next steps.

## 6.2. Data format and nonparametric models

### 6.2.1. Data format

Table 6.1 shows the data structure and data types that NP4VTT requires. NP4VTT operates on data for two-attribute-two-alternative choice experiments. In this format, each choice observation comprises at least six input variables: the travel costs and travel times of the two alternatives, plus the choice. Additionally, each respondent must have a unique identifier, which is repeated for multiple choice situations answered by a same respondent.

| Variable | Unique identifier of the respondent | Choice | Cost alternative 1 | Time alternative 1 | Cost alternative 2 | Time alternative 2 |
|---|---|---|---|---|---|---|
| Data type | Integer | Integer | Float | Float | Float | Float |
| | 1 | 1 | 25 | 30 | 30 | 20 |
| | 1 | 1 | 25 | 40 | 20 | 45 |
| | 1 | 2 | 25 | 40 | 15 | 45 |
| | 2 | 1 | 15 | 45 | 20 | 30 |
| | 2 | 2 | 20 | 30 | 25 | 20 |
| | … | … | … | … | … | … |

*Table 6.1: Data structure required by NP4VTT*

NP4VTT distinguishes between cross-sectional, balanced and unbalanced panel data. Cross-sectional data is when each respondent answers exactly one choice situation. Balanced panel data is when all respondents answered two or more choice situations and all respondents answered the same number of choice situations. Hence, unbalanced panel data is when all respondents answered two or more choice situations, but the number of choice situations answered across respondents differs. NP4VTT automatically identifies whether the dataset contains cross-sectional, balanced or unbalanced panel data. In case the dataset does contain either cross-sectional or unbalanced panel data, only the Local constant, Local logit and RV models are enabled.

Dominant choice tasks are not permitted: each choice observation must have a *slow and cheap* alternative and a *fast and expensive* alternative. In the case of a dominant alternative, it must be removed. NP4VTT runs integrity checks before estimation and will prompt error messages if dominant alternatives are present in the data.

NP4VTT computes the Boundary-Value-of-Travel-Time (BVTT). The BVTT is the implicit price of time in a two-attribute-two-alternative choice task (Cameron & James, 1987). Often the BVTT is perceived as a valuation threshold (Ojeda-Cabral et al., 2016) , meaning that a respondent choosing the fast and expensive alternative reveals having a VTT above the BVTT; a respondent choosing the slow and cheap alternative reveals having a VTT below the BVTT. The formula to compute the BVTT is given in equation (6.1), where $t_1$ and $c_1$ denote the travel time and travel cost of the *slow and cheap* alternative and $t_2$ and $c_2$ denote the travel time and travel cost of the *fast and expensive* alternative.

$$BVTT = -\frac{c_1 - c_2}{t_1 - t_2} \tag{6.1}$$

### 6.2.2. Local constant model

The Local constant model is pioneered by Fosgerau (2006, 2007) for studying the VTT distribution from cross-sectional data. The Local constant model is an approach based on the regression model $y = f(BVTT) + \varepsilon$, in which the aim is to get an approximation of the function $f$ in a nonparametric fashion using a kernel density (Nadaraya-Watson) estimator.

We formalise the Local constant model used in this paper as follows. Let y be an indicator variable which equals one if a respondent chooses the slow and cheap alternative and zero otherwise. Define the conditional utility of each alternative as $\alpha_t t_i + \alpha_c c_i$, in which $\alpha_t$ and $\alpha_c$ are random parameters independent across choice tasks, while the subscript $i$ denotes the slow and cheap (1) and fast and expensive (2) alternatives. Then, the condition under which a respondent chooses the slow and cheap alternative is given in equation (6.2):

$$\alpha_t t_1 + \alpha_c c_1 > \alpha_t t_2 + \alpha_c c_2 \tag{6.2}$$

As a result, y can be rewritten as equation (6.3):

$$\begin{aligned} y &= 1\{\alpha_t t_1 + \alpha_c c_1 > \alpha_t t_2 + \alpha_c c_2\} \\ &= 1\left\{\frac{\alpha_t}{\alpha_c} < -\frac{c_1 - c_2}{t_1 - t_2}\right\} \\ &= 1\left\{\frac{\alpha_t}{\alpha_c} < BVTT\right\} \end{aligned} \tag{6.3}$$

equation (6.3) suggests that we observe a respondent choosing the slow and cheap alternative, i.e., $y = 1$), when his/her unobserved VTT (represented by $w = \alpha_t/\alpha_c$), is lower than the BVTT. Conversely, a respondent chooses the fast and expensive alternative when his/her VTT is higher than the BVTT.

Furthermore, notice that $P(y = 1) = P(w < BVTT) = F_w(BVTT)$, where $F_w(\cdot)$ is a Cumulative Distribution Function (CDF) of $w$. Therefore, we can define the model as in equation (6.4):

$$y = F_w(BVTT) + \eta \tag{6.4}$$

The objective is to get an estimate of $F_w(BVTT)$. Fosgerau (2006) proposes to approximate $F_w$ in a nonparametric way, using the Nadaraya-Watson estimator. Following Fosgerau (2007), given a point $x_0$ defined in the support of the VTT, the estimate of $F_w$ around $x_0$ is given by equation (6.5):

$$\hat{F}_w(X_0) = \frac{\sum_( i \leq N)K\left(\frac{BVTT_i - x_0}{h}\right)y_i}{\sum_( i \leq N)K\left(\frac{BVTT_i - x_0}{h}\right)}, \tag{6.5}$$

where $K(\cdot)$ is a kernel function and $h$ is a user-defined smoothing bandwidth parameter. The idea behind of equation (6.5) is to have a weighted average of the probability on each point of the VTT support. $K(\cdot)$ weights the distance between each point of the BVTT and $x_0$. BVTTs close to the support point $x_0$, $K(\cdot)$ weight comparatively heavily, and vice versa, BVTTs that are far from the support point weight comparatively lightly. The bandwidth parameter h controls the smoothness of $\hat{F}_w$. A large value of $h$ will result in a smooth estimate of the CDF, at the cost of underfitting; a small value of $h$ will result in a comparatively more erratic estimate of the CDF.

NP4VTT implements a standard normal (gaussian) kernel function for the Local constant model. The construction of the VTT distribution is done by estimating $\hat{F}_w$ for the mid points of a user-defined grid of VTT support points. The user controls the minimum and maximum values of the grid, as well as the number of support points in between.

### 6.2.3. Local logit

The Local logit model is a model for cross-sectional data first proposed by Fan et al. (1995), and later pioneered by Fosgerau (2007) in the VTT literature. As the name suggests, the Local logit model involves estimation of a series of "local" logits. In this context, "local" refers to the notion that the logit models are estimated on a subset of the data. In the VTT context, this means logit models are estimated on subsets that are created based on the BVTT. For instance, a first subset may contain all choice observations for which holds 0 €/hr $<$ BVTT $<$ 10 €/hr, a second subset may contain all choice observations for which holds 10 €/hr $\leq$ BVTT $<$ 20 €/hr, and so on. The centre of the first bin then sits at 5 €/hr, of the second bin at 15 €/hr, and so. On each of these subsets, a Logistic regression (hence logit) is estimated.

This logit model comprises two coefficients: an intercept for the mid-point and a linear term capturing the distance from the centre of the bin. In the log-likelihood function, the weight of each data point is computed using a kernel function. The further away (in terms of BVTT space) a data point is from the centre of the bin, the lower the impact of the prediction of that data point in the log-likelihood function. After all, a data point further away from the bin centre contains less information on the VTT at the bin. In NP4VTT, we use a triangular kernel function. This means that the contribution of an observation to the log-likelihood function decreases linearly with the distance from the centre of the bin.

### 6.2.4. Rouwendal model

Rouwendal et al. (2010) propose a nonparametric model to estimate the VTT and the Values-of-Statistical-Life (VOSL) from SC data consisting of three attributes: cost, time and safety. NP4VTT implements a recent adaption of this model for balanced panel data, proposed by (van Cranenburgh & Kouwenhoven, 2021, , appendix B) to estimate the VTT from two-alternative-two-attribute data. This nonparametric model is built on two assumptions. First, each respondent has a VTT that is constant across the presented choice tasks. Second, each choice that is made by a respondent is subject to a given probability $q$ of being inconsistent with the respondent's underlying VTT. Therefore, the probability of observing a series of T choices $Y_n = \{y_{n1}, y_{n2}, \dots, y_{nT}\}$ for a respondent $n$ is given by equation (6.6):

$$P(Y_n|\nu) = q^{\tau_n}(1-q)^{T-\tau_n}, \tag{6.6}$$

where $\tau_n(\nu)$ denotes the number of choices that is consistent with when the respondent's VTT equals $\nu$. The unconditional probability for observing a respondent $n$ to making the series of choices $Y_n$ is computed by integrating over $\nu$. Practically, we discretise the VTT space in $B$ evenly spaced bins, and we integrate numerically, as shown in equation (6.7):

$$P(Y_n) = \sum_{b=1}^{B} f(\nu_b) q^{\tau_n} (1-q)^{T-\tau_n}, \tag{6.7}$$

where $f(\tau_b)$ denotes the probability density function of the VTT at $\tau_b$. Estimating this model involves estimating one density parameter per bin, plus the probability for an inconsistent choice $q$. An estimate of the CDF is obtained by computing the cumulative sum of the density parameters, previously converted to probability space.

### 6.2.5. ANN-based VTT model

van Cranenburgh & Kouwenhoven (2021) propose an ANN-based approach to uncover the VTT distribution (henceforth referred to as ANN-based VTT model), which is implemented in NP4VTT for balanced panel data. This ANN-based VTT model builds on the notion that the VTT can be inferred through finding the BVTT that makes the respondent indifferent between the slow and cheap and fast and expensive alternatives. If a respondent is indifferent between the slow and cheap and the fast and expensive alternative, then the BVTT must equal the VTT of the respondent. This approach takes the following steps to recover the BVTT that makes a respondent indifferent.

First, the data are reorganised. For each respondent, a randomly selected choice observation is singled out as the dependent choice task. The remaining $T-1$ choice observations, including the associated choices, are used as independent variables. Second, an ANN is trained to predict the choice in the dependent choice task. In other words, the ANN is trained to learn a mapping $g(\cdot)$, based on the choices and BVTTs of the $T-1$ choice tasks and the BVTT of the dependent choice task. Equation (6.8) details such mapping, where $BVTT_{-r}$ and $y_{-r}$ denote respectively the BVTTs and choices of the $T-1$ choice observations that serve as independent variables, and $BVTT_r$ denote the BVTT of the dependent choice task:

$$P_r = g(BVTT_{-r}, y_{-r}, BVTT_r). \tag{6.8}$$

Third, after training the ANN, it is used to simulate the effect of $BVTT_r$ on the choice probablility. The VTT for each respondent is recovered by finding the $BVTT_r$ that yields $P_r = 0.5$. Finally, the recovered VTTs of all respondents in the sample are taken together to produce the VTT distribution.

## 6.2.6. Logistic regression-based VTT model

This model is proposed by van Cranenburgh & Kouwenhoven (2021) as a variation of their ANN-based VTT model, and implemented in NP4VTT for balanced panel data. It is motivated by the observation that the ANN-based VTT model cannot be readily interpreted because of the opaqueness of the ANN. The main idea of this model is to create a transparent counterpart, by replacing the ANN with a Logistic regression. In other words, in this model, the mapping $g(\cdot)$ of equation (6.7) is learned by a simple linear regressor. Doing so will decrease the model's flexibility, leading to deteriorated model performance (e.g., lower log-likelihood), but will increase the interpretability of the model. Because of its simple linear structure, van Cranenburgh & Kouwenhoven (2021) show that the regression problem reduces to the form described in equations (6.9) and (6.10):

$$P_r = \frac{1}{1+\exp(-V_n)},, \tag{6.9}$$

$$V_n = \delta + \beta_{y,BVTT} \sum_{t=1}^{T-1} y_{nt}^{FE} \cdot BVTT_{nt} + \beta_{BVTT} \cdot BVTT_{nr}. \tag{6.10}$$

In equation (6.10), $\delta$ is an intercept and $y_{nt}^{FE}$ is an indicator equal to 1 if respondent $n$ chooses the fast and expensive alternative in choice task $t$. $\beta_{y,BVTT}$ is interpretable as the marginal effect of a choice for the fast and expensive alternative in a choice task with $BVTT_{nt}$. $\beta_{y,BVTT}$ is expected to be positive. $\beta_{BVTT}$ captures the marginal effect of the $BVTT_r$ of the dependent choice task. As a higher $BVTT_r$ lowers the probability of choosing the fast and expensive alternative. Hence, we expect $\beta_{BVTT}$ to be negative.

## 6.2.7. Random Valuation

While the Random Valuation (RV) model is not a nonparametric model, it is nonetheless added to this software package to compute a benchmark VTT. The RV model is a parametric model that allows to compute the mean VTT in cross-sectional two-alternative-two-attribute data, first proposed by Cameron & James (1987) and described in Ojeda-Cabral et al. (2016) and Ojeda-Cabral & Chorus (2016). The RV model assumes that the choice task is a "time market" in which the price is the BVTT of equation (6.1). Then, a respondent chooses the slow and cheap alternative if their VTT is lower than the BVTT of the choice task, otherwise, they choose the fast and expensive alternative. Adding an additive Extreme Value stochastic term, the choice probabilities of the RV model are represented by equation (6.11):

$$y = 1\{VTT < BVTT + \varepsilon\} \tag{6.11}$$

Following Ojeda-Cabral & Chorus (2016), the utility of each alternative in the RV model is parametrized as in equations (6.12) and (6.13):

$$U_1 = \mu \cdot BVTT + \varepsilon_1 \tag{6.12}$$

$$U_1 = \mu \cdot VTT + \varepsilon_2 \tag{6.13}$$

where $\mu$ is a scale parameter. Since $\varepsilon_1$ and $\varepsilon_2$ are Extreme Value stochastic terms, the RV choice probabilities collapse to a binary logit model. The mean VTT is directly obtained from the estimation results since it enters to the logit model as a parameter to be estimated.

## 6.3. The NP4VTT software

NP4VTT is provided as a Python 3 package. Users can install NP4VTT from the Python Package Index (PyPi) using the regular procedure to instal packages (i.e., `python -m pip install py-np4vtt` in the command line interface). NP4VTT requires Python version 3.8 or higher and depends on the following packages:

- Pandas version 1.3.1 or higher,
- SciPy version 1.7.1 or higher,
- Scikit-learn version 1.0.2 or higher,
- Matplotlib version 3.5.1 or higher, and
- Numdifftools version 0.9.40 or higher.

NP4VTT is developed as open-source software. This means that users have complete access to the source code of NP4VTT. They can download the source code and suggest changes and additions. The source code of NP4VTT is stored in a Git repository and can be accessed by anyone. While we, as maintainers of NP4VTT, aim to provide a reliable tool for research and education, this software comes with no warranty. Thus, neither the authors nor the Delft University of Technology are liable for any consequences from the use of the software, waiving that responsibility to the users.

Figure 6.1 details the structure of the NP4VTT package. The main module (pynp4vtt) contains six submodules that contain each model, plus three submodules dedicated to arranging variables (`data_format`), creating the necessary arrays from a Pandas data frame (`data_import`) and utilitarian functions (utils). The submodules that contain each model are:

1. Local constant (see Section 6.2.2): `model_lconstant`,
2. Local logit (see Section 6.2.3): `model_loclogit`,
3. Rouwendal model (see Section 6.2.4): `model_rouwendal`,
4. ANN-based VTT model (see Section 6.2.5): `model_ann`,
5. Logistic regression (see Section 6.2.6): `model_logistic`, and
6. Random Valuation model (see Section 6.2.7): `model_rv`.

Each model submodule contains a configuration class and a model class. The purpose of a configuration class is to receive the specific parameters of the correspondent model, perform integrity checks (e.g., the number of support points of the Local logit model must be a positive integer), and pass that information to the correspondent model class. Model classes contain the routines and methods to prepare specific arrays and estimate their correspondent model (`run`). After estimating a model (i.e., using the method `run`), the model class stores specific output to compute the VTT distribution based on the configuration parameters and the data. Table 6.2 describes the configuration parameters and outcomes of each model. A complete description of the classes and functions included in NP4VTT is provided in the reference manual as supplementary material to this paper.

*Table 6.2: Input of configuration class and model class outputs*

| Model and model submodule | Configuration class and inputs | Description of configuration inputs | Model class and output |
|---|---|---|---|
| Local constant:<br>model_lconstant | ConfigLConstant:<br>minimum [float]<br>maximum [float]<br>supportPoints [integer]<br>kernelWidth [float] | The minimum, maximum and number of support points of the VTT grid in which the supportPoints-1] estimates of the CDF will be estimated, plus the kernel width of the nonparametric (Nadaraya-Watson) estimator. | ModelLConstant:<br><br>At initialisation:<br>vtt_grid: VTT grid of points<br>vtt_mid: mid-points of the VTT grid<br><br>After estimation:<br>p: set of estimates of the CDF evaluated at each mid point of the VTT grid. |

| | | | |
|---|---|---|---|
| | | | vtt: set of VTTs of size NP, derived from p<br>est_time: the estimation time in seconds. |
| Local logit:<br>model_loclogit | ConfigLocLogit:<br>minimum [float]<br>maximum [float]<br>supportPoints [integer] | The minimum, maximum and number of support points of the VTT grid in which the [supportPoints-1] estimates of the CDF will be estimated. | ModelLocLogit:<br><br>At initialisation:<br>vtt_grid: VTT grid of points<br>vtt_mid: mid-points of the VTT grid<br><br>After estimation:<br>p: set of estimates of the CDF evaluated at each interval of the VTT grid.<br>vtt: set of VTTs of size NP, derived from p<br>ll: log-likelihood value at the optimum<br>est_time: the estimation time in seconds. |
| Rouwendal model:<br>model_rouwendal | ConfigRouwendal:<br><br>minimum [float]<br>maximum [float]<br>supportPoints [integer]<br>startQ [float] | The minimum, maximum and number of support points of the VTT grid in which the [supportPoints -1] estimates of the CDF will be estimated, in addition to the starting value (startQ) of the probability of consistent choice | ModelRouwendal:<br><br>At initialisation:<br>vtt_grid: VTT grid of points<br>vtt_mid: mid-points of the VTT grid<br><br>After estimation:<br>q_est: raw estimate of the probability of consistent choice<br>q_se: Std. Error of q_est<br>q_prob: (logit) probability of q_est<br>x: density parameter estimates<br>se: standard errors of x<br>p: set of estimates of the CDF evaluated at each interval of the VTT grid<br>vtt: set of VTTs of size NP, derived from p<br>init_ll: value of log-likelihood function at starting values<br>ll: value of log-likelihood function at the optimum<br>exitflag: convergence result (if zero, estimation was successful)<br>est_time: the estimation time in seconds. |
| ANN-based model:<br>model_ann | ConfigANN:<br><br>hiddenLayerNodes [list]<br>trainingRepeats [int] | The topology of the neural network (hiddenLayerNodes), the number of training repeats and the number of shuffles per repeat | ModelANN:<br><br>At initialisation: |

| | | | |
|---|---|---|---|
| | shufflesPerRepeat [int]<br>seed [hashable] | | X_train, X_test, X_full: input data (BVTT and T-1 choices) for training, testing and full sample<br>y_train, y_test, y_full: output data (T choice) for training, testing and full sample<br><br>After estimation:<br>ll_list: array of log-likelihoods at convergence per training repeat<br>r2_list: array of rho-square values per training repeat<br>vtt_list: array with VTTs per respondent per training repeat<br>est_time: the estimation time in seconds.<br>avg_time: average estimation time per repetition. |
| Logistic regression: model_logistic | ConfigLogistic:<br><br>startIntercept [float]<br>startParameter [float]<br>startScale [float]<br>maxIterations [int]<br>seed [hashable] | The starting values of the parameters to be estimated in the Logistic regression model, and the number of iterations of the optimization routine. | ModelLogistic:<br><br>At initialisation:<br>(none)<br><br>After estimation:<br>x: model estimates (scale, intercept and parameter<br>se: standard errors<br>vtt: VTT per respondent, based on x.<br>init_ll: log-likelihood at starting values<br>ll: log-likelihood value at convergence<br>exitflag: convergence result (if zero, estimation was successful)<br>est_time: the estimation time in seconds. |
| Random Valuation: model_rv | ConfigRV:<br><br>mleScale [float]<br>maxIterations [int]<br>mleVTT [float] | The starting values of the parameters of the RV model, and the number of iterations of the optimization routine. | ModelRV:<br><br>At initialisation:<br>(None)<br><br>After estimation:<br>x: model estimates (scale and VTT)<br>se: standard errors<br>init_ll: log-likelihood at starting values<br>ll: log-likelihood value at convergence<br>exitflag: convergence result (if zero, estimation was successful) |

| | |
|---|---|
| | est_time: the estimation time in seconds. |

Irrespective of the model, the procedure to use NP4VTT consists of three stages. Figure 6.2 depicts these three stages. In the first stage, the user creates the model arrays using the method `make_modelarrays`, which takes a Pandas data frame that contains the dataset and a dictionary that maps each necessary array with the variable names in the dataset as inputs. In the second stage, the user provides configuration parameters through a configuration class and creates the model object. For example, suppose the user wants to estimate a Local logit model. In that case, the user must provide the minimum and maximum VTT values, and the number of support points through the `ConfigLocLogit` configuration class. Then, the user creates a model object using the `ModelLocLogit` model class, the configuration class and the arrays object created with `make_modelarrays`. Some methods store specific arrays once the model object is created that can be accessed by the user. For instance, the `ModelLocLogit` object contains the VTT grid array that can be accessed using the "dot" notation (i.e., `ModelLocLogit.vtt_grid`) and the mid points of the VTT grid (i.e., `ModelLocLogit.vtt_mid`). In the third stage, the user estimates the model using the method `run`. The output of the method run corresponds to the specific outcomes of the estimated model. For example, in a Local logit model, the estimation routine stores the set of estimates of the CDF points at each interval of the VTT grid, the estimated VTT for each respondent of the sample and the final log-likelihood value.

## 6.4. Demonstration of NP4VTT

### 6.4.1. Demonstration data

We demonstrate the use of NP4VTT using the Norway 2009 VTT data (Ramjerdi et al., 2010). This is a balanced panel dataset that contains 5,832 respondents, each having made $T = 9$ choice tasks. Each respondent was presented with a two-alternative choice task, characterised by two attributes: travel time in minutes and travel cost in Norwegian crowns.

Table 6.3 shows one example of a choice tack from this SC experiment. To ease interpretation, we converted the currency from Norwegian krone to Euro, using an exchange rate of 9 NOK = 1 EUR. The minimum and maximum values of the BVTT are 0.67 EUR/hr. and 113.56 EUR/hr., respectively. The mean chosen BVTT is 10.30 EUR/hr.

These data comprise six variables:

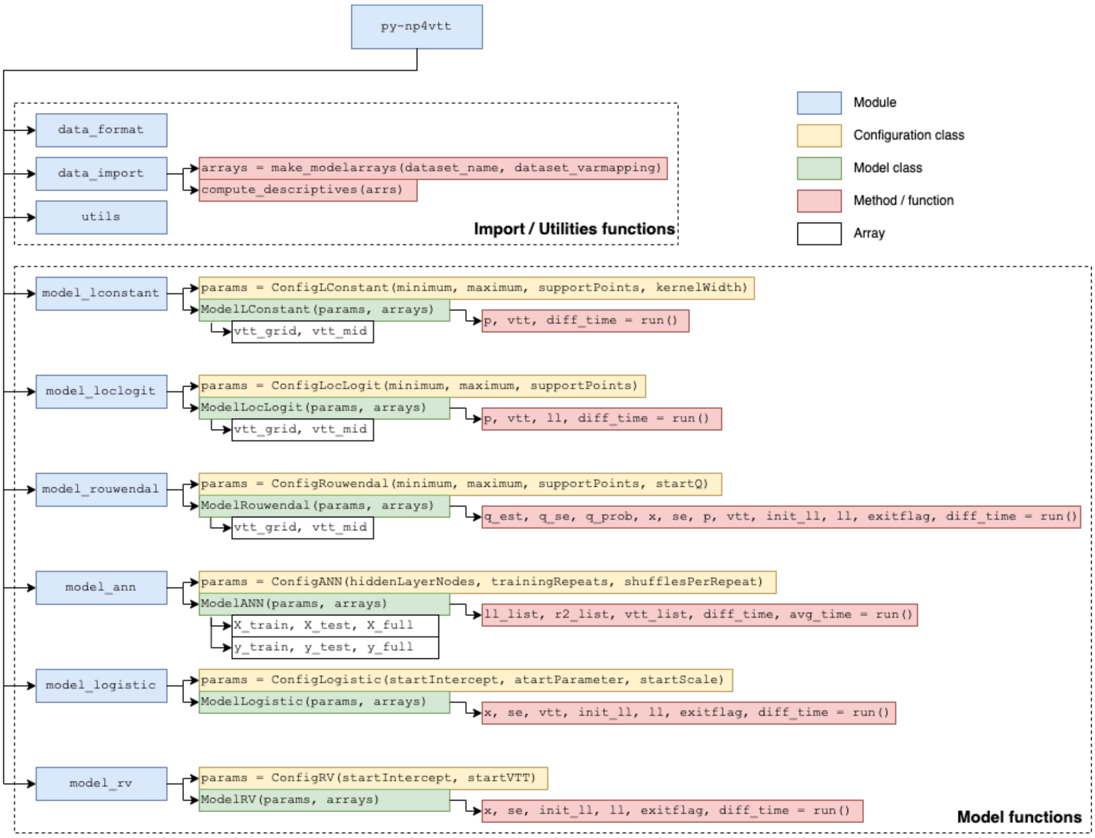

py-np4vtt
Module
Configuration class
Model class
Method / function
Array
data_format
data_import
arrays = make_modelarrays(dataset_name, dataset_varmapping)
compute_descriptives(arrs)
utils
Import / Utilities functions
model_lconstant
params = ConfigLConstant(minimum, maximum, supportPoints, kernelWidth)
ModelLConstant(params, arrays)
p, vtt, diff_time = run()
vtt_grid, vtt_mid
model_loclogit
params = ConfigLocLogit(minimum, maximum, supportPoints)
ModelLocLogit(params, arrays)
p, vtt, ll, diff_time = run()
vtt_grid, vtt_mid
model_rouwendal
params = ConfigRouwendal(minimum, maximum, supportPoints, startQ)
ModelRouwendal(params, arrays)
q_est, q_se, q_prob, x, se, p, vtt, init_ll, ll, exitflag, diff_time = run()
vtt_grid, vtt_mid
model_ann
params = ConfigANN(hiddenLayerNodes, trainingRepeats, shufflesPerRepeat)
ModelANN(params, arrays)
ll_list, r2_list, vtt_list, diff_time, avg_time = run()
X_train, X_test, X_full
y_train, y_test, y_full
model_logistic
params = ConfigLogistic(startIntercept, atartParameter, startScale)
ModelLogistic(params, arrays)
x, se, vtt, init_ll, ll, exitflag, diff_time = run()
model_rv
params = ConfigRV(startIntercept, startVTT)
ModelRV(params, arrays)
x, se, init_ll, ll, exitflag, diff_time = run()
Model functions

*Figure 6.1: Structure of NP4VTT*

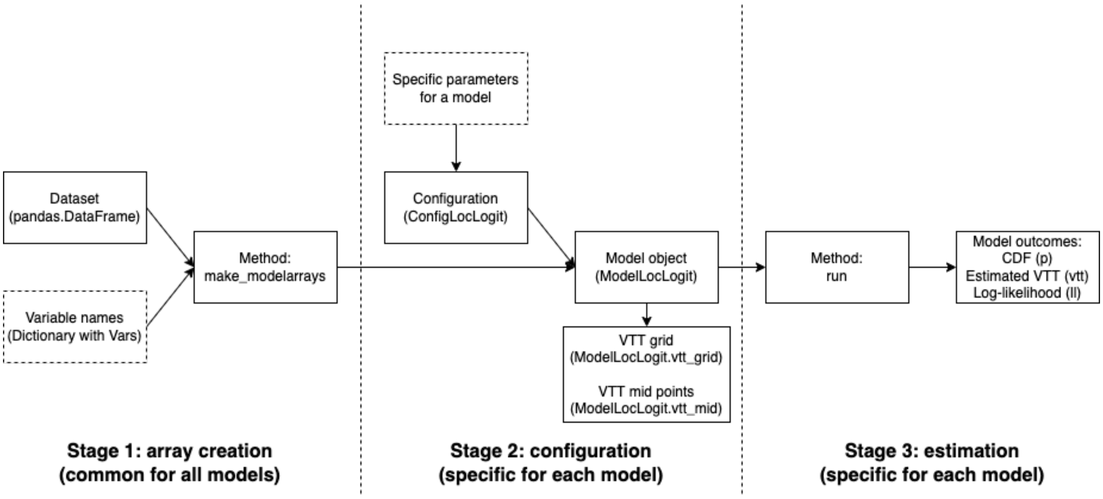


*Figure 6.2: Usage procedure of NP4VTT (using a Local logit as an example*

| | Alternative 1 | Alternative 2 |
|---|---|---|
| Travel time (minutes) | 15 | 12 |
| Travel cost (Euro)* | 8 | 10 |
| **YOUR CHOICE** | **O** | **O** |

*Table 6.3: Example choice task of the Norwegian 2009 VTT data*

- `RespID`: A unique identifier that maps every single respondent of the dataset,
- `Chosen`: The respondent's chosen alternative. It can take values 1 or 2,
- `CostL`: Travel costs of alternative 1 [NOK],
- `CostR`: Travel costs of alternative 2 [NOK],
- `TimeL`: Travel time of alternative 1 [minutes], and
- `TimeR`: Travel time of alternative 2 [minutes].

### 6.4.2. Loading data, array creation and descriptive statistics

First, the user must load the dataset and generate the necessary arrays. Box 6.1 shows a code snippet that exemplifies this process in NP4VTT.

After the user loads the required modules (i.e., Pandas and the sub-modules of NP4VTT for array creation), the data are read from a CSV-file and stored as a Pandas data frame named `df`.

To map the required variables with the names that appear in the dataset, the user must a dictionary object — named in this example as `columnarrays` — with the variable names contained in the data, for the following keys:

- `Vars.Id`: The unique identifier variable,
- `Vars.ChosenAlt`: Choice indicator,
- `Vars.Cost1` and `Vars.Cost2`: Cost variables, and
- `Vars.Time1` and `ars.Time2`: Time variables.

Finally, the user creates the necessary arrays using the `make_modelarrays` function. The arrays are stored in the `model_arrays` object. The function `make_modelarrays` takes the Pandas data frame and the dictionary that maps the variables as inputs. This function outputs a list with the following elements:

```
# Load modules
import pandas as pd
from py_np4vtt.data_format import Vars
from py_np4vtt.data_import import make_modelarrays, compute_descriptives

# Load dataset
df = pd.read_table('../data/Norway2009VTT_demodata.txt')

# Convert travel cost to EUR
NOK2euro_exchange_rate = 9
df[['CostL','CostR']] = df[['CostL','CostR']] .div(NOK2euro_exchange_rate)

# Convert travel time to hours
df[['TimeL','TimeR']] = df[['TimeL','TimeR']] .div(60)

# Create dictionary that maps the required variables with the variables of the dataset
columnarrays = {
    Vars.Id: 'RespID',
    Vars.ChosenAlt: 'Chosen',
    Vars.Cost1: 'CostL',
    Vars.Cost2: 'CostR',
    Vars.Time1: 'TimeL',
    Vars.Time2: 'TimeR',
}

# Create the necessary arrays
model_arrays = make_modelarrays(df, columnarrays)

# Compute descriptives
descriptives = compute_descriptives(model_arrays)
```

*Box 6.1: Code to load the data, create arrays and compute descriptive statistics in NP4VTT*

- `NP`: An integer describing the number of respondents,
- `T`: An integer describing the number of choice tasks per respondent,
- `ID`: A 1-dimensional NumPy array of dimension NP that contains the unique identifiers of each respondent,
- `BVTT`: A 2-dimensional NumPy array of dimension (NP × T), that contains the BVTT computed using equation (6.1) from the observed travel costs and travel time, for each choice task and each respondent,
- `Choice`: A 2-dimensional NumPy array of dimension (NP×T), in which each cell takes value 1 if the respondent chose the expensive and fast alternative, and zero otherwise, and
- `Accepts`: A 1-dimensional NumPy array of dimension NP that contains the number of times a respondent accepted the fast and expensive alternative.

Finally, the function `compute_descriptives` takes `model_arrays` as input and produces a set of descriptive statistics:

- The number of respondents ($NP$) and number of choice tasks ($T$),
- Number of non-traders of the expensive-but-fast and cheap-but-slow alternatives,
- Mean of the chosen BVTT, and
- Minimum and maximum values of the BVTT in the data.

## 6.4.3. Estimation

After the necessary arrays are created, we estimate the nonparametric models. Each model included in NP4VTT takes the arrays created with the function `model_arrays` plus one object that stores the specific configuration parameters of the model. The model is initialised with this information. Estimation produces a set of arrays that allow the user to describe the VTT distribution, as in Table 6.2.

In the following subsections, we demonstrate NP4VTT by showing how to invoke the following four models: Local constant, Local logit, Rouwendal and ANN-based VTT.

### Estimating a Local constant model

Estimating a Local constant model involves configuring the specific parameters, creating the model object based on the specific configuration and arrays and executing the estimation routine. Box 6.2 details the code to configure and estimate the Local constant model. In this example, we define a VTT grid from 0 to 100 with 21 support points and a kernel width of 2 euros. The Local constant model will get an estimate of the CDF at each mid point of the VTT grid.

```
# Configure a Local constant model
from py_np4vtt.model_lconstant import ModelLConstant, ConfigLConstant
config = ConfigLConstant(
    minimum=0,
    maximum=100,
    supportPoints=21,
    kernelWidth = 2)

# Create Local constant model object
lc = ModelLConstant(config, model_arrays)

# The VTT grid and mid points can be accessed using the \dot' notation
vtt_grid = lc.vtt_grid
vtt_mid = lc.vtt_mid

# Estimate the Local constant model
p, vtt, est_time = lc.run()
```

*Box 6.2: Configuration and estimation of a Local constant model*

The object `ConfigLConstant` performs integrity checks to avoid invalid entries for the configuration parameters. Then, the user creates the Local constant object using the object `ModelLConstant` that receives the configuration object and the necessary arrays. The model is stored in the object called `lc`. The Local constant model object creates an array called `vtt_grid`, corresponding to the VTT grid as specified in the configuration parameters, as well as its corresponding mid points (`vtt_mid`). The VTT mid points are computed to allow the user to describe the VTT distribution in plots as the Local constant model computes the probabilities pointwise, thus complicating the interpretation of histograms. The VTT grid is stored in the Local constant model object and can be accessed using the “dot” notation (i.e., `lc.vtt_grid`).

To start the estimation routine, the user executes the method “run” (i.e., `lc.run()`). After completion, the estimation routine returns the following objects:

- `p`: set of estimates of the CDF evaluated at the mid point of the VTT grid,
- `vtt`: set of VTTs of size NP, derived from p, and

- `est_time`: the estimation time in seconds.

The user can use the model outputs to describe the VTT distribution. In Section 6.4.4, we demonstrate a visualisation of the VTT distribution.

### Estimating a Local logit model

As with all models of NP4VTT, estimating a Local logit model involves configuring the specific parameters, creating the model object, and executing the estimation routine. Box 6.3 details the code to configure and estimate a Local logit model. In this example, we define a VTT grid between 0 and 100, with 21 support points. The Local logit model will estimate the CDF on equally sized intervals with a length of 5.

```
from py_np4vtt.model_loclogit import ModelLocLogit, ConfigLocLogit
config = ConfigLocLogit(
    minimum=0,
    maximum=100,
    supportPoints=21)

# Create the Local logit model object
loclogit = ModelLocLogit(config, model_arrays)

# The created VTT grid and midpoints can be accessed using the ’dot’ notation
vtt_grid = lc.vtt_grid
vtt_mid = loclogit.vtt_mid

# Estimate the Local logit model
p, vtt, ll, est_time = loclogit.run()
```

*Box 6.3: Configuration and estimation of a Local logit model*

First, the user creates the configuration object `ConfigLocLogit` that verifies that the parameters of the VTT grid are valid. Then, the Local logit model object is created using the configuration and the arrays using the object `ModelLocLogit`. The Local logit model object creates the VTT grid (`vtt_grid`) as specified by the configuration parameters and its corresponding mid points (`vtt_mid`). The VTT mid points are computed to allow the user to describe the VTT distribution in plots, as the Local logit model estimates the CDF at intervals of the VTT grid. The VTT grid and mid points can be accessed using the “dot” notation.

To estimate the Local logit model, the user executes the “run” method. The outputs of the estimated Local logit model are:

- `p`: set of estimates of the CDF evaluated at each interval of the VTT grid,

- `vtt`: set of VTTs of size NP, derived from p,
- `ll`: the log-likelihood function at the optimum of the estimation process, and
- `est_time`: the estimation time in seconds.

The user can visually describe the VTT distribution using the estimates of the CDF and the VTT mid points or directly use the estimated VTT. Section 6.4.4. illustrates the VTT distribution for this specific example.

### Estimating the Rouwendal model

Box 6.4 details the code to configure and estimate the Rouwendal model. The configuration of the Rouwendal model parameters is done with the `ConfigRouwendal` object. We define a VTT grid from 0 to 100 with 21 support points. Hence, the Rouwendal model will estimate the CDF on equally sized intervals of 5 units. Additionally, we set the starting value of the probability of consistent choice in `q = 0.9`.

```
# Configure Rouwendal model
from py_np4vtt.model_rouwendal import ConfigRouwendal, ModelRouwendal
config = ConfigRouwendal(
    minimum= 0,
    maximum= 100,
    supportPoints= 21,
    startQ= 0.95)

# Create Rouwendal model object
rouwendal = ModelRouwendal(config, model_arrays)

# VTT grid and VTT mid points can be accessed using the ’dot’ notation
vtt_grid = rouwendal.vtt_grid
vtt_grid

vtt_mid = rouwendal.vtt_mid
vtt_mid

# Estimate the Rouwendal model
q_est, q_se, q_prob, x, se, p, vtt, init_ll, ll, exitflag, est_time = rouwendal.run()
```

*Box 6.4: Configuration and estimation of the Rouwendal model*

After creating the configuration object, the user creates the Rouwendal model object `ModelRouwendal` using the configuration object and the arrays. The model object creates the VTT grid (`vtt_grid`) as specified by the configuration parameters and its corresponding mid points (`vtt_mid`), as the Rouwendal model estimates the CDF at intervals of the VTT grid. The VTT grid and mid points can be accessed using the "dot" notation.

The estimation of the Rouwendal model is done using the run method. The estimated Rouwendal model returns the following outputs:

- `q_est`: raw estimate of the probability of consistent choice,
- `q_se`: standard error of `q_est`,
- `q_prob`: (logit) probability of `q_est`,
- `x`: density parameter estimates,
- `se`: standard errors of `x`,
- `p`: set of estimates of the CDF evaluated at each interval of the VTT grid,
- `vtt`: set of VTTs of size NP derived from p,
- `init_ll`: Value of log-likelihood function at starting values. Starting values of density parameters are equal to zero,
- `ll`: Value of log-likelihood function in the optimum,
- `exitflag`: Convergence result. If `exitflag=0`, the optimisation succeeded. Otherwise, check the configuration parameters, and
- `est_time`: the estimation time in seconds.

The user can describe the VTT distribution with the set of estimates of the CDF at each VTT mid point or directly plot a histogram of the estimated VTT. Section 6.4.4 illustrates the VTT distribution for the Rouwendal model compared with the other models employed in this example.

### Estimating an ANN-based model

The final example is the configuration and estimation process of an ANN-based model. Box 6.5 shows the code for configuring an ANN-based model with our example. In this example, we specify an ANN with two hidden layers and ten hidden nodes per layer. Additionally, we define five estimation repeats and 50 random shuffles of the estimation data per repeat. Optionally, the user can set the random seed for being able to replicate results purposes by adding the parameter seed in the configuration object.

To estimate the ANN-based model, the user executes the run method. The estimated ANN-based model returns the following outputs:

```
# Configure the ANN-based model
from py_np4vtt.model_ann import ModelANN, ConfigANN
config = ConfigANN(
    hiddenLayerNodes=[10, 10],
    trainingRepeats= 5,
    shufflesPerRepeat= 50,
    seed = None)

# Create the ANN-based model
ann = ModelANN(config, model_arrays)

# Access to data (e.g., training input data) using the 'dot' notation
X_train = ann.X_train

# Estimate the ANN-based model
ll_list, r2_list, vtt_list, est_time, avg_time = ann.run()
```

*Box 6.5: Configuration and estimation of an ANN-based model*

- `ll_list`: array of log-likelihoods at convergence per training repeat,
- `r2_list`: array of Rho-squared values per training repeat,
- `vtt_list`: array of VTTs per respondent per training repeat,
- `est_time`: the estimation time in seconds, and
- `avg_time`: the average estimation time per repetition.

The user can use the outputs to depict the VTT distribution. In Section 6.4.4, we show the results of the VTT distribution for the ANN-based model, together with the other models.

### 6.4.4. Recovering and visualising the VTT distribution

Table 6.4 summarises the estimation time of each nonparametric model. The estimations were performed in a MacBook Pro 2019 laptop with a 4-core Intel Core i5 CPU with 2.4GHz and 8GB of RAM. Similar results were obtained with a Linux machine with similar characteristics. The local constant and local Logit models have an estimation time of less than 1 second. In contrast, the Rouwendal' model took approximately one minute (68.99 seconds) to be estimated, and the ANN-based VTT model took almost 4 minutes (224 seconds) to be trained. The larger estimation time of the Rouwendal's model is explained by the computation of the standard errors. For the ANN-based model, the increased training time is explained because it relies on repeated training, while on average, each repetition takes around 45 seconds.

| Model | Estimation time (in seconds) |
|---|---|
| Local constant | 0.09 secs. |
| Local Logit | 0.71 secs. |
| Rouwendal's model | 68.99 secs. |
| ANN-based VTT model | Total: 223.91 secs. - average/repetition: 44.78 secs. |

*Table 6.4: Estimation time per model*

Figure 6.3 presents empirical CDFs (left) and histograms (right) obtained from the four models we used to demonstrate NP4VTT. The upper row shows the results of the Local constant model; the second row shows the Local logit model; the third row shows the Rouwendal model; the lower row shows the ANN-based VTT model. In these plots, the x-axis shows the VTT [€/hr.]; the y-axis depicts the estimate of the CDF (left) and the count (right).

Being able to produce and compare the results of multiple nonparametric models provides a profound understanding of the VTT distribution. But, it is important to interpret the results considering the sort of nonparametric model. For instance, we see that the Rouwendal predicts a considerably thinner tail than the Local constant and Local logit models. In the Rouwendal model, the thickness of the tail is a result of the assumption that respondents have a fixed probability of making decisions that are inconsistent with their underlying VTT. This probability is not a function of the BVTT. In this model, the thickness on the very end of the tail is thus not the result of observations for the fast and expensive alternative at the very end of the tail. Furthermore, the estimation process of the ANN-based model is enriched by repeated estimations, data expansion and random shuffling. In contrast, the Local constant model and Local logit model do not assume any underlying process regarding inconsistent choices, they do not consider any panel structure, and they do not enrich the data, strictly limiting themselves to get an estimate of the CDF in a given set of VTT points.

## 6.5. Conclusion

This paper introduces NP4VTT, a new Python package to estimate nonparametric models for inference of the VTT distribution. This package includes the following nonparametric models: Local constant, Local logit, Rouwendal, ANN-based VTT model, and Logistic regression. The modular set-up of NP4VTT allows to incorporate other nonparametric models in the future. We hope this package lowers the burden for researchers of using these powerful models to analyse the shape of the VTT distribution.

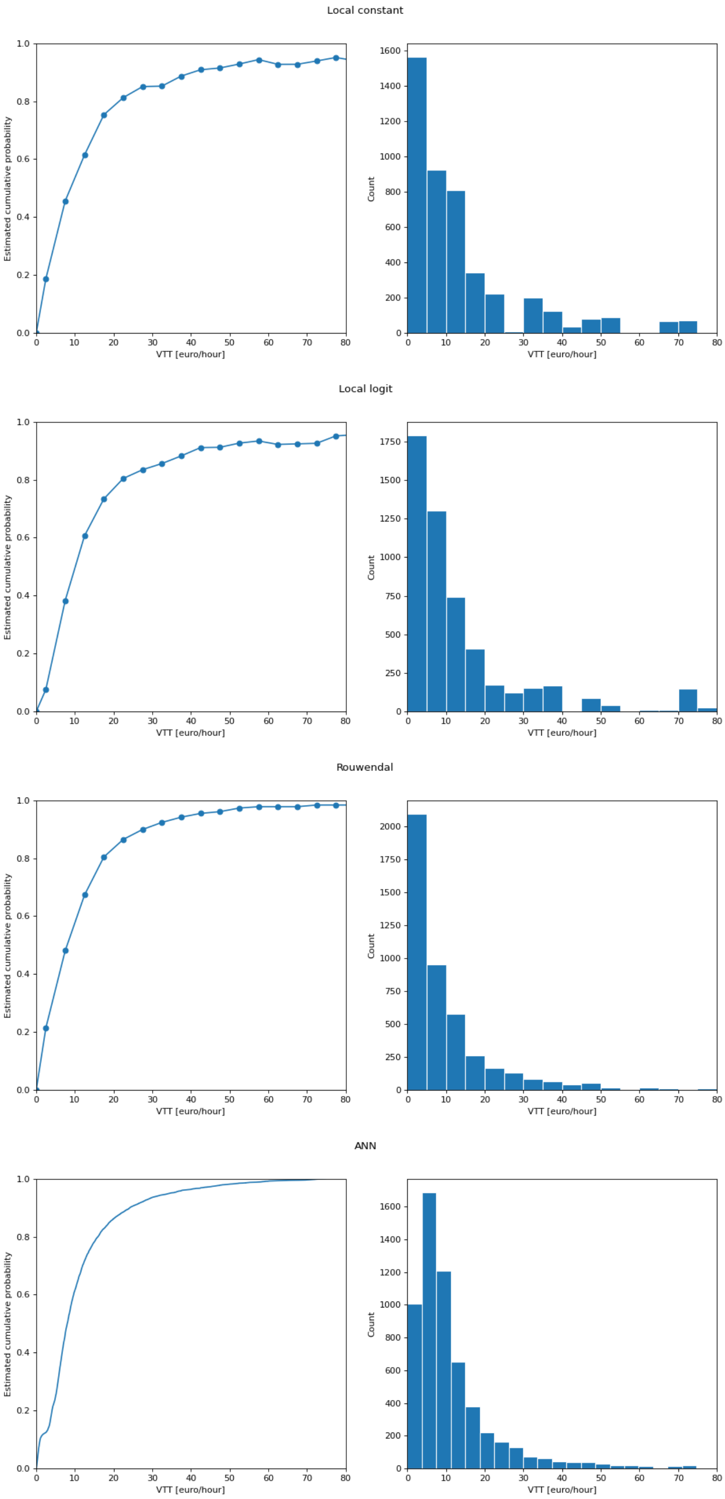


*Figure 6.3: Recovered VTT distributions using local constant (top), Local logit (second), Rouwendal (third) and ANN-based VTT model (bottom)*

From a substantive point of view, our results highlight the sensitivity of the recovered shape of the VTT distribution. While the advantage of nonparametric models is that the analyst does not need to assume the shape of the distribution, they do unequivocally produce the same shape of the VTT. The differences between the recovered distributions of the nonparametric models have various causes, such as whether the model accounts for a panel structure and how the model deals with stochasticity. We believe future research must dig further into the robustness of the VTT shape recovery, using both parametric and nonparametric approaches. We hope our Python package facilitates in this effort and encourages the development of new nonparametric models.

## Acknowledgements

We would like to thank Askill H. Halse for sharing the Norwegian 2009 VTT data with us and allowing us to make these data publicly available for demonstrating this software. José Ignacio Hernandez would like to thank João Pizani Flor for his valuable support and knowledge during the development of this software.

## 6.A. NP4VTT manual

The original article includes the NP4VTT manual as a single PDF file. For reasons of succintness, I do not include the complete manual in this thesis chapter. The interested reader can download the manual in the following link: https://ars.els-cdn.com/content/image/1-s2.0-S1755534523000283-mmc1.pdf

# Bibliography

Bierlaire, M. (2018) A short introduction to PandasBiogeme, *École polytechnique fédérale de Lausanne*, p. 22.

Börjesson, M., J. Eliasson (2014) Experiences from the Swedish Value of Time study, *Transportation Research Part A: Policy and Practice*, 59, pp. 144–158.

Cameron, T. A., M. D. James (1987) Efficient Estimation Methods for ”Closed-Ended” Contingent Valuation Surveys, *The Review of Economics and Statistics*, 69(2), pp. 269–276.

Carson, R., R. Mitchell, M. Hanemann, R. Kopp, S. Presser, P. Ruud (2003) Contingent valuation and lost passive use: Damages from the Exxon Valdez oil spill, *Environmental and Resource Economics*, 25(3), pp. 257–286.

Fan, J., N. E. Heckman, M. P. Wand (1995) Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-Likelihood Functions, *Journal of the American Statistical Association*, 90(429), pp. 141–150.

Fosgerau, M. (2006) Investigating the distribution of the value of travel time savings, *Transportation Research Part B: Methodological*, 40(8), pp. 688–707.

Fosgerau, M. (2007) Using nonparametrics to specify a model to measure the value of travel time, *Transportation Research Part A: Policy and Practice*, 41(9), pp. 842–856.

Fosgerau, M., K. Hjorth, S. V. Lyk-Jensen (2007) The danish value of time study, *Kgs. Lyngby, Denmark.*

Hess, S., A. Daly, M. Börjesson (2020) A critical appraisal of the use of simple time-money trade-offs for appraisal value of travel time measures, *Transportation*, 47(3), pp. 1541–1570.

Hess, S., D. Palma (2019) Apollo: A flexible, powerful and customisable freeware package for choice model estimation and application, *Journal of Choice Modelling*, 32, p. 100170.

Kouwenhoven, M., G. C. de Jong, P. Koster, V. A. C. van den Berg, E. T. Verhoef, J. Bates, P. M. J. Warffemius (2014) New values of time and reliability in passenger transport in The Netherlands, *Research in Transportation Economics*, 47, pp. 37–49.

Ojeda-Cabral, M., R. Batley, S. Hess (2016) The value of travel time: Random utility versus random valuation, *Transportmetrica A: Transport Science*, 12(3), pp. 230–248.

Ojeda-Cabral, M., C. G. Chorus (2016) Value of travel time changes: Theory and simulation to understand the connection between Random Valuation and Random Utility methods, *Transport Policy*, 48, pp. 139–145.

Ramjerdi, F., S. Flügel, H. Samstad, M. Killi (2010) Value of time, safety and environment in passenger transport–Time, *TØI report B*, 1053.

Rouwendal, J., A. de Blaeij, P. Rietveld, E. Verhoef (2010) The information content of a stated choice experiment: A new method and its application to the value of a statistical life, *Transportation Research Part B: Methodological*, 44(1), pp. 136–151.

Small, K. A. (2012) Valuation of travel time, *Economics of Transportation*, 1(1), pp. 2–14.

StataCorp, L. P. (2005) Stata base reference manual, *College Station: StataCorp LLC*.

van Cranenburgh, S., M. Kouwenhoven (2021) An artificial neural network based method to uncover the value-of-travel-time distribution, *Transportation*, 48(5), pp. 2545–2583.

# Chapter 7

# Conclusion

This thesis provides five studies that propose and make available new data-driven methods to study individual choice behaviour in Participatory Value Evaluation (PVE) experiments and Discrete Choice Experiments (DCEs). The main goal of this thesis is *To investigate the extent that data-driven methods can be used for analysing individual choice behaviour from SC experiments, either to complement theory-driven choice models, or alternatives to theory-driven choice models; and to provide methodological and substantive contributions for such purposes.*. This research goal is scoped to two specific SC experiments: PVE and DCEs.

This chapter provides some conclusions and implications derived from the studies of the present thesis: Section 7.1 summarises the conclusions of each study; Section 7.2 provides some overall conclusions, implications and a final reflection about the findings of this thesis; finally, Section 7.3 provides some directions for further research.

## 7.1. Conclusions per individual study

### Study 1: A large-scale deployment of a Participatory Value Evaluation experiment

> *This study introduces PVE experiments to the reader and illustrates how they work in a real-life application. In addition, this study shows the standard approach to model PVE experiments using a theory-driven choice model based on Kuhn-Tucker models, what behaviourally-relevant outcomes are obtained and interpreted, and which challenges emerge from analysing PVE experiments with these models.*

This study presents a large-scale PVE experiment conducted to elicit the preferences of 30,000 Dutch citizens for relaxing COVID-19 measures. The modelling of this PVE experiment is done with a Kuhn-Tucker theory-driven choice model based on utility theory (Dekker et al., 2019).

The modelling results show, for instance, that Dutch citizens assign equal value to a reduction of 100 deaths among individuals younger than 70 years old and a reduction of 168 deaths among individuals aged 70 years or older. Yet, the estimated parameters suggest that respondents assign a considerably smaller value to the impacts of each measure (represented by taste parameters) than the value assigned to the measures themselves (represented by policy-specific constants). The combination of alternatives that generates the highest societal welfare is to re-open businesses (excluding the hospitality industry), re-open contact professions, and allow direct family members from other households to have social contact again.

The qualitative results show that this PVE experiment is received with high regard by citizens: nearly 60 percent of respondents indicate that their participation in this experiment heightens their awareness of the consequences associated with relaxing COVID-19 measures, and almost 80 percent of the respondents perceive this method as a positive approach to involve citizens in government decision-making processes. Despite the latter, a majority (69 percent) of respondents still feel that the government should give greater weight to expert advice when making decisions.

## Study 2: Data-driven methods to assist choice models for Participatory Value Evaluation experiments

> *This study addresses* ***RG1:*** *To examine the extent that data-driven methods can be used as a complement to theory-driven choice models for PVE experiments, and to develop methodological tools for this purpose.*

This study proposes three procedures based on Association Rules (AR) learning and Random Forest (RF) to assist the specification of a theory-driven portfolio choice model for PVE experiments (Bahamonde-Birke & Mouter, 2019). Firstly, AR learning is combined with a methodological-iterative (MI) algorithm to identify and incorporate relevant interactions between chosen alternatives in the PVE experiment. The found interactions are incorporated as interaction parameters in the portfolio choice model. Secondly, RF is combined with the MI algorithm to identify the most (least) important attributes of the PVE experiment and determine the inclusion or exclusion of attribute-specific parameters in the portfolio choice model. Thirdly, RF is used to test the validity of the behavioural assumptions of theory-driven choice models. This is done by comparing the rankings of the most likely chosen combinations of alternatives.

On the one hand, this study shows that the assisted portfolio choice models reach a model fit improvement of no more than 1.3 decimal points of Rho-Squared, compared with a non-assisted choice model. On the other hand, the assisted models provide new behaviourally-relevant insights: the interaction terms of the AR-assisted portfolio choice model have a behavioural interpretation as positive/negative synergies between chosen alternatives, and the RF-assisted portfolio choice model leaves the statistically significant specific attribute effects per alternative of the PVE experiment. Finally, it is found that the AR-assisted portfolio choice model is closer to the "true" ranking of most likely chosen portfolios than the RF-assisted model.

In summary, this study shows how two specific data-driven methods, AR learning and RF, can be used to assist the specification of theory-driven choice models. The assisted choice models result in model fit improvements and provide behaviourally-relevant insights that otherwise would take considerably longer time and higher computational resources to be found.

## Study 3: Explainable artificial intelligence to study Participatory Value Evaluation experiments

> *This study addresses **RG2**: To examine the extent that data-driven methods can be used as alternatives to theory-driven choice models for PVE experiments, and to develop methodological tools for this purpose.*

This study analyses the preferences of Dutch citizens for reimposing COVID-19 measures in the Netherlands using explainable artificial intelligence methods (XAI). Specifically, the data is modelled with XGBoost, a supervised ensemble machine learning model. Then, SHAP (SHapley Additive exPlanations), an XAI method, identifies what explains the respondents' support for COVID-19 measures. SHAP is used to identify: 1) the most (least) important covariates that explain the support for each COVID-19 measure, on average, and 2) observed heterogeneity across respondents or COVID-19 measures and potential non-linear effects for different groups of respondents.

Results show that, on average, the support for COVID-19 measures is primarily associated with the respondents' age, their weight to citizen advice versus scientific advice, and their perceived risk of getting very ill from COVID-19. Further inspection of individual SHAP values for these variables shows several forms of observed heterogeneity across respondents, such as clusters of people with similar preferences, sparse distributions and non-linear (e.g., U-shape, piecewise linear) effects of specific variables. Notably, these results provide a considerably higher level of detail when compared with conventional modelling approaches for PVE experiments, namely choice models and LCCA (Latent Class Cluster Analysis).

Overall, this study shows how a specific data-driven modelling approach based on XAI can be used on PVE experiments data, what new behaviourally-relevant insights can be obtained from such an approach and how these insights are compared with those obtained from conventional modelling approaches.

## Study 4: An economically-consistent discrete choice model based on artificial neural networks

> *This study addresses **RG3**: To develop a new discrete choice model based on data-driven methods that balances flexibility to learn the utility function from the data, with consistency with economic assumptions.*

This study introduces the "Alternative-Specific and Shared weights Neural Network" (ASS-NN) model. The ASS-NN is a discrete choice model based on Artificial Neural Networks (ANNs) that combines the flexibility of ANNs with consistency with RUM theory and fungibility of money ("one euro is one euro"), which guarantees equal marginal utility of costs across different alternatives. Consequently, the outcomes obtained from the ASS-NN (e.g., marginal utilities and willingness to pay measures) are also consistent with such assumptions. The use of the ASS-NN is illustrated using Monte Carlo simulations and empirical data from the Swissmetro dataset to obtain estimates of marginal utilities, the value of travel time (VTT) and the value of waiting time (VoWT).

The Monte Carlo analyses show that the ASS-NN model successfully recovers the true utility functional form and accurately predicts the marginal utilities and the VTT directly from the data. When applied to empirical data, the ASS-NN model outperforms conventional multinomial logit models under different utility specifications. Furthermore, the predicted marginal utilities are consistent with the fungibility of money assumption, whereas the marginal utility of travel time and headway vary across different travel models. Notably, the ASS-NN model predicts that public transport modes (i.e., Swissmetro and train) are more attractive for shorter trips (in terms of travel time), while cars become more appealing for longer trips. Regarding welfare measures, the ASS-NN predicts that respondents assign a higher average VTT to public transport trips compared to car trips for short-time trips, while this tendency is reverted as the travel time increases. Finally, the ASS-NN predicts that the VoWT for Swissmetro trips is higher than for train trips, contradicting the predictions of the multinomial logit models.

Overall, this study shows that the ASS-NN model is a promising data-driven alternative to theory-driven choice models. The ASS-NN provides behaviourally-relevant insights while balancing the advantages of ANNs to learn from the data with consistency with economic assumptions.

### Study 5: A new software package to estimate nonparametric models to compute the value of travel time distribution from binary choice experiments

> *This study addresses **RG4**: To develop a new software tool to estimate and compare the outcomes of different data-driven methods simply and conveniently.*

This study introduces NP4VTT, a Python package that provides five nonparametric models to estimate the VTT distribution from two-attribute-two-alternative DCEs. These models are: 1) Local Constant (Fosgerau, 2006), 2) Local Logit (Fosgerau, 2007), 3) Rouwendal's model (Rouwendal et al., 2010), 4) an ANN-based model (van Cranenburgh & Kouwenhoven, 2021), and 5) a Logistic regression model based on the ANN-based model. NP4VTT provides researchers with a unified syntax, enabling them to easily and conveniently estimate the VTT distribution using nonparametric models.

To illustrate the use of NP4VTT, data from the Norwegian VTT study (Ramjerdi et al., 2010) is used to estimate and compare the VTT distribution using four of the nonparametric models provided in this package. The results of this application show that the recovered VTT distribution consistently follows the same shape across models, and differences are attributed to modelling factors (e.g., whether the panel structure is considered) or stochasticity.

## 7.2. Overall conclusions and implications

This thesis provides shows how data-driven methods can be used for studying individual choice behaviour in PVE experiments and DCEs. Furthermore, new methodological tools are provided, namely new data analysis methods and new software tools. The main research goal and sub-goals are achieved in substance, as shown in Section 7.1.

Yet, it is relevant to put the findings and conclusions of this thesis in perspective. The interested reader (e.g., a choice modeller, a policymaker) shall wonder: are the new findings provided in this thesis worth their costs (e.g., a higher effort to implement new methods, higher estimation time)? Do the data-driven methods this thesis investigates lead to considerably different conclusions than a conventional choice model? Naturally, these questions contain a subjective component, as some researchers may focus more on model fit and predictive power. In contrast, others might be interested in new insights for studying choice behaviour or decision-making. Below, I provide three overall conclusions and a final perspective derived from the findings of this thesis, as well as some implications for choice modellers, applied researchers and policymaking:

## Model fit improvements are modest compared to those from conventional, non-assisted choice models.

Throughout this thesis, it is found that using data-driven methods leads to model fit improvements, compared with using conventional, non-assisted choice models. Nevertheless, from the author's perspective, such improvements shall be considered "modest" for predicting choice behaviour. Specifically, when AR learning and RF are used to assist a portfolio choice model (Chapter 3), the increase of Rho-Squared is no more than two decimal points of Rho-Squared, compared with a non-assisted model. The ASS-NN model (Chapter 5) attains an increase of no more than three decimal points of Rho-Squared, compared with conventional multinomial logit models.

These findings provide new evidence to a longstanding debate in the choice modelling community: is the data-driven methods' predictive performance consistently superior to that from conventional choice models? Such question has been revisited for years in literature (Bentz & Merunka, 2000; Alwosheel, 2020; Wang et al., 2021; Ali et al., 2023). For instance, Bentz & Merunka (2000) finds that feed-forward ANNs reach three more decimal points of Rho-Squared[1], compared with a multinomial logit model, similar to the results of this thesis. Alwosheel (2020) also finds that ANNs did not reach considerable model fit gains compared with conventional discrete choice models applied on stated choice experiments in the transport field. Wang et al. (2021) compares the prediction performance of several discrete choice and machine learning models, finding that the former reach only three to four percentage points less accuracy than the best-performing machine learning models. In a more recent study, Ali et al. (2023) finds that choice models outperform ANNs and gradient-boosted machines in the context of vehicle ownership decisions.

[1] In their study, Rho-Squared is presented as the amount of uncertainty or $U^2$.

These findings may serve as a warning to modellers and applied researchers interested in predicting choice behaviour: data-driven methods do not necessarily outperform choice models, and when they do, the gains are not that high. Naturally, this does not mean that data-driven methods shall be entirely discarded, as two to three points of model fit gains might be relevant in specific contexts, for instance when human lives or health gains are at stake Mulderij et al. (2021); Rotteveel et al. (2022); Mouter et al. (2022). Yet, the question of whether to consider (or discard) the data-driven methods provided in this thesis goes beyond the scope of this work.

## Modellers and practitioners count with new information for improving models and for decision-making.

A key result of this thesis is the considerable number of new outcomes obtained from data-driven methods. These findings concern, namely, interactions between chosen alternatives (Chapter 3), variable importances (Chapters 3 and 4), individual-level effects and observed heterogeneity across individuals (Chapter 4) and marginal effects and welfare measures for specific individuals and groups (Chapters 5 and 6).

For choice modellers, these findings contribute to the literature of assisted specification of choice models (Hillel et al., 2019; Shiftan & Bekhor, 2020; Ortelli et al., 2021; Ghorbani et al., 2023), as well as providing outcomes that can be used for supporting choice modellers in other ways than what is shown in this thesis. For instance, the outcomes obtained from SHAP values (Chapter 4) can guide the specification of random parameters or specify nonlinear forms of taste parameters for specific attributes or covariates of a theory-driven choice model. In addition, choice modellers now count with outcomes that can be compared and contrasted with the outcomes of choice models. By doing so, it is possible to confirm (or not) potential model specification issues. Some suggested comparisons are: the average effects observed from SHAP values, compared with the taste parameters of a choice model (Chapter 4), or the predicted marginal utilities of the ASS-NN with those from choice models (Chapter 5).

For applied researchers and policymakers, these findings imply that more information to support decision-making is available. For instance, AR learning can be used to investigate further bundling preferences for shopping (Sharpe & Staelin, 2010), vacation and hotelling (van Cranenburgh et al., 2014; Dominique-Ferreira & Antunes, 2019) and to support the design of Mobility-as-a-Service (MaaS) systems (Reck et al., 2020; Ho et al., 2021). As another example, SHAP can be used to identify observed heterogeneity in healthcare-related choice experiments, as this topic has gained interest in literature (see, for instance Zhou et al., 2018; Vass et al., 2022; Karim et al., 2022).

However, these findings also imply that researchers should elucidate what parts of the newly found information are relevant for making decisions and the extent such information can be synthesised for effective communication. For instance, there is a limited understanding in the literature about how machine learning models can be used for policymaking (Hillel et al., 2021). Applied researchers and policymakers may question if they actually need the disaggregated information provided by the methods of this thesis, compared with conventional choice models, which are often easier to interpret, they are representative of the whole population and are entirely built in a theory of choice behaviour. Further research is needed to synthesise the new insights derived from data-driven methods effectively, especially when such insights are planned to be used in policymaking, where time and space restrictions (i.e., in policy reports or to inform high-ranked public officers) are more common, and insights must be communicated efficiently.

### New data-driven methods, code and software are available in the public domain.

Part of the primary goal of this thesis is to provide new methods for data-driven analysis of choice behaviour. This goal is achieved by the specific data-driven methods proposed in Chapters 2 to 5 and by the new software tool (NP4VTT) in Chapter 6. Furthermore, following the research directions from Alwosheel (2020) and in line with standard practices of the machine learning community, an effort is made to make the methods provided in this thesis openly available for the general public in three ways: 1) publishing each study in open-access scientific journals, 2) when possible, using open-source tools, programming languages (i.e., Python) and datasets, and 3) when possible, providing the source code in public-domain repositories. By doing so, these methods can be more accessible for choice modellers, applied researchers and policymakers interested in using them.

Despite the latter, solely pushing for making data-driven methods openly accessible would not be enough to encourage the choice modelling community to use them. In the author's view, a critical aspect to consider is the homogeneity between terms employed in data-driven methods (and in the particular case of machine learning) with the concepts of choice modelling. Data-driven methods, particularly machine learning, and choice modelling, hold several shared concepts (van Cranenburgh et al., 2022; Hillel et al., 2021), albeit with different names (e.g., sigmoid and logit function, training and estimation). These semantic differences represent a barrier for researchers more familiar with choice modelling concepts to explore the potential of data-driven methods, even though they can find them in the public domain.

### Data-driven methods complement choice models; they do not substitute them

A final question worth doing is if, based on the findings of this thesis, data-driven methods are in a maturity stage that allows them to be alternatives (i.e., to replace) theory-driven choice models. Based on the findings of this thesis, in the author's view, data-driven methods complement choice models rather than substituting them. This thesis shows the considerable amount of insights that can be obtained from data-driven methods for improving new models and decision-making. However, such methods' potential for predictive tasks is not considerably different from a theory-driven choice model. Furthermore, researchers still face challenges in summarising all the new insights from the studies presented in this thesis in ways that can be easy to understand and present, which is one of the main strengths of theory-driven choice models: their interpretability, despite their simplicity.

Naturally, some of the challenges for data-driven methods expressed here (and in the preceding sections of this thesis) may have an eventual solution in the future, and they may require new research or even a new thesis. In the author's view, choice models are still a valuable, parsimonious and robust modelling paradigm for studying individual choice behaviour that can be complemented by data-driven methods, including the ones provided in this thesis.

## 7.3. Further research directions

Across its chapters, this thesis provides some directions for new research on data-driven methods for PVE experiments and DCEs. Likewise, the conclusions of this chapter show that, while data-driven methods present considerable new insights of behavioural interest, some challenges must be addressed to make these methods a viable alternative to theory-driven choice models. Below, I provide a non-exhaustive list of directions for further research:

- One of the conclusions of this thesis is that using the proposed data-driven methods result in modest model fit improvements. While some explanations are inferred in the same conclusions (see Section 7.2), further research is needed in order to shed light on the reasons behind these results. A good starting point is on the thesis of Alwosheel (2020), who suggests that that ANNs yield to better prediction performance in revealed preference data. Future research could investigate whether the data-driven methods proposed by this thesis lead to higher predictive performance on revealed preference datasets than in stated choice experiments data.

- This thesis concludes that, although the proposed data-driven methods provide a considerable number of behavioural insights for policymaking, synthesising that information is still challenging to provide meaningful advice. Further research should be conducted find new forms to synthesise and present the information obtained from the data-driven methods proposed by this thesis. Some specific ideas are: investigate the extent that the information obtained from SHAP can be embedded in an online environment for public consultation; or developing a method to visualise the interactions between chosen alternatives of a PVE experiment that are identified by AR learning.

- While this thesis provides NP4VTT as a concrete contribution to make data-driven methods more accessible to the choice modelling community (Chapter 6), in the author's view, there is still room for more software for the same purposes. A good start would be to develop a software package that integrates discrete choice models based on data-driven methods, such as ANNs (Wang et al., 2020a,b; van Cranenburgh & Kouwenhoven, 2021), including the ASS-NN proposed by this thesis.

- The feature importance used for the random forest model in Chapter 3 presents three key limitations compared with other types of explanations, such as SHAP, which is used in Chapter 4. Firstly, feature importances must be computed in the training dataset. SHAP, on the other hand, can be applied either in any data instance, either the training data, a holdout (test) sample, or a hypothetical instance. Secondly, random forests feature importances could be overestimated for numerical variables or with a high cardinality (i.e., with several attribute levels, in the case of our application). Thirdly, random forests feature importances can only inform how much important is a variable for the model, but it does not provide any information about whether the variable is associated with positive or negative effects, while SHAP does.

- The methods proposed in this thesis are either theory-agnostic (Chapters 4 and 6) or they are embedded into a random utility maximisation (RUM) framework (Chapters 3 and 5). However, alternative theories of choice behaviour studied in choice modelling were not considered, such as prospect theory (e.g., reference dependence), regret theory or models for moral decision-making. While the choice modelling field has longtime studied and developed models for alternative forms of choice behaviour, the methodological research in this area for data-driven methods is rather scarce, with the notable exception of van Cranenburgh & Alwosheel (2019), while a recent study (Smeele et al., 2023) points on the relevance of machine learning for studying moral decision-making.

- The methods proposed by this thesis are only conceived to work with tabular (numerical) data, which is the conventional data employed by theory-driven choice models. However, richer data types, namely images or texts, were not explored. In the author's opinion, there is a clear opportunity to draw new research in this area, especially in the case of PVE experiments where, for instance, a considerable amount of written motivations are collected but not used for modelling (Chapter 2). To address this, a potential research direction could be to extend the work of Liscio et al. (2021) and use Natural Language Processing (NLP) methodologies to identify relevant patterns from PVE experiments data and integrate them into a portfolio choice model.

- An aspect that is yet not addressed in literature can be summarised in the following question: do PVE experiments really need new data analysis tools? In the author's view, other aspects of PVE experiments have been longtime not considered, namely: how PVE experiments experimental design should be conducted; to what extent respondents of PVE experiments pay (or not) attention to all the information presented in the choice task; to what extent respondents do trade-offs between alternatives, attributes and budget allocations in PVE experiments? While some of such questions exceed the main scope of this thesis, if they are not addressed, neither theory-driven choice models nor data-driven methods will be able to find more relevant information from PVE experiments data, aside from the findings already identified in this work.

# Bibliography

Ali, A., A. Kalatian, C. F. Choudhury (2023) Comparing and contrasting choice model and machine learning techniques in the context of vehicle ownership decisions, *Transportation Research Part A: Policy and Practice*, 173, p. 103727.

Alwosheel, A. S. A. (2020) Trustworthy and Explainable Artificial Neural Networks for Choice Behaviour Analysis.

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

Bentz, Y., D. Merunka (2000) Neural networks and the multinomial logit for brand choice modelling: A hybrid approach, *Journal of Forecasting*, 19(3), pp. 177–200.

Dekker, T., P. Koster, N. Mouter (2019) The economics of participatory value evaluation.

Dominique-Ferreira, S., C. Antunes (2019) Estimating the price range and the effect of price bundling strategies: An application to the hotel sector, *European journal of management and business economics*, 29(2), pp. 166–181.

Fosgerau, M. (2006) Investigating the distribution of the value of travel time savings, *Transportation Research Part B: Methodological*, 40(8), pp. 688–707.

Fosgerau, M. (2007) Using nonparametrics to specify a model to measure the value of travel time, *Transportation Research Part A: Policy and Practice*, 41(9), pp. 842–856.

Ghorbani, A., N. Nassir, P. S. Lavieri, P. B. Beeramoole (2023) A sparse identification approach for automating choice models' specification, *arXiv preprint arXiv:2305.00912*.

Hillel, T., M. Bierlaire, M. Elshafie, Y. Jin (2019) Weak teachers: Assisted specification of discrete choice models using ensemble learning, in: *hEART 2019: 8th Symposium of the European Association for Research in Transportation. Budapest, Hungary.*

Hillel, T., M. Bierlaire, M. Z. E. B. Elshafie, Y. Jin (2021) A systematic review of machine learning classification methodologies for modelling passenger mode choice, *Journal of Choice Modelling*, 38, p. 100221.

Ho, C. Q., D. A. Hensher, D. J. Reck, S. Lorimer, I. Lu (2021) Maas bundle design and implementation: Lessons from the sydney maas trial, *Transportation Research Part A: Policy and Practice*, 149, pp. 339–376.

Karim, S., B. M. Craig, C. Vass, C. G. Groothuis-Oudshoorn (2022) Current practices for accounting for preference heterogeneity in health-related discrete choice experiments: A systematic review, *PharmacoEconomics*, 40(10), pp. 943–956.

Liscio, E., M. Meer, L. Siebert, C. Jonker, N. Mouter, P. Murukannaiah (2021) Identifying and evaluating context-specific values, *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021*, 10, pp. , 3–7,.

Mouter, N., K. T. Jara, J. I. Hernandez, M. Kroesen, M. de Vries, T. Geijsen, F. Kroese, E. Uiters, M. de Bruin (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, p. 115430.

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

Ortelli, N., T. Hillel, F. C. Pereira, M. de Lapparent, M. Bierlaire (2021) Assisted specification of discrete choice models, *Journal of Choice Modelling*, 39, p. 100285.

Ramjerdi, F., S. Flügel, H. Samstad, M. Killi (2010) Value of time, safety and environment in passenger transport–Time, *TØI report B*, 1053.

Reck, D. J., D. A. Hensher, C. Q. Ho (2020) MaaS bundle design, *Transportation Research Part A: Policy and Practice*, 141, pp. 485–501.

Rotteveel, A. H., M. S. Lambooij, E. a. B. Over, J. I. Hernández, A. W. M. Suijkerbuijk, A. T. de Blaeij, G. A. de Wit, N. Mouter (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), pp. 428–443.

Rouwendal, J., A. de Blaeij, P. Rietveld, E. Verhoef (2010) The information content of a stated choice experiment: A new method and its application to the value of a statistical life, *Transportation Research Part B: Methodological*, 44(1), pp. 136–151.

Sharpe, K. M., R. Staelin (2010) Consumption effects of bundling: consumer perceptions, firm actions, and public policy implications, *Journal of Public Policy & Marketing*, 29(2), pp. 170–188.

Shiftan, Y., S. Bekhor (2020) Utilizing a random forest classifier for a methodological-iterative discrete choice model specification and estimation, in: *hEART 2020: 9th Symposium of the European Association for Research in Transportation. Lyon, France.*

Smeele, N. V., C. G. Chorus, M. H. Schermer, E. W. de Bekker-Grob (2023) Towards machine learning for moral choice analysis in health economics: A literature review and research agenda, *Social Science & Medicine*, p. 115910.

van Cranenburgh, S., A. Alwosheel (2019) An artificial neural network based approach to investigate travellers' decision rules, *Transportation Research Part C: Emerging Technologies*, 98, pp. 152–166.

van Cranenburgh, S., CG. Chorus, B. van Wee (2014) Vacation behaviour under high travel cost conditions–A stated preference of revealed preference approach, *Tourism Management*, 43, pp. 105–118.

van Cranenburgh, S., M. Kouwenhoven (2021) An artificial neural network based method to uncover the value-of-travel-time distribution, *Transportation*, 48(5), pp. 2545–2583.

van Cranenburgh, S., S. Wang, A. Vij, F. Pereira, J. Walker (2022) Choice modelling in the age of machine learning - Discussion paper, *Journal of Choice Modelling*, 42, p. 100340.

Vass, C., M. Boeri, S. Karim, D. Marshall, B. Craig, K.-A. Ho, D. Mott, S. Ngorsuraches, S. M. Badawy, A. Mühlbacher, et al. (2022) Accounting for preference heterogeneity in discrete-choice experiments: an ispor special interest group report, *Value in Health*, 25(5), pp. 685–694.

Wang, S., B. Mo, S. Hess, J. Zhao (2021) Comparing hundreds of machine learning classifiers and discrete choice models in predicting travel behavior: An empirical benchmark, *arXiv preprint arXiv:2102.01130*.

Wang, S., B. Mo, J. Zhao (2020a) Deep neural networks for choice analysis: Architecture design with alternative-specific utility functions, *Transportation Research Part C: Emerging Technologies*, 112, pp. 234–251.

Wang, S., Q. Wang, J. Zhao (2020b) Deep neural networks for choice analysis: Extracting complete economic information for interpretation, *Transportation Research Part C: Emerging Technologies*, 118, p. 102701.

Zhou, M., W. M. Thayer, J. F. Bridges (2018) Using latent class analysis to model preference heterogeneity in health: a systematic review, *Pharmacoeconomics*, 36, pp. 175–187.

# Summary

Since its origins in the 1970s, choice modelling has become an important field of study in diverse areas, including transportation, health economics, environmental economics and marketing. Choice modellers have developed several methods to collect and model individual choices. Researchers and policymakers use such methods to understand individual preferences in diverse contexts, derive economic values or predict behaviour.

Over the years, the field of choice modelling has been developed in two key areas. Firstly, choice modellers have developed new data collection tools to account for more realistic forms of decision-making. While discrete choice experiments (DCEs) are still popular and highly customisable, they force respondents to choose among mutually-exclusive alternatives, which may not reflect how individuals choose in real life. In response, new SC experiments have been proposed to incorporate more realistic forms of decision-making, such as Participatory Value Evaluation (PVE). In a PVE experiment, respondents select a combination of alternatives without surpassing resource constraints. Secondly, while theory-driven models based on utility theory, e.g., random utility maximisation (RUM) or Kuhn-Tucker, are still the norm to model choice behaviour, there is a broader recognition that individual' behaviour is ultimately unknown from the analyst perspective, data-driven methods can help to uncover such behaviour.

Despite the latter, to the author's knowledge, three methodological and practical challenges are still unresolved in the literature. Firstly, no research has been done to explore the potential of data-driven methods to analyse data from SC experiments outside DCEs, and in particular for PVE experiments, either as complements to improve the specification of choice models or as standalone data analysis methods. Secondly, while data-driven methods for discrete choices (and DCEs) are available in the literature, such methods either sacrifice their flexibility to learn from the data to satisfy consistency assumptions or vice versa, hindering a more widespread use of these models in real-life policy applications. Thirdly, there is a lack of software tools to estimate and compare data-driven methods easily and conveniently, hindering their widespread

use.

Considering these challenges, this thesis further investigates how data-driven methods can be used for analysing individual choice behaviour from SC experiments, either to complement theory-driven choice models or alternatives to theory-driven choice models; and to provide methodological and substantive contributions for such purposes. This thesis scopes its research to two specific SC experiments: PVE and DCEs.

Chapter 2 introduces the reader to how PVE experiments are applied in real life, how they are conventionally analysed with theory-driven choice models and which challenges emerge from analysing PVE experiments with these models. In this PVE application, almost 30,000 Dutch citizens advised the government on which COVID-19 should be relaxed. Data is analysed with a theory-driven choice model based on utility theory. The modelling results show, for instance, that citizens assign equal value to reduce 100 deaths of people younger than 70 years old to reducing 168 deaths of people of 70 years or older. Yet, the estimated parameters suggest that respondents assign a considerably smaller value to the impacts of each measure (represented by taste parameters) than the value assigned to the measures themselves (represented by policy-specific constants). Citizens' preferred combination of measures is to re-open businesses other than the hospitality industry and allow direct family members to have social contact again. The qualitative results show that citizens highly regard PVE as a participation method, yet they still feel the government should prioritise expert advice when making decisions.

Chapter 3 proposes three procedures based on association rules (AR) learning and random forests (RF) to assist the specification and test the validity of the assumptions of theory-driven choice models for PVE experiments. First, a methodological-iterative (MI) algorithm is combined with AR learning to identify relevant interactions between chosen alternatives of the PVE experiment and incorporate them in a portfolio choice model. Second, the MI algorithm is combined with RF to identify the variable importances and decide the inclusion/exclusion of attributes in the portfolio choice model. Third, RF is used to test the validity of the behavioural assumptions of theory-driven choice models. Results show that the assisted portfolio choice models reach small model fit improvements compared with non-assisted models. However, the assisted models provide new behaviourally-relevant insights, namely relevant interaction terms that can be interpreted as positive and negative synergies or specific attribute effects per alternative. Finally, the assisted model with AR learning is shown to be closer to the "true" choice behaviour of the PVE experiment. These findings can be used by choice modellers to improve the specification of models and by policymakers to obtain new insights of behavioural relevance.

Chapter 4 shows how XGBoost and SHAP -a machine learning model and explainable artificial intelligence method, respectively- can be used to analyse PVE experiments data as an alternative to theory-driven analysis. The analyses are done with data from a PVE experiment to study the preferences of Dutch citizens for reimposing COVID-19 measures. The analyses identify the most (least) important covariates that explain the support for COVID-19 measures, as well as observed heterogeneity across respondents and measures. Results show that the support for COVID-19 measures is primarily associated with the respondents' age, their weight to citizen advice versus scientific advice, and their perceived risk of getting ill from COVID-19. In addition, several forms of observed heterogeneity are identified, such as clusters of people with similar preferences or non-linear effects of each covariate. These results provide considerably more detail than conventional modelling approaches for PVE, namely theory-driven choice models and latent class cluster analysis. Policymakers can use this information for tailoring policies or building information campaigns to increase support for COVID-19 measures.

Chapter 5 proposes the "Alternative-Specific and Shared weights Neural Network" (ASS-NN) model. The ASS-NN model is a new discrete choice model based on artificial neural networks that balances flexibility to learn utility functions from the data while satisfying consistency with RUM and fungibility of money (i.e., "one euro is one euro"). The ASS-NN model is tested with Monte Carlo experiments and empirical data from the Swissmetro dataset. Results from the Monte Carlo experiment show that the ASS-NN model successfully recovers the true utility functional form and accurately predicts the marginal utilities and the value of travel time (VTT). When the ASS-NN is applied to empirical data, it outperforms multinomial logit models in terms of model fit and provides marginal utility values consistent with the fungibility of money assumption. The ASS-NN predicts that respondents assign a higher average VTT to public transport trips (i.e., train and Swissmetro) for short-time trips, being this reverted as the travel time increases. Overall, the ASS-NN is a promising data-driven alternative to theory-driven choice models that provides behaviourally-relevant insights while balancing the advantages of ANNs to learn from the data with consistency with economic assumptions.

Chapter 6 introduces NP4VTT, a new software tool that provides five nonparametric models to uncover the VTT distribution from two-attribute-two-alternative DCEs. NP4VTT provides researchers with a unified syntax, enabling them to easily and conveniently estimate the VTT distribution using nonparametric models. The use of NP4VTT is illustrated using data from the Norwegian VTT study. The results show that the recovered VTT distribution consistently follows the same shape across models, and differences are attributed to modelling factors or stochasticity.

This thesis concludes by highlighting that while the primary research goal and sub-goals are achieved, the relevance of the findings and conclusions shall be put into perspective. Firstly, using data-driven methods lead to modest model fit improvements. Thus, researchers or policymakers interested in predicting behaviour should not expect considerable differences compared with conventional choice models. Secondly, new insights of behavioural interest are found. Choice modellers could benefit from these insights to contrast or further assist the development of choice models, and policymakers count with new and more detailed information for decision-making. However, researchers now may have to elucidate what parts of the newly found information are relevant for making decisions, as well as the extent such information can be synthesised for effective communication. Thirdly, while this thesis makes more data-driven methods available, there are still challenges to making these methods more amicable to researchers accustomed to the concepts and structure of the choice modelling community. In conclusion, data-driven methods complement the insights from theory-driven choice models, but they do not substitute them. Theory-driven choice models are still a valuable, parsimonious and robust modelling paradigm for studying choice behaviour that can be complemented by data-driven methods, such as the ones proposed in this thesis.

José Ignacio HERNÁNDEZ HERNÁNDEZ

# Samenvatting

Sinds het ontstaan in de jaren 1970 is keuzemodellering een belangrijk studiegebied geworden op verschillende gebieden, waaronder transport, gezondheidseconomie, milieueconomie en marketing. Keuzemodelleurs hebben verschillende methoden ontwikkeld om individuele keuzes te verzamelen en te modelleren. Onderzoekers en beleidsmakers gebruiken dergelijke methoden om individuele voorkeuren in verschillende contexten te begrijpen, economische waarden af te leiden of gedrag te voorspellen.

In de loop der jaren heeft het vakgebied van de keuzemodellering zich op twee belangrijke gebieden ontwikkeld. Ten eerste hebben keuzemodelleurs nieuwe instrumenten voor gegevensverzameling ontwikkeld om rekening te houden met meer realistische vormen van besluitvorming. Hoewel discrete keuze-experimenten (DCE's) nog steeds populair en zeer aanpasbaar zijn, dwingen ze respondenten om te kiezen tussen elkaar uitsluitende alternatieven, wat misschien niet weerspiegelt hoe individuen in het echte leven kiezen. Als reactie hierop zijn er nieuwe SC-experimenten voorgesteld om meer realistische vormen van besluitvorming op te nemen, zoals Participatory Value Evaluation (PVE). In een PVE-experiment kiezen respondenten een combinatie van alternatieven zonder de beperkte middelen te overschrijden. Ten tweede, terwijl theoriegedreven modellen gebaseerd op nutstheorie, zoals random nutsmaximalisatie (RUM) of Kuhn-Tucker, nog steeds de norm zijn om keuzegedrag te modelleren, is er een bredere erkenning dat het gedrag van individuen uiteindelijk onbekend is vanuit het perspectief van de analist.

Ondanks dit laatste zijn, voor zover de auteur weet, drie methodologische en praktische uitdagingen nog niet opgelost in de literatuur. Ten eerste is er geen onderzoek gedaan naar het potentieel van datagestuurde methoden om gegevens van SC-experimenten buiten DCE's, en in het bijzonder voor PVE-experimenten, te analyseren, hetzij als aanvulling om de specificatie van keuzemodellen te verbeteren, hetzij als zelfstandige methoden voor gegevensanalyse. Ten tweede, hoewel er datagestuurde methoden voor discrete keuzes (en DCE's) beschikbaar zijn in de literatuur, offeren dergelijke methoden ofwel hun flexibiliteit op om te leren van de data om te voldoen aan consistentiehypothesen of vice versa, wat een meer wijdverspreid gebruik van deze modellen in reële beleidstoepassingen belemmert. Ten derde is er een gebrek aan softwaretools om datagestuurde methoden gemakkelijk en handig in te schatten en te vergelijken, wat hun wijdverbreide gebruik belemmert.

Met het oog op deze uitdagingen onderzoekt deze dissertatie verder hoe datagestuurde methoden gebruikt kunnen worden voor het analyseren van individueel keuzegedrag uit SC-experimenten, als aanvulling op theoriegedreven keuzemodellen of als alternatief voor theoriegedreven keuzemodellen; en om methodologische en inhoudelijke bijdragen te leveren voor dergelijke doeleinden. Dit proefschrift richt zijn onderzoek op twee specifieke SC-experimenten: PVE en DCE's.

Hoofdstuk 2 laat de lezer zien hoe PVE-experimenten in het echte leven worden toegepast, hoe ze conventioneel worden geanalyseerd met theoriegedreven keuzemodellen en welke uitdagingen naar voren komen bij het analyseren van PVE-experimenten met deze modellen. In deze PVE-toepassing hebben bijna 30.000 Nederlandse burgers de regering geadviseerd over welke COVID-19 moet worden versoepeld. De gegevens zijn geanalyseerd met een theoriegestuurd keuzemodel gebaseerd op nutstheorie. De modelresultaten laten bijvoorbeeld zien dat burgers evenveel waarde hechten aan het verminderen van 100 sterfgevallen van mensen jonger dan 70 jaar als aan het verminderen van 168 sterfgevallen van mensen van 70 jaar of ouder. Toch suggereren de geschatte parameters dat respondenten een aanzienlijk kleinere waarde toekennen aan de effecten van elke maatregel (weergegeven door smaakparameters) dan de waarde die wordt toegekend aan de maatregelen zelf (weergegeven door beleidsspecifieke constanten). De combinatie van maatregelen waar burgers de voorkeur aan geven, is het heropenen van andere bedrijven dan de horeca en directe familieleden weer sociale contacten laten hebben. De kwalitatieve resultaten laten zien dat burgers PVE als participatiemethode hoog waarderen, maar toch vinden ze dat de overheid bij het nemen van beslissingen voorrang moet geven aan advies van experts.

Hoofdstuk 3 stelt drie procedures voor die gebaseerd zijn op het leren van associatieregels (AR) en random forests (RF) om de specificatie te ondersteunen en de geldigheid van de aannames van theoriegedreven keuzemodellen voor PVE-experimenten te testen. Ten eerste wordt een methodologisch-iteratief (MI) algoritme gecombineerd met AR-leren om relevante interacties tussen gekozen alternatieven van het PVE-experiment te identificeren en op te nemen in een portfoliokeuzemodel. Ten tweede wordt het MI-algoritme gecombineerd met RF om het belang van variabelen te identificeren en te beslissen over het al dan niet opnemen van attributen in het portfoliokeuzemodel. Ten derde wordt RF gebruikt om de geldigheid van de gedragshypothesen van theoriegedreven keuzemodellen te testen. De resultaten tonen aan dat de ondersteunde portfoliokeuzemodellen kleine verbeteringen van de model fit opleveren in vergelijking met niet-ondersteunde modellen. De ondersteunde modellen leveren echter nieuwe gedragsrelevante inzichten op, namelijk relevante interactietermen die kunnen worden geïnterpreteerd als positieve en negatieve synergieën of specifieke attribuuteffecten per alternatief. Tot slot blijkt dat het geassisteerde model met AR-leren dichter bij het "ware" keuzegedrag van het PVE-experiment staat. Deze bevindingen kunnen gebruikt worden door keuzemodelleurs om de specificatie van modellen te verbeteren en door beleidsmakers om nieuwe inzichten te verkrijgen die relevant zijn voor het gedrag.

Hoofdstuk 4 laat zien hoe XGBoost en SHAP - respectievelijk een machine learning model en een verklaarbare kunstmatige intelligentie methode - gebruikt kunnen worden om data van PVE-experimenten te analyseren als alternatief voor theoriegedreven analyse. De analyses worden gedaan met gegevens van een PVE-experiment om de voorkeuren van Nederlandse burgers voor het opnieuw opleggen van COVID-19 maatregelen te bestuderen. De analyses identificeren de belangrijkste (minst belangrijke) covariaten die de steun voor COVID-19 maatregelen verklaren, evenals de waargenomen heterogeniteit tussen respondenten en maatregelen. De resultaten tonen aan dat de steun voor COVID-19-maatregelen vooral samenhangt met de leeftijd van de respondenten, hun gewicht in de schaal leggen bij burgeradvies versus wetenschappelijk advies, en hun waargenomen risico om ziek te worden van COVID-19. Bovendien worden verschillende vormen van waargenomen heterogeniteit geïdentificeerd, zoals clusters van mensen met vergelijkbare voorkeuren of niet-lineaire effecten van elke covariaat. Deze resultaten geven aanzienlijk meer details dan conventionele modelbenaderingen voor PVE, namelijk theoriegedreven keuzemodellen en latente klasse clusteranalyse. Beleidsmakers kunnen deze informatie gebruiken voor het afstemmen van beleid of het opzetten van informatiecampagnes om de steun voor COVID-19 maatregelen te vergroten.

Hoofdstuk 5 stelt het "Alternative-Specific and Shared Weights Neural Network" (ASS-NN) model voor. Het ASS-NN-model is een nieuw discreet keuzemodel op basis van kunstmatige neurale netwerken dat de flexibiliteit om nutsfuncties uit de gegevens te leren in balans brengt met consistentie met RUM en vervangbaarheid van geld (d.w.z. "één euro is één euro"). Het ASS-NN-model wordt getest met Monte Carlo-experimenten en empirische gegevens van de Swissmetro-dataset. De resultaten van het Monte Carlo-experiment laten zien dat het ASS-NN-model met succes de ware nutsfunctievorm terugvindt en nauwkeurig de marginale nutsfuncties en de waarde van de reistijd (VTT) voorspelt. Wanneer de ASS-NN wordt toegepast op empirische gegevens, presteert het beter dan multinomiale logitmodellen in termen van model fit en levert het marginale nutswaarden die consistent zijn met de fungibiliteit van geld-aanname. De ASS-NN voorspelt dat respondenten een hogere gemiddelde

VTT toekennen aan reizen met het openbaar vervoer (d.w.z. trein en Swissmetro) voor reizen met een korte reistijd. Over het geheel genomen is de ASS-NN een veelbelovend datagestuurd alternatief voor theoriegestuurde keuzemodellen dat gedragsrelevante inzichten biedt en tegelijkertijd de voordelen van ANN's om te leren van de data in balans brengt met consistentie met economische aannames.

Hoofdstuk 6 introduceert NP4VTT, een nieuw softwareprogramma dat vijf niet-parametrische modellen biedt om de VTT-verdeling van DCE's met twee attributen en twee alternatieven bloot te leggen. NP4VTT biedt onderzoekers een uniforme syntaxis, waarmee ze eenvoudig en gemakkelijk de VTT-verdeling kunnen schatten met behulp van niet-parametrische modellen. Het gebruik van NP4VTT wordt geïllustreerd aan de hand van gegevens uit het Noorse VTT-onderzoek. De resultaten tonen aan dat de teruggevonden VTT-verdeling consistent dezelfde vorm heeft voor alle modellen en dat verschillen worden toegeschreven aan modelleringsfactoren of stochasticiteit.

Dit proefschrift sluit af door te benadrukken dat, hoewel het primaire onderzoeksdoel en de subdoelen zijn bereikt, de relevantie van de bevindingen en conclusies moeten worden gerelativeerd. Ten eerste leidt het gebruik van datagestuurde methoden tot bescheiden verbeteringen van de model fit. Onderzoekers of beleidsmakers die geïnteresseerd zijn in het voorspellen van gedrag hoeven dus geen aanzienlijke verschillen te verwachten in vergelijking met conventionele keuzemodellen. Ten tweede zijn er nieuwe inzichten gevonden die van belang zijn voor het gedrag. Keuzemodelleurs zouden kunnen profiteren van deze inzichten om de ontwikkeling van keuzemodellen te contrasteren of verder te helpen, en beleidsmakers rekenen met nieuwe en meer gedetailleerde informatie voor besluitvorming. Onderzoekers zullen nu echter moeten ophelderen welke delen van de nieuw gevonden informatie relevant zijn voor het nemen van beslissingen en in hoeverre dergelijke informatie kan worden gesynthetiseerd voor effectieve communicatie. Ten derde, hoewel dit proefschrift meer datagestuurde methoden beschikbaar maakt, zijn er nog steeds uitdagingen om deze methoden toegankelijker te maken voor onderzoekers die gewend zijn aan de concepten en structuur van de keuze-modellering gemeenschap. Concluderend kunnen we stellen dat datagestuurde methoden de inzichten van theoriegedreven keuzemodellen aanvullen, maar niet vervangen. Theoriegedreven keuzemodellen zijn nog steeds een waardevol, eenvoudig en robuust modelparadigma voor het bestuderen van keuzegedrag dat kan worden aangevuld met datagedreven methoden, zoals die in dit proefschrift worden voorgesteld.

José Ignacio HERNÁNDEZ HERNÁNDEZ

# About the author

Jose Ignacio Hernandez is a Chilean economist (Bachelor of Economics) with a masters degree of environmental and natural resource economics of the University of Concepción, Chile. Before his PhD, Jose Ignacio worked as a research assistant and lecturer of microeconomics, econometrics and environmental valuation, with a focus on stated choice experiments to quantify the economic benefits of conservation programmes for ecosystem services.

During his PhD, Jose Ignacio explored the extent that data-driven methods can be made more interpretable and accessible to study stated choices in Participatory Value Evaluation experiments. Aside from his PhD research, Jose Ignacio further contributed on research conducted with several universities and research centers in the Netherlands, such as Erasmus University Rotterdam, Wageningen University, the National Institute for the Public Health and the Environment (RIVM) and Populytics.

# TRAIL Thesis Series

The following list contains the most recent dissertations in the TRAIL Thesis Series. For a complete overview of more than 275 titles see the TRAIL website: www.rsTRAIL.nl.

The TRAIL Thesis Series is a series of the Netherlands TRAIL Research School on transport, infrastructure and logistics.

Hernandez, J.I., *Data-driven Methods to study Individual Choice Behaviour: with applications to discrete choice experiments and Participatory Value Evaluation experiments*, T2023/14, October 2023, TRAIL Thesis Series, the Netherlands

Aoun, J., *Impact Assessment of Train-Centric Rail Signaling Technologies*, T2023/13, October 2023, TRAIL Thesis Series, the Netherlands

Pot, F.J., *The Extra Mile: Perceived accessibility in rural areas*, T2023/12, September 2023, TRAIL Thesis Series, the Netherlands

Nikghadam, S., *Cooperation between Vessel Service Providers for Port Call Performance Improvement*, T2023/11, July 2023, TRAIL Thesis Series, the Netherlands

Li, M., *Towards Closed-loop Maintenance Logistics for Offshore Wind Farms: Approaches for strategic and tactical decision-making*, T2023/10, July 2023, TRAIL Thesis Series, the Netherlands

Berg, T. van den, *Moral Values, Behaviour, and the Self: An empirical and conceptual analysis*, T2023/9, May 2023, TRAIL Thesis Series, the Netherlands

Shelat, S., *Route Choice Behaviour under Uncertainty in Public Transport Networks: Stated and revealed preference analyses*, T2023/8, June 2023, TRAIL Thesis Series, the Netherlands

Zhang, Y., *Flexible, Dynamic, and Collaborative Synchromodal Transport Planning*

*Considering Preferences*, T2023/7, June 2023, TRAIL Thesis Series, the Netherlands

Kapetanović, M., *Improving Environmental Sustainability of Regional Railway Services*, T2023/6, June 2023, TRAIL Thesis Series, the Netherlands

Li, G., *Uncertainty Quantification and Predictability Analysis for Traffic Forecasting at Multiple Scales*, T2023/5, April 2023, TRAIL Thesis Series, the Netherlands

Harter, C., *Vulnerability through Vertical Collaboration in Transportation: A complex networks approach*, T2023/4, March 2023, TRAIL Thesis Series, the Netherlands

Razmi Rad, S., *Design and Evaluation of Dedicated Lanes for Connected and Automated Vehicles*, T2023/3, March 2023, TRAIL Thesis Series, the Netherlands

Eikenbroek, O., *Variations in Urban Traffic*, T2023/2, February 2023, TRAIL Thesis Series, the Netherlands

Wang, S., *Modeling Urban Automated Mobility on-Demand Systems: an Agent-Based Approach*, T2023/1, January 2023, TRAIL Thesis Series, the Netherlands

Szép, T., *Identifying Moral Antecedents of Decision-Making in Discrete Choice Models*, T2022/18, December 2022, TRAIL Thesis Series, the Netherlands

Zhou, Y., *Ship Behavior in Ports and Waterways: An empirical perspective*, T2022/17, December 2022, TRAIL Thesis Series, the Netherlands

Yan, Y., *Wear Behaviour of A Convex Pattern Surface for Bulk Handling Equipment*, T2022/16, December 2022, TRAIL Thesis Series, the Netherlands

Giudici, A., *Cooperation, Reliability, and Matching in Inland Freight Transport*, T2022/15, December 2022, TRAIL Thesis Series, the Netherlands

Nadi Najafabadi, A., *Data-Driven Modelling of Routing and Scheduling in Freight Transport*, T2022/14, October 2022, TRAIL Thesis Series, the Netherlands

Heuvel, J. van den, *Mind Your Passenger! The passenger capacity of platforms at railway stations in the Netherlands*, T2022/13, October 2022, TRAIL Thesis Series, the Netherlands

Haas, M. de, *Longitudinal Studies in Travel Behaviour Research*, T2022/12, October 2022, TRAIL Thesis Series, the Netherlands

Dixit, M., *Transit Performance Assessment and Route Choice Modelling Using Smart Card Data*, T2022/11, October 2022, TRAIL Thesis Series, the Netherlands

Du, Z., *Cooperative Control of Autonomous Multi-Vessel Systems for Floating Object*

*Manipulation*, T2022/10, September 2022, TRAIL Thesis Series, the Netherlands

Larsen, R.B., *Real-time Co-planning in Synchromodal Transport Networks using Model Predictive Control*, T2022/9, September 2022, TRAIL Thesis Series, the Netherlands

Zeinaly, Y., *Model-based Control of Large-scale Baggage Handling Systems: Leveraging the theory of linear positive systems for robust scalable control design*, T2022/8, June 2022, TRAIL Thesis Series, the Netherlands

Fahim, P.B.M., *The Future of Ports in the Physical Internet, T2022/7*, May 2022, TRAIL Thesis Series, the Netherlands

Huang, B., *Assessing Reference Dependence in Travel Choice Behaviour*, T2022/6, May 2022, TRAIL Thesis Series, the Netherlands

Reggiani, G., *A Multiscale View on Bikeability of Urban Networks*, T2022/5, May 2022, TRAIL Thesis Series, the Netherlands

Paul, J., *Online Grocery Operations in Omni-channel Retailing: opportunities and challenges*, T2022/4, March 2022, TRAIL Thesis Series, the Netherlands