



HAL
open science

Performance of econometric and Machine Learning models for the economic study of discrete consumer choices

Nikita Gusarov

► **To cite this version:**

Nikita Gusarov. Performance of econometric and Machine Learning models for the economic study of discrete consumer choices. Economics and Finance. Université Grenoble Alpes [2020-..], 2024. English. NNT: 2024GRALE001 . tel-04620475

HAL Id: tel-04620475

<https://theses.hal.science/tel-04620475>

Submitted on 21 Jun 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : SE - Sciences Economiques

Spécialité : Sciences économiques

Unité de recherche : GAEL - Laboratoire d'Economie Appliquée de Grenoble

Performances des modèles économétriques et de Machine Learning pour l'étude économique des choix discrets de consommation

Performance of econometric and Machine Learning models for the economic study of discrete consumer choices

Présenté par:

Nikita GUSAROV

Direction de thèse :

Iragaël JOLY

MAITRE DE CONFERENCES HDR, Université Grenoble Alpes

Directeur de thèse

Pierre LEMAIRE

PROFESSEUR, Université Grenoble Alpes

Co-encadrant de thèse

Rapporteurs :

André DE PALMA

PROFESSEUR DES UNIVERSITES EMERITE, CY Cergy Paris Université

Maria BÖRJESSON

PROFESSEURE, VTI Swedish Transport Research Institute

Thèse soutenue publiquement le **19 février 2024**, devant le jury composé de :

Nadine MASSARD

PROFESSEURE DES UNIVERSITES, Université Grenoble Alpes

Présidente

Iragaël JOLY

MAITRE DE CONFERENCES HDR, Université Grenoble Alpes

Directeur de thèse

André DE PALMA

PROFESSEUR DES UNIVERSITES EMERITE, CY Cergy Paris Université

Rapporteur

Maria BÖRJESSON

PROFESSEURE, VTI Swedish Transport Research Institute

Rapporteuse

Michel BIERLAIRE

PROFESSEUR, École Polytechnique Fédérale de Lausanne

Examineur

Michel SIMIONI

DIRECTEUR DE RECHERCHE, INRAE centre occitanie - montpellier

Examineur

Invités :

Pierre LEMAIRE

PROFESSEUR, Université Grenoble Alpes



Mr. Nikita Gusarov (2023)

All rights reserved

Abstract

This thesis is a cross-disciplinary study of discrete choice modelling, addressing both econometrics and machine learning (ML) techniques applied to individual choice modelling. The problematic arises from insufficient points of contact among users (economists and engineers) and data scientists, who pursue different objectives, although using similar techniques.

To bridge this interdisciplinary gap, the PhD work proposes a unified framework for model performance analysis. It facilitates the comparison of data analysis techniques under varying assumptions and transformations. The designed framework is suitable for a variety of econometrics and ML models. It addresses the performance comparison task from the research procedure perspective, incorporating all the steps potentially affecting the performance perceptions.

To demonstrate the framework's capabilities we propose a series of 3 applied studies. In those studies the model performance is explored face to the changes in (1) sample size and balance, resulting from data collection; (2) changes in preferences structure within population, reflecting incorrect behavioural assumptions; and (3) model selection, directly intertwined with the performance perception.

Keywords: Data science; Preference studies; Artificial datasets; Econometrics; Machine learning; Consumer choice

Acknowledgements

This PhD thesis was accomplished with financial support from University Grenoble Alpes (UGA) and collaboration of Grenoble Applied Economics Laboratory (GAEL) and G-SCOP.

This work would be incomplete without the assistance provided by members of Data Intelligence Laboratory (LID) at Polytechnique Montréal, as well as the cooperation of multiple researchers from universities of Montréal, Laval and Quebec.

This work was partially financed by MITACS Globalink research mobility program, providing supplementary funds for the duration of stay at Polytechnique Montréal.

The Grenoble Alpes Research - Scientific Computing and Data Infrastructure (GRICAD) equally merits a mention for the infrastructure access and technical support in accomplishment of this thesis. Most of the computations presented in this paper were performed using the GRICAD infrastructure (<https://gricad.univ-grenoble-alpes.fr>), which is supported by Grenoble research communities.

This work would not be accomplished without the support from all surrounding me.

Summary

| | |
|---|-----------|
| Abstract | i |
| Acknowledgements | ii |
| Introduction | 1 |
| 1. Choice modelling: At the intersection of disciplines | 4 |
| 1.1. Current state of Choice Modelling | 5 |
| 1.2. The specificity of the discrete choice modelling | 6 |
| 1.2.1. Choice Analysis in Economics | 8 |
| 1.2.2. Discrete Choice Modelling | 16 |
| 1.2.3. Concluding remarks | 36 |
| 1.3. Taxonomy issues | 37 |
| 1.3.1. Existing taxonomies | 37 |
| 1.4. The vocabulary and terminology | 42 |
| 1.4.1. Models and modelling | 42 |
| 1.4.2. Performance | 45 |
| 1.5. Conclusion | 48 |
| 2. A universal performance comparison framework | 50 |
| 2.1. A need for unified methodology | 51 |
| 2.2. Performance comparison issues | 52 |
| 2.2.1. Scientific procedures | 53 |
| 2.2.2. Performance evaluation | 58 |
| 2.2.3. First framework elements | 61 |
| 2.3. Data constraints and simulation | 63 |
| 2.3.1. Data acquisition | 64 |
| 2.3.2. Experimental design and sources of bias | 68 |

| | |
|--|-----------|
| 2.4. Models and their capabilities | 70 |
| 2.4.1. Statistical models | 73 |
| 2.4.2. Data transformation | 74 |
| 2.4.3. Algorithms | 75 |
| 2.4.4. Software choice | 76 |
| 2.5. Framework presentation | 78 |
| 2.5.1. Data analysis | 79 |
| 2.5.2. Complete framework | 80 |
| 2.6. Existing studies in framework context | 83 |
| 2.6.1. Applied study procedure | 83 |
| 2.6.2. Theoretical innovation introduction procedure | 84 |
| 2.6.3. Theoretical study procedure | 86 |
| 2.7. Concluding remarks | 87 |
| | |
| 3. Framework in action: Case studies | 89 |
| 3.1. Introduction | 90 |
| 3.2. Case 1: Theoretical assumptions | 91 |
| 3.2.1. Introduction | 91 |
| 3.2.2. Methodology | 92 |
| 3.2.3. Results | 96 |
| 3.2.4. Conclusion | 101 |
| 3.2.5. Discussion | 102 |
| 3.3. Case 2: Dataset acquisition | 104 |
| 3.3.1. Research question | 104 |
| 3.3.2. Methodology and context | 105 |
| 3.3.3. Application | 108 |
| 3.3.4. Conclusion | 113 |
| 3.3.5. Discussion | 113 |
| 3.4. Case 3: Statistical modelling | 115 |
| 3.4.1. Introduction | 115 |
| 3.4.2. Performance comparison and model selection | 116 |
| 3.4.3. Application | 124 |

| | |
|--|------------|
| 3.4.4. Conclusion | 130 |
| 3.4.5. Discussion | 130 |
| Conclusion | 132 |
| Glossary | 135 |
| Acronyms | 135 |
| Special terms | 138 |
| Tables | 141 |
| List of Figures | 141 |
| List of Tables | 142 |
| Index | 144 |
| Bibliography | 146 |
| Appendices | 167 |
| A. Bibliometric study | 167 |
| A.1. Motivation and research objectives | 168 |
| A.2. Data collection and preliminary analysis | 168 |
| A.3. Advanced analysis | 172 |
| A.3.1. General information | 172 |
| A.3.2. Keywords | 173 |
| A.3.3. Co-occurrences | 177 |
| A.3.4. Citations | 179 |
| A.4. Analysis by subdomain | 182 |
| A.4.1. Policy | 183 |
| A.4.2. Preferences or attitudes | 186 |
| A.5. Conclusion | 188 |
| B. Extracting economic information from Neural Networks | 190 |
| B.1. Statistical and Machine Learning perspective | 191 |

| | |
|--|------------|
| B.2. Introduction to Neural Networks | 191 |
| B.2.1. Artificial Neuron and Perceptron | 191 |
| B.2.2. Adaline | 195 |
| B.2.3. Multilayer Perceptron | 196 |
| B.2.4. Convolutional Neural Network (CNN) | 202 |
| B.3. NN in Choice Analysis | 202 |
| B.3.1. CNN design for MNL imitation | 204 |
| B.3.2. Alternative Utility Specific DNN (ASU-DNN) | 206 |
| B.3.3. Extracting interpretable information from NN | 207 |
| B.4. Conclusion | 209 |
| | |
| C. Independence from Irrelevant Alternatives | 210 |
| C.1. Traditional formulation of IIA | 211 |
| C.2. History and ambiguity of the IIA | 212 |
| C.2.1. IIA(A) by Arrow (1951) | 212 |
| C.2.2. IIA(RM) by Radner and Marschak (1954) | 213 |
| C.2.3. IIA(L) by Luce (1957) | 213 |
| C.2.4. Contraction consistency by J. F. Nash (1950) | 214 |
| C.2.5. Criticism of the minimax decision theory by Savage (1951) | 214 |
| C.3. Linking the IIA with reality | 214 |
| C.4. Mitigation of the IIA inconsistency | 215 |
| C.4.1. IIA tests and validation | 215 |
| C.4.2. IIA treatment | 217 |
| C.5. Conclusion | 218 |
| | |
| D. Research practices: Unstructured interviews | 219 |
| D.1. Motivation and research objectives | 220 |
| D.2. Methodology | 221 |
| D.2.1. Target audience | 222 |
| D.2.2. Survey | 222 |
| D.3. Results | 224 |
| D.4. Conclusion | 225 |

| | |
|--|------------|
| E. Software packages | 227 |
| E.1. ‘dcesimulatr’: A DCE simulation toolset | 228 |
| E.2. ‘performancer’: Performance estimation functions collection | 228 |
| F. Synthesis in French | 229 |
| Introduction | 230 |
| F.1. Modélisation du choix : À la croisée des disciplines | 232 |
| F.2. Un framework universel de comparaison des performances | 234 |
| F.3. Le framework en action : Études de cas | 236 |
| F.4. Conclusion | 238 |
| Table of contents | 241 |

Introduction

With the development of computational devices and increasing data availability more novel and resource heavy data analysis methods are introduced. In particular, the advances in statistical learning (Hastie, Tibshirani, and Friedman 2009) and data science (Donoho 2017) of the past decades have resulted in propagation of Machine Learning (ML) techniques in application to resolution of economic problematic. The most resource demanding models of previous decades can be executed in several minutes and current research is more and more focused on the big data and analysis automation, be it for policy evaluation tasks or economic agent behaviour modelling. The number of available data analysis strategies may make it complicated for the non-experts to select the optimal solution (Athey and Imbens 2019). For example, even the seemingly banal case of binary consumer choice data analysis may be approached with tools starting from basic sample differences tests, to the more sophisticated regression analysis and finally to the complex supervised classifiers with implementation of boosting algorithms. Each of the enumerated options has its own advantages and weaknesses and inexperienced user may easily overlook some of those elements. To address this issue there is an important need for better understanding of the various models' strength and weak points.

However, addressing the general issue of model performance comparison without any particular context would be extremely difficult. In economic disciplines there exist multiple application scenarios and use-cases, each having extremely specific requirement in terms of a toolset selection. Such fundamental work would require an extensive knowledge of both the models and the economic application specificities, as the model usage can rarely be analysed without any application context. What is more, every year the number of available models grows as more and more complex tools addressing narrow use-cases emerge, which puts the creation of a unified compendium of all the available models outside the scope of any limited in time study. In order to limit the scope of our study, we will focus our attention on the discrete choice model family in the context of the individual choice studies.

This limitation will establish a baseline for the discussion. The choice modelling focuses on the exploration of the behaviour analysis, be that individual or an other type of decision maker. It frames a rather limited, compared to all other available model families, number of techniques. Those may be summarised as the Classification methods using the Statistical Learning (SL) terminology.

At this point it is important to outline the key problematic and difficulties associated with the interdisciplinary model performance comparison task. Those may be separated into two main groups: (1) technical complexities of the available toolset implementation and usage; and (2) conceptual differences imposed by the heterogeneity among the use-cases and users.

First of all, we operate under assumption that the available toolset descriptions may appear to be extremely complex for non-proficient users. In other words, we assume that every model outside the undergraduate or graduate level of expertise might require a learning effort from the target audience. To justify this assumption let us take a look at the presentation of one of the baseline models widely

used for choice analysis nowadays - the Multinomial Logit (MNL) backed up by the Random Utility Maximisation (RUM) framework. While nearly every handbook available presents this tool in a guided and accessible way (Agresti 2013), the original work introducing this toolset (McFadden 1974) is far more complex to inexperienced readers. The modern tools require an advanced expert knowledge to be used and the up to date literature is mostly oriented towards the proficient users.

Effectively, it is possible to encounter some technical notes or guides, which attempt to fill the existing gap between most recent scientific and baseline educational literature, forming a layer of *advanced* knowledge sources. Nevertheless, such *advanced* supports rarely provide enough information to the reader and typically are biased or incomplete. A very interesting illustration of this may be drawn from the attempts to implement the machine learning methodology in economic studies. For example, in the works of Athey and Imbens (2019) or Mullainathan and Spiess (2017) we encounter the *guidelines* for economists on the usage of machine learning toolset. However, while in both cases the publications provide interesting discussions on machine learning toolset usability for economists, both miss out the learning curve steepness for economists unfamiliar with those advanced techniques.

Secondly, we may observe an extreme ambiguity and inconsistency in the vocabulary varying by community. The different domains and branches of science, even though using quite similar tools, may have different comprehension of the theoretical implications behind them. The most basic example in this case would be the line drawn between the *classification* and *discrete choice analysis* tasks. While the toolset implemented for those tasks are typically nearly identical (Agresti 2007; Hastie, Tibshirani, and Friedman 2009), the conceptual differences make it relatively difficult to merge the available knowledge onto a common support. While in both cases the handbooks introduce relatively similar concepts, among which the binary and multinomial logistic regressions, the presentation varies drastically. The introduction of other potential applications for seemingly identical toolset as the preference modelling (Fürnkranz and Hüllermeier 2010) or choice analysis in health economics (Soekhai et al. 2019) only increases the number of divergent terminologies.

Moreover, not only the practical side differs, but the most basic terms may be understood under different perspectives. One of the most speaking illustrations in this case is the ambiguous *model* term, which may be understood differently depending on the context. Theoretical, statistical, mathematical, econometric and economic *models* appear in the literature and all of them might be referred as simply *a model* given the specificity of work. For example, the work of Sfeir, Rodrigues, and Abou-Zeid (2022) have a term *model* present directly in the title of the publication: “*Gaussian process latent class choice models*” - referring to the family of the statistical choice models. The same can be said about the work of El-Badawy, Elharoun, and Shahdah (2021): “*Captivity impact on modelling mode choice behaviour*”. However, this time the delimitation of the *model* term is more ambiguous, as it remains unclear from the title whether it concerns the theoretical models of choice behaviour or the statistical side of the question. A more complex example may be drawn from the work of Lee, Derrible, and Pereira (2018), where different configurations of neural networks are compared with the Multinomial Logit *model*. While the MNL *model* is relatively well delimited in the literature, the neural networks part allows for more flexibility in the definition choice due to its particularly complex modular structure.

With this work, we attempt to organise the existing knowledge from different domains in cross-disciplinary study. The **Chapter 1** of the work is consecrated to the definition of the study’s scope. It overviews the various miss-understandings which appear when the different disciplines are united. The definitions of the vocabulary to be used latter will be given. It includes as well some preliminary insight into the

history of both: (1) tool development and (2) choice theory. **Chapter 2** shifts the focus over the performance comparison task. We address one by one the major elements of the scientific procedure which may potentially impact the observed model performance as well as the perception of the later. From the research question target metrics to the data collection and model selection, we overview each of the steps that may potentially impact the perceived performances. This overview will bring us to the definition of novel performance comparison framework. Finally, in **Chapter 3** we offer a selection of case-studies, accompanied by a reflection on their implementation in relation to the proposed framework. Those are mostly conference papers making use of the performance comparison framework to address and explore the different issues in choice modelling.

Giving a particular focus to the individual choice modelling this work contributes to the economic literature, specifically to the experimental economics literature on the discrete choice experiments data analysis. The proposed framework for model performance assessment and analysis may be further extended outside the direct scope of this study through a series of generalisations. Theoretically, the proposed solution may be equally implemented in different application scenarios. Among which specifically: (1) health economics with the extensive usage of choice modelling and discrete choice experiments; (2) marketing with the focus on optimisation and preferences analysis; (3) economics of innovation, focusing on individual preferences for innovative goods and services; and (4) strategic decision making analysis in the context of industrial economics, as the choice modelling may be extended to other subjects.

1. Choice modelling: At the intersection of disciplines

Choice modelling is employed in economics to analyse and understand individuals' decision-making processes when faced with various alternatives. The choice models help economists and policymakers to perform demand forecasts, tariffication optimisation, design of marketing strategies and public policies optimisation in general. By capturing the factors that influence decisions and estimating the trade-offs individuals make, choice models provide valuable insights into economic behaviour, enabling researchers to predict responses to changes in policies, prices, and other variables.

The *Discrete Choice Modelling (DCM)* answers most of the objectives of economists. They are fundamental tools for studying individual behaviour in economic settings and are widely used to inform policy decisions, market strategies, and economic research. Nevertheless, with the emergence of novel statistical approaches to supervised classification and the increasing popularity of interdisciplinary studies more and more studies attempt to use the new techniques for choice modelling. Modern computing allows models that were highly resource-intensive in the past to be executed within minutes today. Current research increasingly emphasizes large datasets and the more complex analysis procedures.

Unfortunately, such increasing number of available data analysis strategies may make it extremely difficult for the non-experts to select the optimal solution. Moreover, the heterogeneous background of those modelling techniques, as well as differences in vocabulary make it particularly difficult to select the best option for a non-proficient user. To address this issue there is a need for a better understanding of the various models' strength and weaknesses face to different economic questions. The key element for this task is the ability to compare and contrast the modelling approaches.

This chapter introduces the reader to the general problematic associated with the consumer choice modelling task in the context of economic studies. The **Section 1.2.2** offers an overview of existing toolset, both technical and theoretical. The following **Sections 1.3 and 1.4** shed light onto other difficulties related to the interdisciplinary context of the study, addressing vocabulary and terminology related issues.

1.1. Current state of Choice Modelling

Discrete Choice Analysis (DCA) is an ensemble of quantitative research techniques used to analyse and predict the individual behaviour in choice based tasks (K. Train 2002). It is widely spread across many research fields such as economics (Durlauf and Blume 2010; Athey and Luca 2019), health (Mühlbacher and Bethge 2015), marketing (Coussement, Benoit, and Poel 2010), transportation (Guevara and Ben-Akiva 2013), and environmental science (Daziano and Achtnicht 2014). Regardless of the context it is mainly used to understand individual preferences and decision-making processes, be it the choice of a transportation mode or the preferences for particular attributes within available products. Choice modelling typically involves designing surveys or experiments where respondents make choices among different options, and the collected data is used to estimate models that reveal the underlying preferences and trade-offs individuals consider when making decisions (Ben-Akiva, McFadden, and Train 2019). This approach provides valuable insights for businesses (Bode, Macdonald, and Merath 2022), policymakers (Mihailova et al. 2022), and researchers (Fifer, Rose, and Greaves 2014). It helps them to make informed decisions, develop effective strategies, and understand the drivers behind individual's choices.

Recently the traditional DCM started to adopt some of the complex modelling techniques from the *Machine Learning (ML)* discipline (Hillel et al. 2021; Aboutaleb et al. 2021). This convergence of methodologies has enriched DCM toolset by enhancing its predictive capabilities and expanding its applicability (Danaf et al. 2019). While this fusion of disciplines represents a promising avenue in today's data-rich, complex decision environments, some complications arise. The growing array of data analysis strategies can pose a significant challenge for researchers without expertise in the field, making it increasingly challenging to choose the most suitable solution. Following the results of in-person interviews with practising researchers conducted during this PhD work, it appears that there exist two main strategies in model selection: (1) the searchers apply the models with which it is interesting to work for them; or (2) they use the models with which they are familiar enough to accomplish the given task. While this reasoning hold for experienced researchers, the novices may be limited in the modelling strategy choice even further. Typically this leads them to follow the most main-stream modelling strategies, potentially without a complete understanding of underlying processes. The last choice rendered even more difficult by the diversity of the modern scientific literature on choice modelling and classification techniques.

The diversity in backgrounds among the available modelling techniques and variations in terminology further complicate the process of selecting the right option for casual users. For example, depending on the familiarity with one or another discipline the scientist will search either for *classification* or *choice modelling* techniques. To address this issue there is a need for a better understanding of the various models' strength and weaknesses face to different economic questions. Central to this task is the capability to compare and differentiate between available modelling approaches. As while the comparison of seemingly closely related choice models is relatively easy due to their similar structure, the comparison with completely different methods is much more difficult.

The performance assessment is usually performed in academic papers proposing some novel models or alternative estimation techniques. This tightly interlinks the concept of performance to the model itself. Applied studies tend to adopt a more cautious approach when presenting the procedure and outcomes of performance assessment. Typically, only the most promising model makes it to the pub-

lication or production stage. However, some methodological works, mostly in econometric oriented studies, stay away from this paradigm and explore individual effects elicitation (M. Bierlaire, Bolduc, and McFadden 2008) or the ability to correctly derive some composite metrics (Rose and Bliemer 2013). This accentuates the discrepancy present in the literature. Moreover, the performance concept is far from being the only ambiguous term in the literature. The definitions of the simplest concepts such as *model*, *Machine Learning* or *scientific procedure* may be understood differently depending on the background of the reader in the context of an interdisciplinary work. Therefore, dedicating one of the introductory sections to specifying the terminology is essential.

However, such task cannot be carried out without any prior on the application field. The application domain as well as the dominating associated literature would outline the basic principles for the terminology specification. At the same time, the associated fields of interest will influence our definitions, shaping and adjusting them. The current state of the existing literature underscores the imperative to meticulously systematize the terminology that will be employed throughout this manuscript.

This first chapter is dedicated to the introduction of fundamental concepts and terminology that will be employed in subsequent sections of this manuscript. Starting with a basic introduction to DCM discipline, this chapter establishes a baseline for further discussion of the performance comparison. The eventual differences between the application domains (economics, management, sociology), as well as the different epistemology paradigms (Machine Learning and econometrics) will be outlined. The issues of model taxonomy construction in the context of an interdisciplinary work will equally be addressed. Finally, the model concept definition complexity will be introduced transitioning to the performance comparison task complexities presentation.

1.2. The specificity of the discrete choice modelling

Before delving into the core discourse, it is imperative to provide the reader with a foundational understanding of the Discrete Choice Modelling (DCM), with a specific focus on consumer choice modelling and DCM in general. In economics, the history of the individual decision making modelling may be traced as far as Luce (1957) works, and the later introduced state of art by McFadden (1974). However, it is essential to acknowledge that other disciplines, such as management studies and sociology, occasionally diverge from the economic perspective on this subject. Furthermore, disparate knowledge acquisition strategies influence the available toolkit and conceptual framework for data analysis.

In practice, the DCM is not limited to the economic applications, but extends further and finds place in biology or geo-sciences related studies. In relation to the economics, the choice modelling techniques are known to be applied in transportation research (Ojeda-Cabral, Hess, and Batley 2018), health economics (Reckers-Droog, Exel, and Brouwer 2021) and social sciences (Hilhorst et al. 2022) in general¹. The broader generalisation of discrete modelling as classification techniques may be encountered in even more diverse set of disciplines and applications: fraud detection (Baesens, Höppner, and Verdonck 2021), image classification and many more (Gong, Zhong, and Hu 2021).

The common points across all of the applications that make the choice modelling to stand out from a more general classification oriented research is the context of the individual behaviour modelling. This narrowing of problematics typically adds a number of theoretical assumptions and restrictions to

¹Here we unite the sociology and psychology related studies, focusing on human behaviour modelling and exploration.

account for. For example, the data used for such applications has a very specific structure, representing a combination of the decision maker and the alternatives. While more general classification tasks do not offer any distinction on the inputs, the discrete choice modelling task typically fractions inputs separating them onto the individual and alternative related. The response variable in the discrete choice modelling is typically constrained in a discrete output space

Regardless of the restrictions imposed by DCM context, the available modelling strategies are numerous. Each application scenario counts several rather specific models, which rely on the in-depth understanding of the respective domains and the associated problematic. For example, in Preference Learning (PL) (Fürnkranz and Hüllermeier 2010) models are mostly focused on elicitation of the order effects and ranking of alternatives. The mode choice analysis (Hall and Hillier 2003) is focalised on the exclusive choice representation at the same time. This perfectly illustrates as seemingly closely related tasks. Such models are often referred as theory driven analysis methods. There exist a number of classification techniques that are common across all the fields, but usually they lack the flexibility for each task specific application. Typically such general models are implemented with a focus on the data, without an extensive exploration of underlying theoretical implications, assuring the label of data driven methods for this category. This second category nowadays incorporates the more complex ML methods as well (Bai 2022; García-García et al. 2022).

Introduce above contrast between the works of Fürnkranz and Hüllermeier (2010) and Hall and Hillier (2003) underscores an important point in the DCM applications. While the commonalities exist in DCM implementation across disciplines, the differences arise due to the specific contexts and objectives of each use-case. Among the common elements we may identify the focus on individual, or agent², performing the choice under a set of behavioural assumptions. Those common elements may usually be traced to the works of Luce (1957) on individual decision making. The second common element, as stated previously, includes the usage of rather specific data format, containing information on both the decision maker and the available alternatives (McFadden 1974). Finally, the reliance on decision theory is justified by the desire to obtain some interpretable insights into underlying decision-making process.

The differences in DCM application are typically related to the application cases. The economics is often focused on demand forecasts, market behaviours (Ndebele, Marsh, and Scarpa 2019), and policy implications (Janssen and Hamm 2012). Sociology explores choices influenced by societal and cultural factors (Mouter et al. 2021), studying collective decision-making. The environmental and health related studies in contrast shift focus to the resource³ usage problems, investigating choices related to public health (Walrave, Waeterloos, and Ponnet 2020), environment and sustainability (Hannus 2020). Those differences in research questions explored affect the nature of both explicative and outcome variables. While marketing applications might be focused on the ordered choice situations or alternative rankings, the transportation research typically focuses on exclusive choice situations. The requirements to the model precision also drastically vary across the disciplines. While in some general cases a baseline DCM models might suffice, the health related applications might have particular requirements in terms of model predictive precision, due to the increased cost of error (Huls and de Bekker-Grob 2022). The same applies to the different requirements to the confidence intervals of the

²As some studies may focus on corporate decision making, exploring the managerial strategic decision making (Haile, Tirivayi, and Tesfaye 2020).

³Be it human or environmental resources.

identified effects: while some studies might be completed with plain effects identification (Gundlach et al. 2018), some public policy related cases may require greater precisions, for tariffication purposes for example (Dubernet and Axhausen 2020).

In this section, the focus shifts to the history of choice modelling and classification. First there is a brief historical overview of the development of choice modelling, along with an exploration of prevailing trends in the broader choice modelling literature. Subsequently, an overview is provided for the prevalent modelling techniques found in contemporary literature, most of which are introduced in their standard forms. This section serves as a presentation of the *state-of-the-art*, acquainting the reader with naming conventions and facilitating familiarity with diverse modelling approaches.

1.2.1. Choice Analysis in Economics

Discussing economics related DCM toolset in the context of interdisciplinary research poses inherent challenges, particularly in the contemporary context where interest in interdisciplinary studies has drastically increased. This complexity is further complicated when addressing domains characterized by extensive usage of statistical toolset. In this part of the work we are going to focus on the interdisciplinary dimension of the DCM studies, attempting to link those elements in the context of the economic studies. We are going to observe the different DCM related modelling strategies that exist in the literature, as well as the different underlying theories of individual behaviour that support those methods.

Typically, the statistical approaches may be roughly divided into two classes nowadays: (1) data driven and (2) theory driven methods. The discussion about such separation may be traced up to Varian (1994) and Breiman et al. (2001). In more traditional terminology, the data driven methods are typically united under the term Machine Learning in these works, while opposed to the theory driven techniques (typically separated into field-specific disciplines such as: econometrics, psychometrics, biometrics or sociometrics). Nowadays, specifically in economics, this problematic becomes more and more discussed. All this due to the fact that previously complex statistical solutions become easily available through user-friendly software, such as Stata, or high-level accessible programming languages, among which R, Python or Julia. Among the recent works offering this discussion we can refer to: Varian (2014), Lipton (2017), Agrawal, Gans, and Goldfarb (2019), Athey (2018) (and further Athey and Imbens (2019)) or Haghani et al. (2021a). However, even given the numbers of emerging studies, we encounter a lack of consensus on what are the differences of the two paradigms, which is confirmed by some of the meta-studies (Hillel et al. 2021). This is only accentuated by the number of studies attempting to merge the practices: Q. Wang et al. (2020), Vijayakumar and Cheung (2019), Ortelli et al. (2021), Aboutaleb et al. (2021) and Ish-Horowicz et al. (2019). Moreover, there already existed some registered attempts to replace traditional tools (such as Multinomial Logistic Regression or MNL) by Machine Learning even in early 2000 (Bentz and Merunka 2000). These attempts make it even more complex to trace a clear line between the two paradigms.

The task of describing the available techniques in DCA is far more complex than that. Such situation arises because the statistical toolset, as well as the available software solutions, are tightly linked to their field of application. This creates a rather closed scientific community around it. For example, in the specific context of DCM, various academic disciplines concentrate on what may appear to be closely related objectives. Economics (Agresti 2007; Green 2018), marketing (Bentz and Merunka 2000;

Coussement, Benoit, and Poel 2010), sociology (Molina and Garip 2019), psychology (Hurtubia et al. 2014; Guillon 2020) and preference learning (Fürnkranz and Hüllermeier 2010; Domshlak et al. 2011; Pigozzi, Tsoukiàs, and Viappiani 2016; Erişkin 2021) - all of these fields are closely related in their selection of modelling techniques. Nevertheless, with time the vocabulary diverges across application domains, alongside with the overall focus of the models and particular tasks and objectives. Even though nearly all of the listed fields know quite well the existence of works of Luce (1957) or McFadden (1974), the modern modifications of these techniques significantly vary.

As this study predominantly centers on applications in individual behaviour modelling and economics, our primary focus will be on discussing the contemporary status of statistical modelling within this particular discipline. The first discipline related issue⁴ resides in the existence of multiple potential modelling objectives in each of the application fields. In economics we can encounter the tasks of aggregated modelling⁵ or the exploration of economic indicators (Wong, Yang, and Szeto 2021; Reckers-Droog, Exel, and Brouwer 2021). The later one can further be partitioned into multiple categories depending on the indicators' nature: be that (1) simple coefficient estimate (Ludwig et al. 2021), or (2) a rather complex composite estimations requiring simulation or further estimate transformations (Michaud, Llerena, and Joly 2012; Scholz et al. 2015; Hynes et al. 2021; Goff 2021). This difference in objectives behind modelling procedure poses some interesting questions about statistical tools. For instance, it becomes extremely difficult to create a unified taxonomy for all the available techniques and models: (1) some models may be used to answer several questions at a time with limited efficiency, while (2) other are designed to fulfil only a limited number of specific research objectives.

As evident, the task of conjoint analysis of the diverse tools is extremely delicate. These tools are typically designed to address specific types of inquiries, and each field of application possesses its unique set of terminologies and practices for employing statistical methods, often evolving to respond to domain-specific issues. Consequently, it is impractical to directly compare two entirely dissimilar statistical tools outside of any context. Although there is a number of attempts to compare the different models performances in particular use cases: Baldi et al. (2000), Bodea and Garrow (2006), Karlaftis and Vlahogianni (2011), Askin and Gokalp (2013), Hrnjic and Tomczak (2019), Schulz, Speekenbrink, and Shanks (2014), Mohammadi et al. (2021) and many more. Some of the works use simulated data for performance testing purposes, some make focus primarily on the real world datasets. However, the typical restriction of those studies resides in the case specific, limited testing procedure.

Having established the study's context, the attention can be directed to the specific modelling techniques frequently employed in this field. In the upcoming subsections, the goal is to offer a detailed and extensive review of the *current issues, questions, theoretical models* and *statistical modelling approaches*. In the following subsections, the aim is to provide the reader with a comprehensive and thorough literature review of the existing The aim is to familiarise the reader with the DCM background and outline the key components which are essential for the further reading.

1.2.1.1. Application fields, problematic and research questions

To offer the most comprehensive perspective on the subjects under study in the academic literature, the issues addressed through DCM in particular, an extensive literature review procedure was imple-

⁴Here we speak about *discipline related* assuming that this issue is common across multiple disciplines and not exceptionally to econometrics.

⁵For example, aggregated demand modelling (McFadden 1974).

mented. This literature exploration takes its roots in existing studies, addressing usage of ML toolset in economics and Choice Modelling (CM) in particular. Such limited literature study approach relies on the existing bibliometrics reviews and the state of art works in the various application fields. With exception of the works of Athey et al. (2019) and Mullainathan and Spiess (2017), providing the general culture on the ML methodology implementation in economics studies, one of the key works to consider is the bibliometrics review performed by Haghani, Bliemer, and Hensher (2021).

In their work Haghani, Bliemer, and Hensher (2021) provide an overview of the landscape of econometric discrete choice modelling research. The study includes works detected on Web of Science (WoS) from 1900 to 2020 (27 July 2020), with a total of 14237 items, including articles, books and book chapters. As stated in the work “*a set of common candidate terms that characterise discrete choice modelling studies were listed and suitable combinations that did not produce false positives were subsequently chosen*” in an iterative manner. Meaning that candidate terms were explored one by one and if false positive were numerous then combined with (“*choice*” OR “*preference*”) to minimise them. This approach is rather limited as pointed out by the authors, they were unable to explore a full spectrum of the available topics due to resource limitations. The novel choice modelling techniques, such as: *decision field theory* (Hancock, Hess, and Choudhury 2018), *quantum probability* (Hancock et al. 2020), *prospect theory* (Kahneman and Tversky 2012) or *game theory* (Austen-Smith and Banks 1998) - were kept out of scope, because they rarely use the same keywords as more classical choice modelling oriented works. Another limitation was that application of *discrete models* was considered only for *choice* problems and applications. Leaving all the applications such as *vehicle crush analysis* or *classification* tasks out of scope. All these studies have their own limitations, but offer valuable information for each of the selected topics under assumptions of the authors’ expertise. Among the identified in the cited work fields we encounter: (1) *transportation*, (2) *health economics*, (3) *environmental studies* and (4) *consumer studies*. The findings introduced by Haghani, Bliemer, and Hensher (2021) frame the following part of the literature exploration procedure, where the focus is shifted to an independent exploration of the literature.

The second part involves exercise of an automated bibliometrics review using a relatively similar approach. Nevertheless, the resulting analysis is not a simple replication of existing systematic literature reviews, but rather a self-contained bibliometrics study. Through automated procedure of Web of Science database querying and clearly defined filtering rules the most relevant research topics and questions⁶ are identified. This study is limited due to several problems encountered in filtering rule definition procedure. However, when combined the two stages of literature review produce reliable enough results for the purposes of this thesis. The final dataset consists of 65654 items, which are analysed simultaneously. The inclusion of the start of the first months of the year 2022 induces some potential biases into the replicability of our research, because it makes it more difficult to obtain the same results as newer publications appear. However, we make an assumption that such new publications should not affect our conclusion in a significant manner.

While focusing on the journal and publisher supplied information we obtain quite expected results. As we can see in the Table A.3, the main publishers are: (1) *Elsevier* regrouping the publications related to economics, management and transportation; (2) *Springer Nature*, which encompasses publications related to ecology and biology oriented articles; (3) *Wiley* and (4) *Taylor & Francis*. Those key publishers are followed by *Sage*, *Mdpi*, *Emerald Group* and *Oxford University Press*, each amounting for more

⁶A detailed version of this bibliometrics study is available in Appendix A.

Table 1.1.: Sources composition

| (a) By publisher | | (b) By discipline | |
|-------------------|-------|--|-------|
| Publisher | Share | Discipline | Share |
| Elsevier | 22.20 | Economics | 8.04 |
| Springer Nature | 11.50 | Environmental Sciences | 6.24 |
| Wiley | 10.68 | Transportation | 5.04 |
| [h] T&F | 7.87 | Statistics Probability | 4.88 |
| Sage | 4.43 | Ecology | 4.39 |
| Mdpi | 3.68 | Environmental Studies | 4.09 |
| Emerald Group | 2.27 | Public Environmental Occupational Health | 3.89 |
| Oxford University | 1.95 | Geoscience Multidisciplinary | 3.44 |
| W&W | 1.34 | Transportation Science Technology | 3.44 |
| IEEE | 1.27 | Management | 3.37 |

than 1000 items in the dataset. As we can see in the Table A.3b, even though we have excluded a significant number of articles oriented towards biology and natural sciences, our final dataset regroups a lot of publications oriented towards: *Environmental Sciences*, *Ecology*, *Geoscience* and *Environmental Occupational Health*. Our filter puts *Economics* onto the first place, alongside with tightly related disciplines such as: (1) *Transportation*, (2) *Management*, (3) *Business* and (4) *Sociology*; the last two not entering into the first 10 disciplines. Because our main research pattern focuses on the statistical tools, we expectedly encounter among the dominating publication domains the *Statistics and Probability*, followed by *Operations Research* and *Computer Science*. Please note, that we show in the corresponding tables only the first and most prominent entries of the corresponding lists⁷.

The identified fields correlate with those identified in other bibliometrics studies. The main difference is that in our case the Consumer studies are divided into *Economics* and *Marketing*, because the domains differ by their approach to modelling, treated questions and focus of their studies. The further division may be traced with the *Economics* field, switching the focus either towards the *industrial economics* or *consumer studies*. With other points there may be some ambiguity, because some of the *Environmental studies* may be considered as economics research, while other are more correlated with *Geoscience* domain or even *Biology* (when the focus is made on biodiversity). For *Economics* related studies, among the potential topics of interest we encounter:

- Individual choice modelling in general
- Policy making
- Preference studies
- Market analysis
- Attitudes assessment
- Demand modelling (Aggregated demand modelling)
- Modelling of economic agent's behaviour: individuals (students, respondents), households, firms, companies

To the traditional purely economics problematic, we may add *Sociology*, *Psychology* and *Transportation* topics. Those are domains which contrary to biology and geoscience studies have a closely related methodology to our main topic: discrete choice analysis of behaviour. The cluster combining the first

⁷The complete datasets might be found on GitHub.

two topics focuses primarily on the *causal effect detection*. The articles included into our sample address topics of various *Disorders* and *Depression*, as well as *Disability*, *Violence* and *Peer effects* among the individuals. The last one, *Transportation* related cluster, focuses on *Crash* detection and related policies and *Traffic* analysis. We can assume that those research is mostly oriented on public policy proposals and adoption, which makes this cluster potentially interesting for us.

Finally, the keywords may be regrouped into more general topics. For example:

- *Model* and *Performance*, which underline the topic of modelling in general, as well as the particular target in prediction tasks
- *Impact* and *Determinants*, which represent another facet of modelling objectives, focusing on the explanation and causal effects understanding
- *Behaviour*, *Attitudes* and *Willingness to Pay (WTP)* - those regroup the topic of understanding of the individual behaviour, which is rather common in choice modelling
- *Management*, which stands aside from other *Economics* related disciplines, although the explored questions and used techniques are closely related
- *Demand*, which unites the market analysis in general

For general analysis we explore the citations count on the single document level. Thus we will be able to exclude the less cited works from our future analysis, which inevitably excludes some of the most recent works as well. The number of documents in our collection is at 65654 articles at this stage. We define the minimal citation number limit at 0.1% level of all documents (rounded to 66 citations), which drastically reduces our document selection to mere 5741 works. The most notorious 1000 works, based on weighted link strengths, were selected and the main cluster containing 971 document is explored.

The Figure A.7 offers an overview of the citation map using a density representation. This map allows us to detect the most prominent clusters and dependencies among the cited works. In the center we encounter the biggest cluster of *Epidemiology* and *Biology* related modelling articles, which focus on different ecological and environmental questions. For example, while Friedman (2001) focuses on technical aspects and proposes a gradient boosting method for model estimation, Allouche, Tsoar, and Kadmon (2006) focuses on more applied question related to accuracy of species distribution models. Dormann et al. (2013) explores the ways to combat the collinearity, and Firth (1993) proposes a methodology for bias reduction in maximum likelihood estimates.

The “branches” distancing from the central cluster are more discipline specific. On the left side we encounter the cluster related to *Geoscience*: Ayalew and Yamagishi (2005) describes a GIS-based logistic regression for landslide detection. In the upper part of the figure we encounter more advanced *ML* techniques in application to the engineering and technical disciplines: Chen et al. (2014) uses Deep-Learning techniques for classification of Hyperspectral data. Finally, the cluster representing the most interest for us is rightmost branch: McFadden and Train (2000) introducing the Mixed MNL models for discrete response data analysis, which is one of the key works in *Choice Modelling*.

Let us focus on economics related cluster as depicted on Figure A.8. As we can see the works of McFadden and Train (2000), Albert and Chib (1993), Hausman, Hall, and Griliches (1984) and Greene and Hensher (2003) for the gravity center on the figure. Those are the most cited works in this cluster.

To better understand their nature we should probably explore the topics addressed by those works.

⁸The colors on graph are used to separate the entries belonging to different clusters.

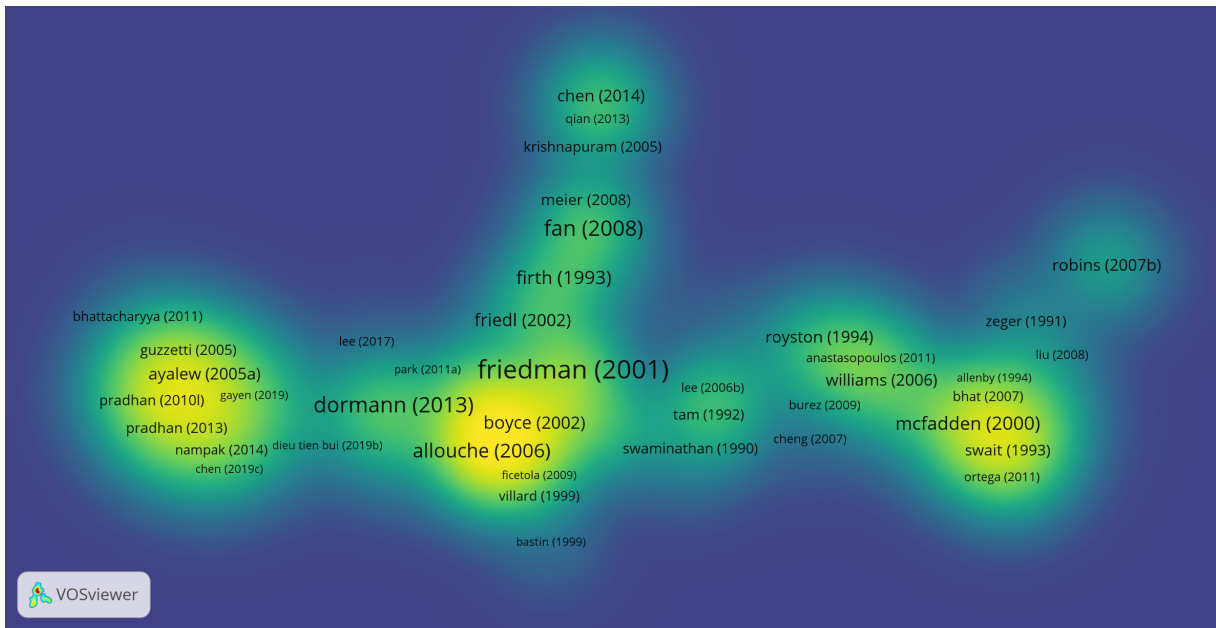


Figure 1.1.: Citation map on document level (1975 - 2021)

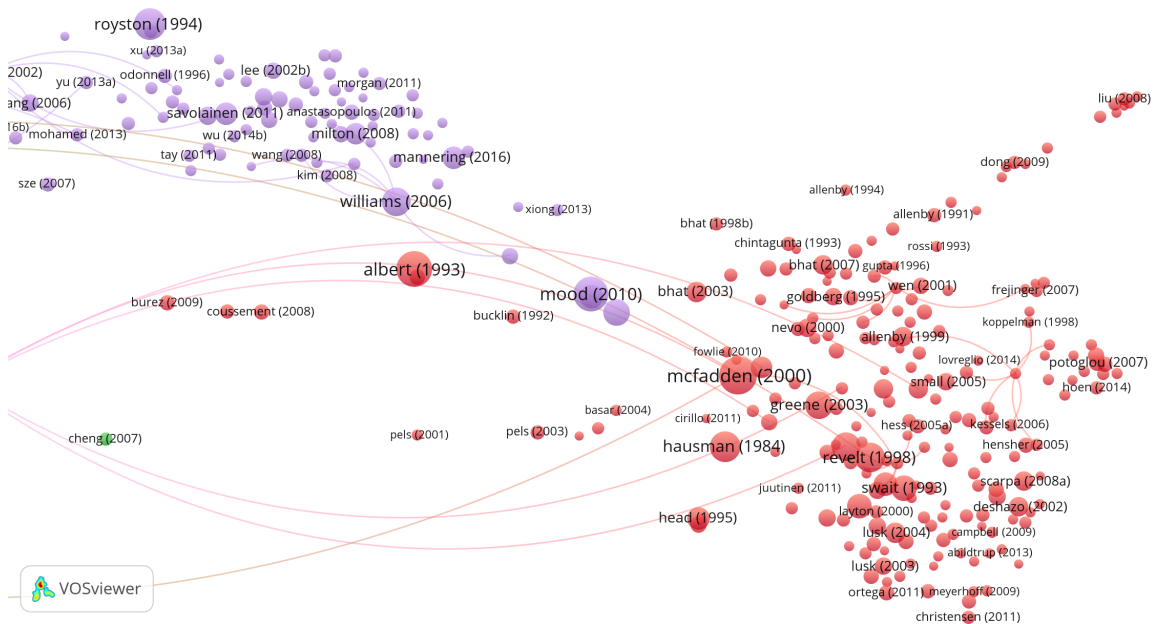


Figure 1.2.: Citation map on document level - focus on Economics (1975 - 2021)⁸

The list of mostly cited works within this cluster comprises mostly theoretical and methodological publications, which perfectly explains their high citation score. The works are mostly consecrated to the discussion of the advanced modelling techniques at the time, that in popularity and availability nowadays.

Among the technical topics, one of the key concepts is the Mixed Logit (or *Mixed MNL* in some more precise cases). This is an advanced modelling technique allowing the introduction of heterogeneity into the Logistic Regression (or Multinomial Logistic Regression) coefficient estimates and thus bypassing some of the technical limitations of the baseline model. McFadden and Train (2000) offers a general overview of the Mixed Logit (MMNL) modelling of the discrete response data; David A. Hensher and Greene (2003) describes the Mixed Logit from the state of practice perspective. Some of the other studies embed similar discussion into more applied work: Revelt and Train (1998) analysing household appliance choice, or Brownstone, Bunch, and Train (2000) illustrating the usage of Mixed Logit models with mixed Stated Preferences (SP) and Revealed Preferences (RP) data. Finally, some of the works represent state of the art for some of the disciplines, as for example the article of Head, Ries, and Swenson (1995) analysing the industrial location choice; or Allenby and Rossi (1998) describing marketing models of consumer heterogeneity.

We also encounter a number of other rather advanced theoretical topics. For example, Albert and Chib (1991) describes a framework for Bayesian Analysis (BA) of *binary* an *polychotomous* data⁹. Hausman, Hall, and Griliches (1984) offers a discussion on MNL model specification testing and validation. Boxall and Adamowicz (2002) and later Greene and Hensher (2003) describe a Latent Class Model for Discrete Choice Analysis, which incorporates some of the ideas of semi-parametric estimation techniques. Lusk and Schroeder (2004) and Brownstone, Bunch, and Train (2000) explore the different implications and usage of data from different sources. Bhat (2001) and Bhat (2003) offers a discussion on the Maximum Likelihood (ML) estimation numerical implementation, with help of quasi-random or Halton sequences.

This sufficiently illustrates the currents state of the existing literature on DCM techniques, their usage and applications. The work relies in part on existing systematic literature overviews on the topic, among which the particular attention is given the work of Haghani, Bliemer, and Hensher (2021), which covers a substantial period from 1900 to 2020 of DCM related literature. The review is completed by another systematic literature review performed separately, exploring DCM techniques usage, with a particular emphasis on the economics related works. The study identifies the key disciplines of DCM application, among which: Economics, Management, and Environmental Sciences, which aligns with other bibliometrics studies. However, the limitations should be acknowledged, including resource constraints and the exclusion of some alternative choice modelling techniques. The study is divided into two main parts: a review based on existing literature on the topic and an automated bibliometrics review. Despite limitations, the combined results might be considered as sufficiently reliable and representative, correctly depicting the current state of DCM in economics.

⁹A *polychotomous* variable is a variable that can have more than two values, the term is often used in contrast to the *binary* variables that take only two values. Polychotomous variables can be ordered, unordered, or sequential.

1.2.1.2. Target metrics of interest

Once the most common problematic and applications are outlined, but before we proceed with the performance analysis, it is important to understand the key target metrics. In this context, the term target metrics refers to the crucial values and indicators that play a pivotal role in addressing the research query. To illustrate this concept, consider a straightforward example: suppose a researcher is interested in analysing the overall demand for traffic analysis. In such a scenario, the primary emphasis would be placed on predicting market shares (Michel Bierlaire and Krueger 2020) and evaluating their external validity. Consequently, for such task the performance would be directly linked to this particular model's output, which is denominated target metrics. For a more precise definition of the target metric concept within the scope of this study, the following definition can be provided.

Target metrics a numerical or logical value(s) obtained as a result of data analysis, serving to answer the research question

As one can guess, the target metrics are case-specific and may differ drastically depending on the particular application. According to our literature review, those metrics may be roughly divided into several groups: (1) the values reliant on the predictions (Bergantino, Capurso, and Hess 2020; Zhao et al. 2020), (2) direct effects¹⁰ (Daly, Hess, and de Jong 2012; M. Bierlaire, Bolduc, and McFadden 2008) and (3) derived metrics (González, Román, and Marrero 2021; Thiene and Scarpa 2009). Here we are going to present those different categories one by one.

The first category might be the easiest to present and explain. Most of the statistical learning models are suitable for those purposes, as they offer some sort of mapping of the inputs into output discrete space for the individuals. This offers the predictions, which might be then used either on individual or aggregated level (Coussement, Benoit, and Poel 2010; Zhao et al. 2020). Recommendation tasks (Danaf et al. 2019), churn prediction (Coussement, Benoit, and Poel 2010), market shares assessment (Bergantino, Capurso, and Hess 2020) - all those tasks rely on the model's predictive qualities.

The second group unites all the values directly available through basic model estimation, without any particular transformations. Most of the traditional DCM methods rely on the effect estimation for a particular utility function, which makes those ideally suited for such tasks. The information on the effects' values may be used for multiple diverse purposes and require distinct model qualities. While some studies might be focused on the effects presence, relying on statistical tests for the analysis, other may be interested in the effects' direction or magnitude. As one can imagine those tasks require different precision levels, shaping the performance requirement and consequently the performance perceptions. For example, Khan, Habib, and Jamal (2020) investigate the impact of smartphone application usage on mobility choices, using data from the Smartphone Use and Travel Choice Survey¹¹. Employing a *Latent Class Random Parameter Logit (LCRPL)* model, the research uncovers behavioural insights related to individuals' attitudes, travel characteristics, built environment, and accessibility measures. Notably, in their conclusions authors focus on how the various variables affects the likelihood to increase vehicle kilometres travelled.

Finally, the most complex metrics may require some transformations applied to the direct estimates. For example, the willingness of consumer to pay for a particular alternative's property is among those metrics, as it requires the researcher to analyse the result of transformation of the direct model esti-

¹⁰Sometimes also denoted as estimates, but this term brings more confusion.

¹¹Data collected in 2015 in Halifax, Canada.

mates¹². As an example in this case the work of Tsouros et al. (2021) might be taken. The authors analyse the demand and WTP for *Mobility-as-a-Service (MaaS)* in Greater Manchester, UK. Employing a multinomial logit model, the research assesses user choices and calculates WTP for different MaaS services, including public transport, car-sharing, bike-sharing, and taxi. The findings in this case offer guidance for potential policymakers, emphasizing user preferences, WTP, and the impact of socio-demographic factors and travel habits on MaaS plan choices. Another case of derived metrics is given by the elasticities analysis or cross-elasticities (Birchall and Verboven 2022), as once again the estimates obtained from the model are used to compute further economic indicators, which serve the researcher as a support for further decision making. Those metrics are the most difficult to explore and analyse.

All of the listed above target metrics might be viewed as point values, or as random values with some underlying distribution. This fact multiplies the number of interpretable target metrics even further, as while some researchers might be satisfied with point estimated values, other applied cases will require information on confidence intervals associated to those values. For example, Coussement, Benoit, and Poel (2010) in their churn prediction application with exception of churners share prediction, offer a set of calculated indicators for managerial decision making. Among those indicators one may encounter: discounted profit and marginal profit values. Nevertheless, the provided values are point estimates and have no further information on the confidence intervals or their underlying distributions. Confidence intervals provide a range of values around the point estimate within which the true value of the parameter is likely to lie, with a certain level of confidence. This makes them extremely important in the context of economic applications, where the decision making based on the modelling results entails gains or losses for the decision-maker. There exist an extensive literature on the confidence intervals identification and computation for the various families of available models. For example, Gatta, Marcucci, and Scaccia (2015) systematically compare methods for constructing confidence intervals for WTP measures in choice modelling, including those from other research fields. Their findings illustrate that the commonly used Delta method may not accurately capture skewness in WTP distributions, while the Fieller method and likelihood ratio test inversion method prove more realistic for small samples.

All in all, this illustrates the heterogeneity in the requirements to the modelling techniques that is present in economics field. The various applications and research questions rely on different target metrics and have distinct requirements in terms of precision associated to those indicators. Evidently, the statistical models available to the researchers are typically unable to cover all the requirements at the same time. The modelling strategy choice entails an arbitration between model capabilities: precision against interpretable estimates, estimation speed and efficiency against against the estimates completeness and coverage. This brings us to the need to explore more in depth the available modelling strategies, applicable to DCM tasks.

1.2.2. Discrete Choice Modelling

In the preceding section, a broad sketch of the subjects and specific applications addressed through discrete choice models was presented. This overview of the field effectively highlights the diverse application domains, each characterized by its unique emphasis, research inquiries, and preferred metrics

¹²In this example it is assumed that the model was not estimated in preference space.

of interest. Furthermore, due to these distinctions in applications, the associated terminology varies, often posing challenges for interdisciplinary communication. In this section, the aim is to tackle this issue by establishing a standardized notation that will be employed in the subsequent chapters. Specifically, for the domain of Discrete Choice Modelling (DCM) the widely accepted notation convention introduced by McFadden (1974) is adopted. For other disciplines an attempt is made to bring the notation as much as possible to the DCM one. In particular, for more universal data-driven models the notations, common for respective fields, were changed to correspond better to the notation used in DCM studies.

The notation introduction in this section has a second objective, which is nonetheless important. Not only the basic notation should be introduced, but as the models and their objectives vary, it is also important to make the reader familiar with the different concurrent state of the art techniques specific for the different research purposes. For a better systematisation of the knowledge this part of the work will present: (1) *data driven* techniques present in the general classification oriented literature; (2) *classical theory driven* techniques, which are the most widely spread across the different application domains; and (3) *other theory driven* methods, which are less widespread in the context of traditional Choice Modelling (CM) studies.

Before proceeding, the drastic difference between the data driven and the theory driven should be emphasised. The separation comes in the base approach to data analysis of the two paradigms. Assuming that some decision making process is observed in the real world (Figure 1.3), which involves mapping of the \mathcal{X} input space variable vector to the \mathcal{Y} output space choice. In the most simple case the input space will regroup the available alternatives' attributes and the output space will contain information about the final decision (choice). Eventually, the researcher will not have the full information neither of \mathcal{X} nor about \mathcal{Y} , the observed (measured) representations of those values may be seen as X and Y respectively. At this point the approaches to analysis the influence of X on Y diverge.



Figure 1.3.: Real world

On the one hand, the researcher can use a theory driven approach to data analysis. In this case it's assumed that the underlying decision making process, or its approximation, is known to the researcher. In choice modelling we typically speak about the behavioural model (Figure 1.4a), as the assumptions are made about the model describing the individual's behaviour face to a decision making process. Obviously, if the assumed underlying model is erroneous or significantly flawed the obtained results have very low validity in this case. However, the possibility to estimate the variable effects in the context of a precise behavioural model have their advantages in interpretability and inference.

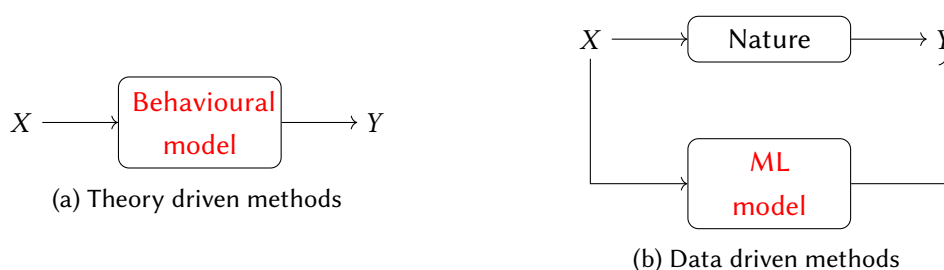


Figure 1.4.: Data analysis approaches

On the other hand, the practitioner can turn to the data driven methodology. The assumptions on the underlying relationships between X and Y are far more permissive. Usually the sole assumption concerns the mapping functions' form, unless the model is not in itself sufficiently flexible to identify the functional form on its own. For the later case we can give the example of Multilayer Perceptron (MLP) and its derivatives, which can potentially approximate any functional form. This leads to *guessing* of the underlying mapping function (Figure 1.4b). While such toolset allows to more easily and sometimes with less work from the analyst to identify the relationships between X and Y , the computation of some target metrics becomes impossible or extremely complicated (burdensome).

Another eventual difference between the two paradigms, which results from the particular use cases, is the usage of specific optimisation techniques. While the general optimisation methods are usually quite similar, such as Maximum Likelihood maximisation, the algorithms differ quite significantly. The data driven methods rely on the quantity of the available data to identify the hidden patterns. This leads to optimisation of the estimation techniques for big or extremely-big datasets, with quite low efficiency in small samples. Such datasets are available for choice modelling applications only in some specific cases, as for example, the RP based datasets for transportation networks. The theory driven methods are oriented to providing as much information as possible in the context of the explored behavioural model. This results in much slower algorithms offering more complete information on the estimates' structure.

The later typically refers to the confidence intervals for estimates. In other words this situation puts researchers before a choice to either be able to work with high-dimensional data and use the ML algorithms speed advantages to maximum, or be capable to derive full information on the estimates and their confidence intervals with respect of the underlying economic theories.

1.2.2.1. Data driven approach and classification

The first group of models focuses on most general and usually more flexible modelling techniques rarely implemented by the economists in their studies. The later fact is justified by the eventual lack of domain specific information in the models' outputs. This model family is usually regarded as not offering enough insight when it comes to the *plain effects* estimation, not even speaking about the more complex composite targets. In other words, the general classification models perform quite well in prediction tasks, but rarely offer sufficient insights when it comes to a more in-depth analysis. This means that such models can produce only a limited number of target metrics, which has an immediate repercussion on the research question types addressed by this model family. The ML techniques are usually viewed by economists as some black boxes, which do not provide any information about the underlying process, even though it is not always the case¹³. It is quite easy to accept their position, as even though the most advanced techniques perform better in terms of predictive power. Such models rarely offer any human-interpretable insight into the decision making process as depicted by classical decision making theories. This subsection will address the differences between the data driven and theory driven models from the perspective of the most traditional classification Machine Learning (ML) models.

In this section our objective is to introduce the reader to the Neural Network (NN) as such models represent the best the family of the data driven techniques. There exist two possible approaches to

¹³Here we may reference the Convolutional Neural Network (CNN)s or Random Forest (RF)s as ML techniques offering some insights into the behavioural implications.

Table 1.2.: ML notation
[H]

| <i>Notation</i> | Definition |
|---|---|
| \mathcal{X} | Input space |
| \mathcal{Y} | Output space |
| $(\mathbf{x}_i, y_i), i \in \{1, \dots, N\}$ | Observation i |
| $\mathbf{x}_i = (x_{i1}, \dots, x_{iR})$ | Explicative variables vector of size R |
| y_i | Outcome variable |
| $\mathcal{S}_N = \{\mathbf{x}_i, y_i\}$ | Sample of N observations |
| \mathcal{D} | Probability distribution |
| $f : \mathcal{X} \rightarrow \mathcal{Y}$ | Function mapping \mathcal{X} to \mathcal{Y} |
| $\mathcal{F} = \{f : \mathcal{X} \rightarrow \mathcal{Y}\}$ | Class of functions |
| $\mathcal{L}(f) = \mathbb{E}[l(f(\mathbf{x}), y)]$ | Generalisation Loss (Error) |
| $\hat{\mathcal{L}}(f(\mathbf{x}), S_N) = \hat{\mathcal{L}}(\omega)$ | Empirical Loss (Error) |
| ω | Parameters of prediction function |

present NNs. In the literature focused on Statistical Learning and data analysis we may encounter the introduction of NN through more simple statistical models (ex. in Hastie, Tibshirani, and Friedman (2009)). The authors start their presentation from the plain linear models, popular in theory driven studies, and extend them to the NN through several generalisation steps. Another representation may be encountered in the community focused on informatics and Machine Learning (ML), where the authors adopt algorithmic approach. Here we are going to use the later discourse tram as it allows to avoid the introduction of all the intermediary models.

Most of the models implemented for data driven analysis rely on the assumptions of *independence* of observations and their *identical distribution*. As one may remark those assumptions are relatively close to the ones encountered in econometrics and Social and Human Sciences (SHS) field. Nevertheless, there is a major difference in the paradigms: while econometricians extend their models to tackle various biases in those two key assumptions, the ML scientists focus on the functional form flexibility in presence of large samples.

The learning problem may hence be formalised: Considering an input space $\mathcal{X} \subset \mathbb{R}^d$ and an output space \mathcal{Y} . The example pairs $(x, y) \in \mathcal{X} \times \mathcal{Y}$ are identically and independently distributed (IID) with respect to an unknown but fixed probability distribution \mathcal{D} . Assuming that only N pairs of $(x_i, y_i), i \in N$ drawn from \mathcal{D} are observed. The *aim* is to construct a function $f : \mathcal{X} \rightarrow \mathcal{Y}$, which predicts an output y for a given x with a minimal error.

Let us define some common notation to be reused in this part of the work:

While we introduce the notation with the individual related subscript i , it will be rarely used in the further text. For simplicity we omit the observation index i so $\mathbf{x}_i = (x_{i1}, \dots, x_{iR}) \Leftrightarrow \mathbf{x} = (x_1, \dots, x_R)$.

Most of the materials in the next several subsections is inspired by the handbooks of Hastie, Tibshirani, and Friedman (2009) and Amini and Usunier (2015). The references to historical works are given directly within the text. For further reading on the topic readers are invited to consult the Appendix B, as well as the abovementioned works.

A. Artificial neuron and Perceptron

Speaking about the particular implementation of the learning algorithms under the form of a NN, we can trace the history to Ramon y Cajal (2002). Nobel prize laureate in 1906 in biology and neuroscience, he remains known as the first one to represent the biological neurons' anatomy. Grace to this particular step in biology domain, the scientific community obtained a new dream - the possibility to artificially reconstruct the neural structure and hence the brain itself. It's in the work of McCulloch and Pitts (1943) that the first mathematical formalisation of a neuron appears (Figure 1.5). Later, many various learning rules were proposed. A most simple formal neuron may be defined with a prediction function $h_\omega \in \mathcal{F}$, which is linear:

$$h_\omega : \mathbb{R}^d \rightarrow \mathbb{R} \quad (1.1)$$

$$\mathbf{x} \mapsto \langle \hat{\omega}, \mathbf{x} \rangle + \omega_0 \quad (1.2)$$

Assuming ω_0 to be included in the vector ω and $x_0 = 1$, we can rewrite the formal rule. The changes may be summarised in graphical form as in Figure 1.5. Here we adopt more familiar for economists graphical representation convention. The observed variables are in squares, the latent construct or intermediary results are in circles and the weights are in plain text.

$$h_\omega : \mathbb{R}^d \rightarrow \mathbb{R} \quad (1.3)$$

$$\mathbf{x} \mapsto \langle \hat{\omega}, \mathbf{x} \rangle = \omega \mathbf{x} \quad (1.4)$$

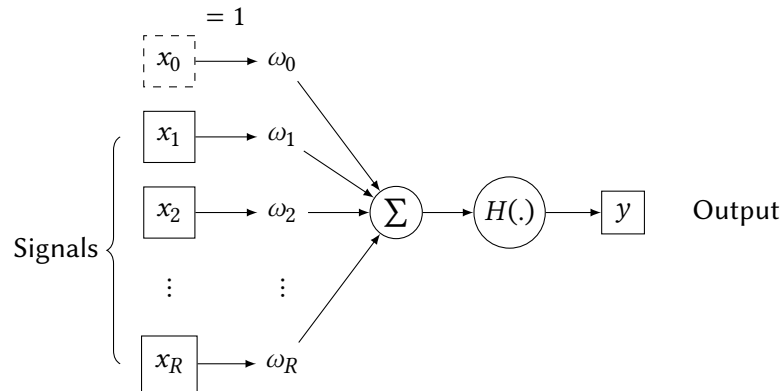


Figure 1.5.: Formal neuron (alternative representation)

Later this model was readapted and tested by Rosenblatt (1958). The linear part of the perceptron was identical to the one proposed previously, but the learning rule was optimised. The model opted to find the best set of parameters $\omega = \{\omega_0, \dots, \omega_R\}$ through minimisation of the distance between misclassified examples to the decision boundary. We may define the objective loss function for the simple *Perceptron* as:

$$\hat{\mathcal{L}}(\omega) = - \sum_{i \in N} y_i(\omega \mathbf{x}_i)$$

As one can see, the simple perceptron is quite close in its linear structure to the basic Linear Regression (LR) model, for which the loss is given as $\hat{\mathcal{L}}(\omega) = -\sum_{i \in N} (y_i - \omega \mathbf{x}_i)^2$. In fact the only differences between those models lie in the specification of the loss function, which drives the optimisation process, and in the learning algorithm, which is used in search for a solution. The plain linear regression relies on matrix operations for coefficients estimation, it is equally assumed that the dataset and effects' dimensionality allows the effects' identification. More about differences between the loss functions may be found in Appendix B, while a systematic review of such functions is presented in the article of Q. Wang et al. (2020).

The statistical models in the restricted context of the classification and Choice Modelling are typically restricted to the models with a *discrete* output. Meaning that the performed modelling task is extremely close to the concept of classification. While the linear model might be used to model discrete binary¹⁴ or ordered outputs, it is advised to use a transformation function (denoted $H(\cdot)$ on Figure 1.5). The common idea across the available transformations is to bound the output values in the probability space (varying in $[0, 1]$ interval), or offer some logic output. Among the most popular transforms we encounter the *sigmoid* (also known as Logit) transformation:

$$H(x) = \frac{e^x}{1 + e^x}$$

B. Multilayer Perceptron (MLP)

With the developments and improvements of simple models some of their drawbacks became apparent (Minsky and Papert 1969). Most of them propose a linear (or sigmoid in case of Logit) separations, whereas in real world such linearly separable problems are few. More elaborate learning algorithms required more complex logical rules, as for example Exclusive OR (XOR) (see Figure 1.6 for the problem presentation) or parity rules. As one can see on the graphical presentation, this problem has no solution through simple hyperplane separation of the input space. In fact it requires two hyperplanes to correctly separate the classes in this case.

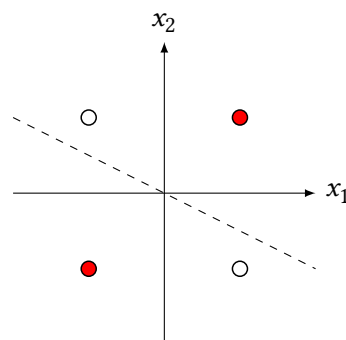


Figure 1.6.: XOR problem

The circuit theory was poorly developed at the time to solve such complex problems. This situation resulted in active search for non-linear models and the specific learning techniques to address the issue.

¹⁴In data driven applications the output levels are typically encoded as $[-1, 1]$ and less frequently as $[0, 1]$.

The invention of Neural Network (NN), also known as Multilayer Perceptron (MLP) or Deep Neural Network (DNN), may be associated with the work of Rumelhart and McClelland (1987). This work is considered to be the first introduction of the backpropagation estimation algorithm to the wide public, although there are ongoing debates about who was the first to invent it. The main idea behind MLP was the possibility to combine simple neurons into a complex system, feeding the outputs of some neurons to other. This methodology resembles to the Project Pursuit Regression (PPR), which in itself is yet another extension of the plain LR model generalisation. Or more accurately, a generalisation of the Generalised Additive Models (GAM), which in its turn extends the capabilities of the Generalised Linear Models (GLM) family. In this particular case the idea is to use the latent space for input variable projection and then predicting the outputs using the latent variables. In functional form the model may be represented as following (Hastie, Tibshirani, and Friedman 2009):

$$f(\mathbf{x}) = \sum_{l=1}^L h_l(z_l) = \sum_{l=1}^L h_l(\omega_l \mathbf{x})$$

Where $z_l = \omega_l \mathbf{x}$, $l \in \{1, \dots, L\}$ represents the elements of L dimensional latent variable space. And $h_l(\cdot)$ is a function mapping the input vector \mathbf{x} to the given dimension components¹⁵. For example the case of 2 layers MLP may be represented as in Figure 1.7, while the representation of MLP may be generalised to any number of layers. In this case we propose a generalisation for a K -class classification problem ($y_k \in \{1, 0\}$, $k \in \{1, \dots, K\}$), where there is a single hidden layer composed of L neurons. This can be generalised even further for a case of S hidden layers.

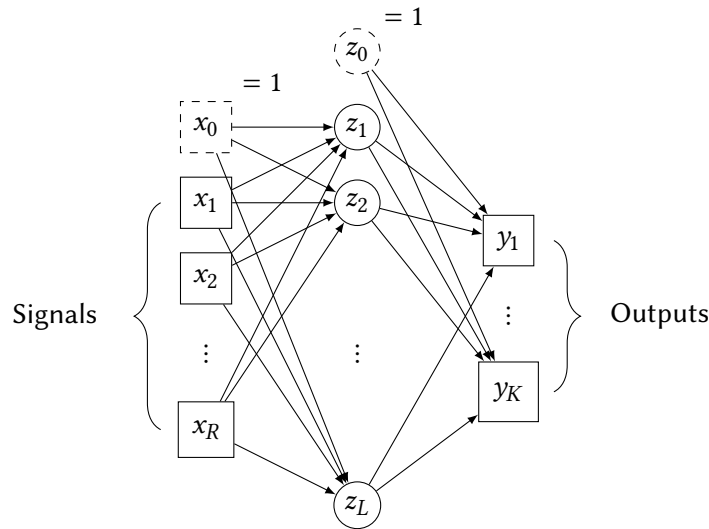


Figure 1.7.: Multilayer Perceptron

In order to formally describe this new model we will need to add index identifiers to our existing notation. For an observation x_i , $i = 1, \dots, R$ taken as input, we can define the element z_l of the hidden layer as:

$$\forall l \in \{1, \dots, L\}, z_l = H^{(1)}(\omega_l^1 x_i) = H^{(1)}\left(\sum_{r=0}^R \omega_{lr}^{(1)} x_{ir}\right)$$

¹⁵As we can not use h_ω due to number of transformations, we update the subscript ω to $l \in \{1, \dots, L\}$.

Where $\omega_l^{(1)}$ is the vector of weights associated with element l of hidden layer. The superscript (1) indicates that this vector belongs to the first matrix of weights, assuming that the elements of hidden layer are not simply linear but undergo some sort of transformation $H(\cdot)$ as well.

The same procedure applies for the output layer, which takes the vector z as input. The Figure 1.7 illustrates the general case of K -class *mono-label* classification, where each element is associated with an indicator vector:

$$\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, y = k \Leftrightarrow y : \forall j \in \{1, \dots, K\} = 1, y_{j \neq k}$$

This vector corresponds to the output layer on Figure 1.7. We can express $y_k, k \in \{1, \dots, K\}$ as:

$$\forall k \in \{1, \dots, K\}, y_k = H^{(2)}(\omega_k^{(2)} z) = H^{(2)}\left(\sum_{l=0}^L \omega_{kl}^{(2)} z_l\right)$$

Or, in a more complete form as:

$$y_k = H^{(2)}\left(\sum_{l=0}^L \omega_{kl}^{(2)} \times H^{(1)}\left(\sum_{r=0}^R \omega_{lr}^{(1)} x_{ir}\right)\right)$$

Once the model defined we would like to return to the similarities with some other statistical techniques. As previously mentioned, the resulting model is a further development of GLM and GAM models denoted Project Pursuit Regression (PPR). This class of models was proposed by Friedman and Stuetzle (1981) as a method of non-parametric multiple regression. The idea was identical to the one behind MLP: to project the input data in the optimal direction before applying smoothing functions.

After the invention of MLP the scientific community was focused on this class of models because of the advantages and flexibility it offered in comparison with more simplistic models such as Ordinary Least Squares (OLS)¹⁶, GLM and even GAM. The main advantage was the possibility to approximate *any* function f , without imposing any additional restrictions and supposition. This allowed to bypass the limitation of the more simple models, for which it was necessary to introduce prior assumptions concerning the defined functional form. Cybenko (1989) was among the first to demonstrate this property for MLP with Sigmoid activation functions. Later, Hornik (1991) demonstrated that the results are not limited to some specific activation functions, but can be generalised for the whole family of the *feed-forward* MLP architecture.

This capacity of a universal approximator of the MLP (and DNN in general) perfectly illustrates the strong points of the data driven modelling approach. However, it also demonstrates the potential pitfalls of such approach, as with the increase in number of parameters to estimate, as well as the number of latent transformations, it becomes more and more complicated to derive meaningful information from the estimates. What is more, it becomes more and more difficult to fine tune and control the model, with the increase in the number of potential transforms and transitions.

This accentuates one of the questions extremely important in the context of NNs, and most ML techniques in general, application: the choice of hyperparameters. As one can observe the number of layers

¹⁶Here we use OLS term instead of plain LR as such definition is more precise in giving the exact optimisation routine.

included into the MLP is not limited by any means, the same goes for the number of neurons within each layer. What is more, each neuron may potentially have some more complex activation function $H(\cdot)$ instead of the plain linear combination of inputs produced by h_ω . This uncertainty makes the hyperparameter choice in neural networks a critical aspect of model development, influencing the network's performance and generalization to unseen data. Hyperparameters are configuration settings external to the model itself and include not only the information about the network configuration, but also the elements specific to the learning algorithm, such as: learning rates and batch sizes¹⁷. The impact of hyperparameter choice extends beyond model accuracy, affecting training speed, computational resources, and the network's ability to handle various complexities in the data. Selecting appropriate hyperparameters is a challenging task, often requiring a balance between in-sample and out-of-sample performance of model¹⁸. A poorly chosen set of hyperparameters may result in a neural network that fails to learn the underlying patterns in the data. Another corner case involves the model memorising the training set leading to poor performances on new data.

The process of hyperparameter tuning involves iterative experimentation, where different combinations of hyperparameter values are tested to identify the configuration that optimizes the model's performance. Techniques like grid search, random search, and more advanced optimization algorithms, such as Bayesian optimization, are commonly employed for this purpose.

Taking the procedure of hyperparameter selection into account may lead to rather interesting results when comparing the model performance to the classic DCM modelling techniques. This may lead to situations, where NNs outperform the DCM model in terms of learning speed and resulting predictive power, but fall far behind once the hyperparameter selection stage is taken into account.

In econometrics and particularly in choice modelling the NNs are recently gaining popularity with the increase in number of work attempting to use this toolset for transportation and behavioural related modelling tasks. Some attempts have already been undertaken in year 2000 to imitate the MNL structure with ML toolset (Bentz and Merunka 2000; Hruschka, Fettes, and Probst 2001), although the conclusions at the time were not satisfactory due to the computation burden and low efficiency in comparison with classic DCM toolset. Recently those studies have regained popularity with the works from S. Wang, Wang, and Zhao (2020), S. Wang, Mo, and Zhao (2020) and S. Wang et al. (2021). While subject to some criticism those works have illustrated the possibility to combine the flexibility of NNs with the RUM-compliant theoretical assumptions. The final version of proposed model includes a DNN with restricted output layer, denoted *Alternative Specific Utility Deep Neural Network (ASUDNN)*. More information about economic information extraction from NNs may be found in Annexe B.

1.2.2.2. Theory driven approach and Classic Choice Modelling

Contrary to the previous section the theory driven approach takes its root primarily in the prior theoretical assumption on the models' structure. Instead of performing a plain approximation of the mapping function, the researchers focus on the validation or confirmation of the given model under theoretically imposed constraints. This obliges us to shift the focus of our discourse to the underlying theories.

Classic Choice Modelling (CM) is also known as Discrete Choice Modelling (DCM) or Discrete Choice

¹⁷Typically the model's weights are updated only once several iterations. The *batch size* determines how often the model weights are updated.

¹⁸One may also view this problem as under-fitting or over-fitting of the resulting model.

Analysis (DCA). Depending on the literature it may have different means and address multiple types of modelling techniques, starting from basic classification to far more sophisticated behavioural oriented models. There exist several theoretical frameworks for modelling individual choices and behaviour. In this subsection we will address only the most widespread Random Utility Maximisation (RUM) framework, attributed to McFadden (1974). The current subsection focuses solely on the RUM framework and related modelling techniques, as well as eventual issues and inconsistencies. An overview of concurrent theoretical frameworks is presented in the next subsection of the manuscript. Concurrent frameworks are presented in less detail than RUM due to their lesser popularity in the literature.

Let us now focus our attention on the RUM framework. As mentioned previously some of its elements may be traced to the work of Luce (1957), the key reference in the literature of choice modelling is the publication of McFadden (1974). In short, the RUM framework relies on the general utility concepts from the economics. Individuals are assumed to choose alternatives that maximize their utility, meaning that the individual always selects the alternative with the highest perceived utility. The utility of an alternative in this case is a latent variable that represents the satisfaction or preference an individual derives from choosing that alternative. The utility is denoted U by convention. Given the uncertainty in measurements and unobserved impacts, resulting in differences between $\{\mathcal{X}, \mathcal{Y}\}$ and $\{\mathbf{X}, \mathbf{Y}\}$, it is also common to separate deterministic part V and unobserved elements, or errors, as ε . The stochastic component in the utility function, ε , corresponds to the term *random* in RUM. This random term captures unobserved factors or random variations in decision-making that are not explicitly modelled. The randomness helps to account for unobservable heterogeneity among individuals and eventual shocks to preferences.

At this point we may switch to formalisation of the theoretical concepts in a mathematical form. Thus for individual i and alternative $j \in \Omega$, where Ω is a choice set of all available alternatives, the utility may be defined as:

$$U_{ij} = V_{ij} + \varepsilon_{ij}$$

Where, V_{ij} is the deterministic utility part. And ε_{ij} represents the error term, an identically and independently distributed Extreme Value (EV) random variable. The usage of the EV as distribution for the unobserved random variables is a rather comfortable solution given the model specific structure, but we will see this later. At this stage the comprehension of utility concept is a more crucial matter. As it will be seen further the restrictions on the error term may be relaxed.

At this point we offer a brief outline of the most common *RUM-compliant* models and model families. RUM-compliance refers to adherence to the assumptions and principles of the RUM framework in discrete choice modelling, regrouping the ensemble of models constructed relying on the RUM framework. Among the first and most widely represented in the literature we encounter the Logistic Regression (Logit), widely used for binary classification. It is among the most widely used techniques in the context of the RUM framework. While the history of sigmoid transformation may be traced to the early 1900, its introduction in the context of choice theory development is attributed to Luce (1957). A state of the art extension to this model, the Multinomial Logistic Regression (MNL) is linked to choice modelling by McFadden (1974) and McFadden and Train (2000). This model extended the binary Logit to polychotomous case, where a single choice from a set of multiple (more than 2) alternatives was being made. It is exactly to McFadden (1974) that the he introduction of Random Utility Maximisa-

tion (RUM) framework is attributed. In the baseline models the consumers optimize (and researchers estimate) an indirect utility function, that “*has a closed graph and is quasi-convex and homogeneous of degree zero in the economic variables*” (McFadden 2001). Applying the standard model to discrete choice requires the consumer’s choice among the feasible alternatives to maximize conditional indirect utility based on some reference alternative, rather than absolute utility. Among other models we encounter Probit, which could be traced up to Bliss (1934), and Multinomial Probit models. There also exist Nested versions (Ben-Akiva and Bierlaire 2003) of all the listed above models, as well as Mixed or Random Parameter (Stern Decembre 1997; K. E. Train 1998; McFadden and Train 2000) extensions for those models.

There exist a further extension based on the EV model family, which is generalisation of the previously introduced RUM approach. The ensemble of such models are reunited into Generalised Extreme Value (GEV) group. Among the examples we can reference the works of Ben-Akiva and Bierlaire (2003) or Fosgerau, McFadden, and Bierlaire (2013). In this subsection we do not take into account the models for ordered or sequential choices. More information on non-RUM compliant DCM tools might be found in the work of Bouscasse, Joly, and Peyhardi (2019), where the frontier between RUM and non-RUM models is traced.

Among the cited extensions many address theoretical inconsistencies between the baseline RUM framework and the observed behaviour. Among them the complex choice structures, addressed through a Nested Logit (NL) model, which arrived as a remedy for Independence from Irrelevant Alternatives (IIA) relaxation¹⁹, as described in Ben-Akiva and Bierlaire (2003). Later extended with more powerful techniques, as for example the Cross-Nested Logit (CNL) (Michel Bierlaire 2006). Or the Mixed Logit (MMNL or XML) models, arriving as even more vast generalisation of RUM models Brownstone and Train (1998), K. E. Train (1998). This model class is extended even further with flexible distributions (K. Train 2016), and other methods. Taking their root in *Generalized RUM models* (J. Walker and Ben-Akiva 2002), we can find some other models. Integrated Choice and Latent Variable (ICLV) models, alternatively known as Hybrid Choice Models (HCM) are described by Ben-Akiva et al. (2002) and later by Abou-Zeid and Ben-Akiva (2014).

The extensions are not limited to the described above models. The *Bayesian Estimation* techniques have also found its way to the choice modelling community, as they may be used for most of the described above models. For example, we can refer to the works of Allenby and Rossi (1998) or Ben-Akiva, McFadden, and Train (2019). Some other research directions address the usage of RUM models with panel data, adding either temporal or spatial dimension to the observations (or in rare cases both of them). One of the reference works is the production of Arellano and Honoré (2001). Dynamic Structural Choice Models (DSCM) are described by Aguirregabiria and Mira (2010). The Dynamic Choice Models (Dyn.CM) are reviewed by Hillel and Frejinger (2021), although only a working paper and not a state of art work. The family of dynamic models is further extended in the literature to Online Choice Models (OCM), which could be updated on the fly taking into account both *intra-* and *inter-*individual heterogeneity in behaviour. For more information see Danaf et al. (2019) and Danaf, Atasoy, and Ben-Akiva (2020).

The Table 1.3 extends previously introduced notation, adding some more discipline specific elements and altering other, while attempting to keep it relatively consistent. For example, to better follow

¹⁹A more detailed overview of the Independence from Irrelevant Alternatives in the literature is available in Appendix C.

Table 1.3.: Discrete Choice Modelling (DCM) notation extension
[H]

| Notation | Definition |
|--|---|
| $i \in \{1, \dots, N\}$ | Individual i index |
| Ω | Set of alternatives |
| $j \in \{1, \dots, J\} : \Omega$ | Alternative j index |
| \mathbf{s}_i | Characteristics of individual i |
| \mathbf{z}_j | Attributes of alternative j |
| $\mathbf{x}_{ij} = (\mathbf{s}_i, \mathbf{z}_j)$ | Explicative variable vector for individual i facing alternative j |
| \mathbf{x}_{ij}^* | Latent or unobserved explicative variables |
| \mathbf{i}_{ij} | Indicators of \mathbf{x}_{ij}^* |
| U_{ij} | Utility of an alternative j for individual i |
| V_{ij} | Deterministic utility of an alternative j for individual i |
| β | Effects of individual characteristics |
| γ_j | Effects of alternative's j attributes |
| α_j | Intercept for alternative's j attributes |

the DCM literature, the explicative variable vector \mathbf{x}_{ij} is separated into two elements: (1) individual characteristics \mathbf{s}_i and (2) alternative attributes \mathbf{z}_i . A particular attention is drawn to the \mathbf{z}_i , as it is not a latent variable anymore, as previously employed in the NNs for hidden layer denomination.

The shift from the more general *classification* oriented notation to the more DCM specific one should be apparent at this point. For example, as the context of DCM studies imposes the presence of decision makers the i index previously used to denote the observations, now corresponds to individuals, assuming that each individuals makes single choice at this point²⁰. The intermediary latent concepts corresponding to theoretical assumptions equally appear in the table. Finally, not only the explicative variable vector \mathbf{x}_{ij} is separated into two parts, but the associated effects, or weights, might also be separated into several elements: (1) individual specific β ; (2) alternative dependent γ_j ; and (3) intercept α_j , corresponding to the ω_0 weight in ML applications.

A. Multinomial Logistic Regression (MNL)

We will not linger on the presentation of the binary Logit case and will proceed outright with the more general model. The MNL model is one of the simplest RUM-compliant models, although it is sometimes criticised for not addressing many behavioural issues. This class of models relies on the hypothesis, that an individual $i \in \{1, \dots, N\}$ maximises his perceived utility over a set of alternatives $j \in \Omega$, as described earlier:

$$U_{ij} = V_{ij} + \varepsilon_{ij}$$

Where the deterministic part is represented by convention by a linear function of individual characteristics \mathbf{x}_i of the given individual. As well as the alternative specific attributes, denoted \mathbf{z}_j .

²⁰Obviously, this assumption may not hold and in more complex DCM models a notation with 3 indices is used, each index corresponding to the individual, choice situation and alternative respectively.

$$V_{ij} = \alpha_j + \beta \mathbf{x}_i + \gamma_j \mathbf{z}_j$$

Both β , representing the alternative specific individual coefficients, and $\gamma_j \forall j$, standing for population-wide attributes effects, are assumed to be fixed across population, meaning that all the individuals have identical preferences and are subject to identical effects. As precised in Agresti (2013) this approach enables discrete-choice models to take into account both: (1) *characteristics* of the agent and (2) *attributes* of the alternatives. More simple models may be imagined if the access to the individual characteristics or alternatives' attributes is limited, resulting in two special cases. The first model captures only alternatives' attributes impacts $U_{ij} = \alpha_j + \gamma_j \mathbf{z}_j$ and is also known under the name of *Conditional Logit*. The second case results in modelling only individual characteristic effects $U_{ij} = \alpha_j + \beta \mathbf{x}_i$. In literature that separates the conditional Logit as a separate case, the later model is denoted as MNL (K. Train 2002). However, in the less specific literature both formulations, as well as their mixtures appear under MNL denomination.

The MNL is based on the assumption that the residuals ε_{ij} are identically and independently distributed (IID) as Gumbel random variables with zero mean and scale parameter σ (scaling factor), which is usually set to 1. Such changes ensure the model identification, which could not be otherwise achieved for means different from zero. The probability of choosing alternative ω_j from among those available $\{\omega_1, \dots, \omega_k\} \in \Omega$ by individual i , can be expressed in closed form as:

$$P_{ij} = \frac{e^{V_{ij}/\sigma}}{\sum_{l=1}^k e^{V_{il}/\sigma}}$$

B. Nested Logistic Regression (NL)

One of the most simple extensions addresses the issue of non-compliance with the Independence from Irrelevant Alternatives property. The IIA property assumes that the choice procedure is not affected by the irrelevant alternatives, meaning that alternatives not considered by the individual should not impact the decision. We are going to skip the most widely spread example illustrating the IIA property inconsistency in the context of the baseline DCM models, also known as the “*red and blue bus paradox*”. For a better understanding of the IIA property treatment the Annexe C is available.

In short, the key concept of the model is that individual may consider a subset of alternatives in a relatively uniform set. For example, in the transportation mode choice modelling it is quite common that individual will primarily consider the choice between the subset of public and private transportation options. And only then select the most appropriate mode within the chosen subset. Such *division* is achieved through a non-uniform covariance structure of ε_{ij} . The model reunites the subsets of alternatives within *nests* with a common covariance, reflecting the complex decision making process. Assuming there exist M nests (subsets) of alternatives. Let us denote $\Omega_m \in \Omega$ the subset of alternatives belonging to the nest $m \in \{1 \dots, M\}$. Assuming the λ_m is a scaling parameter for the nest m , the choice probability for an alternative j in a nest λ_l might be expressed as:

$$P_{ij} = \frac{e^{V_{ij}/\lambda_l} (\sum_{k \in \Omega_l} e^{V_{ik}/\lambda_l})^{\lambda_l - 1}}{\sum_{m=1}^M (\sum_{k \in \Omega_m} e^{V_{ik}/\lambda_m})^{\lambda_m}}$$

The λ_m values should be comprised in $[0, 1]$ interval for RUM consistency, which produces the Nested Logit (NL). A special case of the model may be observed when $\lambda_m = 1 \forall m$ the error terms are IID Gumbel variables and the model is reduced to the baseline MNL model.

C. Mixed Multinomial Logistic Regression (MMNL)

The MMNL is another development and generalisation of MNL, because these models may be constructed using MMNL specification with a correct parametrisation. The model's history may be traced down to the publications of Cardell and Dunbar (1980) and Boyd and Mellman (1980), as well as the works from other disciplines considering the similar structures (Talvitie 1972). More widely referenced academic publications include works of McFadden and Train (2000) and Brownstone, Bunch, and Train (2000). The main difference from the more simple models is that in this case it is assumed that effects vary across population and might even be correlated. The utility specification in this case is constructed identically to simple models, but the deterministic part assumes that effects vary across population. Mathematically the random effects specification is achieved through the parameter vector γ_i , which is unobserved for each i . The γ in this case is assumed to vary in the population following the continuous density $f(\gamma_i | \theta)$, where θ are the parameters of this distribution.

$$U_{ij} = V_{ij} + \varepsilon_{ij} \text{ where } V_{ij} = \alpha_j + \beta \mathbf{x}_i + \gamma_j \mathbf{z}_j$$

The simplest choice of the distribution for the random effects is the Normal distribution, which was used by Michaud, Llerena, and Joly (2012), or more precisely a multivariate Normal distribution, because authors took into account the correlation between coefficients:

$$\gamma_i \sim MVN(\gamma, \Sigma)$$

Obviously the choice of the distribution reposes fully on the researcher. As there might be intentions to bound the effects to be strictly positive or negative. There exist studies offering the possibility to specify flexible mixing distributions (Danaf, Atasoy, and Ben-Akiva 2020; K. Train 2016) or use non-parametric distributions (Vij and Krueger 2017; K. E. Train 2008). Earlier works focused on the specification tests for the fixed distributions (Fosgerau and Bierlaire 2007).

Finally, for estimation the simple Maximum Likelihood is not sufficient, leading users to usage of Maximum Simulated Likelihood (MSLE) or Method of Simulated Moments (MSM).

D. Integrated Choice and Latent Variable Models (ICLV) and Hybrid Choice Models (HCM)

The first description of Integrated Choice and Latent Variable model (Bouscasse 2018) type may be traced to the work of Ben-Akiva et al. (2002). The work extended the baseline MNL with additional *latent* concepts. Depending on the type of those latent variables it is common in the literature to speak about either Latent Class Choice Model (LCCM) (Greene and Hensher 2003) or Latent Variable Choice Model (LVCM) (Ben-Akiva et al. 2002) models. The model class may be seen a union of Structural Equation Models (SEM) and MNL models. While the SEM part provides structural equations for latent

variable distribution approximation. The MNL uses the resulting latent variables as plain inputs for choice modelling purposes.

The structural equation part for a variable x^* is typically given as follows. Here for simplicity we avoid the individual related indexing.

$$x^* = h(x, \omega) + \eta \text{ where } \eta \sim \mathcal{D}(0, \sum_{\eta})$$

The measurement part of the model uses the indicator variables for x^* elicitation:

$$I = g(x, x^*; \alpha) + \epsilon \text{ where } \epsilon \sim \mathcal{D}(0, \sum_{\epsilon})$$

In the following equation we simplify the notation assuming the vector of parameters β regroups both individual and alternative specific elements, as the distinction between the elements is not the key element of the presented model. The utility is defined as in the baseline MNL model:

$$U = V + \varepsilon \text{ where } V = f(x, x^*; \beta) \text{ and } \varepsilon \sim \mathcal{D}(0, \sum_{\varepsilon})$$

Those elements constitute the complete model. Assuming η, ϵ and ε are independent, the joint probability of the outcome y and indicators I , conditional on exogenous variables x is given as:

$$\pi(y, I | x; \alpha, \beta, \omega, \sum_{\eta}, \sum_{\epsilon}, \sum_{\varepsilon}) = \int_{x^*} P(y | x, x^*; \beta, \sum_{\varepsilon}) g(I | x, x^*; \alpha, \sum_{\alpha}) h(x^* | x; \omega, \sum_{\omega}) dx^*$$

Where the first term of the integral corresponds to the actual choice model, the second represents the measurement model and the third stands for the structural equation from the latent variable model. The measurement may be omitted in model application stage as it serves primarily to increase the estimates reliability.

Later in the literature we may encounter another model family which repeats the same ideas as the ICLV: the Hybrid Choice Models (HCM). The HCM family regroups all of the previously mentioned elements and extensions, it is sometimes referred to as the Generalised Choice Models (GCM). The following definition for this model class is proposed in the work of Abou-Zeid and Ben-Akiva (2014) and illustrates the similarity of the model families:

Hybrid Choice Models (HCM) is a modelling framework that attempts to bridge the gap between discrete choice models and behavioural theories by representing explicitly unobserved elements of the decision-making process, such as the influence of attitudes, perceptions, and decision protocols. It integrates discrete choice models with latent (or unobserved) variable models.

1.2.2.3. Other theory driven approaches and irrational behaviour

In the previous subsections the two different approaches to DCM were described: (1) ML methodology focused on generic classification task; and (2) classic DCM approach aligning with econometric

toolset, specifically designed for individual choice analysis. Once the overview is complete the transition should be made to a brief introduction of alternative theories. Evidently, the RUM-compliant models, while popular, are not unique in their availability. The alternative behavioural theories lay ground for concurrent behavioural frameworks, which might or might not use the same modelling techniques as introduced before. While these theoretical frameworks will not make appearance in Chapter 3, they play a crucial role in understanding of current state of choice modelling. All those theories make part of the existing body of scientific literature in Choice Modelling.

While the RUM framework is one of the most popular ones and the easiest to grasp, there exist alternative approaches to modelling the behaviour. Within the existing literature the concept of , traced back up to J. F. Nash (1950), occupies a central position in the majority of available behavioural theories. Even though the rational behaviour became criticised and nowadays more and more studies attempt to bypass the limitations and constraints imposed by this concept, in this task authors still rely on . Most of theories present in this subsection will are the results of such works.

The current trend to oppose the RB theory roots in the assumption that humans are rarely rational (Durlauf and Blume 2010). While modelling the mean behaviour the desired properties may be achieved grace to a sufficiently large sample, the irrational behaviour becomes extremely pronounced in the cases where its encouraged. Among the traditional choice theory key works J. F. Nash (1950) and Luce (1957) works occupy particular place, the later being revisited in Luce (1977). Both Nash and Luce, in their respective domains, provided mathematical frameworks to understand RB. Nash's equilibrium concept is crucial in strategic decision-making, while Luce's work in decision theory and utility theory contributes to the understanding of how individuals make rational choices in various situations. The RB refers to the decision-making process where an individual or agent systematically and consistently chooses actions that maximize their overall gains (or utility), given the available information, preferences, and constraints. From the theory of rational behaviour arises the more widely exploited in the Choice Modelling literature RUM concept, which assumes that individuals maximise the perceived utility defined by deterministic and stochastic components (McFadden 1974). The RUM framework implies that economic agents are fully rational in their strife to maximise their observed utility. However, there are always some limitations in the most general models. In RUM-compliant models the NL, MMNL and HCM models serve exactly this purpose, relaxing some of the basic restrictions imposed in the case of baseline MNL model.

The original versions of classic theories are typically extremely constrained and unrealistic in the real world. For example, in the case of Nash's game the subject may attempt to minimise the cognitive burden in selecting a suboptimal alternative, or prefer to chose the option resulting in worse possible outcome for their adversary. Identically, in the case of Luce's choice model the individuals might be affected by seemingly irrelevant alternatives. Moreover, not only the theory itself may be unrealistic in the particular context, the data collection strategy may also induce errors. Due to an imperfect access to information, external factors and cognitive biases, people are rarely fully rational. And even though in large samples such behaviour can be neglected sometimes (Ludwig et al. 2021), in the small datasets the problems become more and more prominent. Many recent theories attempt to bypass the limitation of the classical theory. As it becomes apparent from the research conducted by Haghani et al. (2021a) recent studies consider various forms of biases.

One of the most popular approaches and direct concurrent to the RUM framework is the Random Regret Minimisation (RRM) behavioural approach by Chorus (2010). It challenges the assumptions

that consumers maximise their utility. The decision rule is then summarised to *choosing the lesser evil*, which can effectively represent the actual behavioural pattern in some cases. Among the most common examples we encounter the choice of transportation mode, or as suggested by David A. Hensher, Greene, and Chorus (2013) the choice among durable goods (ex: vehicle type). Another recent approach is the modelling of behaviour through Random Advantage Maximisation (RAM) theory (Leong and Hensher 2015), which is tightly linked with previously described RRM approach. Both of them belong to a family of *context-dependent* frameworks as suggested by Belgiawan et al. (2019), meaning that they allow to perform better estimations in specific cases where individuals' rationality may be affected by the choice set composition. Some of the more advanced works attempt to incorporate the newer theoretical models with the classical ones. For example, Hess and Chorus (2015) propose a mixture model where RUM and RRM behaviours are united through a latent class structure, making it possible to model the population with different behavioural profiles. There exists as well the recent theory of Quantum Choice (QC) modelling described by Yukalov and Sornette (2017) and implemented in some applied studies (Hancock, Hess, and Choudhury 2018; Gangi and Vitetta 2021). This last framework tackles the issues of irrationality of the individuals choice and solves some of the behavioural paradoxes present in the more classical behavioural theories.

De Palma et al. (2008) extends the traditional choice models to the choice modelling *under uncertainty*. This model family is tightly intertwined with the more general Expected Utility (EU) theory introduced by Kahneman and Tversky (1979), which may be understood as a further extension of the classical utility-based choice theories into the probabilistic field. In the case of EU theory, as described by De Palma et al. (2008) there is a number of paradoxes that go outside of the EU theory scope. Among them: (1) the alias paradox (Allais 1953), illustrating that in reality individual choices do not comply with the Expected Utility theory; (2) rank-dependent EU frameworks of Quiggin (1981) and Schmeidler (1989); (3) loss aversion, introduced by Kahneman and Tversky (2012); and many more. In this part of the work we attempt to cover most of these less popular in DCM context theories. Obviously, the fact that those theories are not dominant in DCM applications, does not in any way mean that there is no active community and ongoing research on each of the topics.

A. Expected Utility (EU)

For the presentation of a standard EU model we adopt the reasoning proposed in the work of De Palma et al. (2008). In this theory the decision maker is assumed to face several alternatives, *prospects*, denoted \mathcal{P}_a , each composed of several uncertain outcomes, also denominated as *events* E_j , known to individual. For a more detailed view on changes in notation, please, see Table 1.4. In the decision theory standard notation: $\mathcal{P}_a \succcurlyeq \mathcal{P}_b$ implies that individual is willing to choose, or prefers, the prospect \mathcal{P}_a from set $\{\mathcal{P}_a, \mathcal{P}_b\}$. Identically to the basic discrete choice theory, assuming the rationality of the individual, in the case of set containing multiple prospects (alternatives) the individual selects the one, which is dominant in the given choice set.

Due to the absence of certain knowledge regarding the true event, the resulting outcome from an individual prospect remains uncertain. This concept is encapsulated in the phrase *decision under uncertainty*. It is usually assumed that the probabilities p_j associated with those outcomes are known to individual. This setting allow to get a probability distribution of a prospect \mathcal{P} : $(p_1 : u_1, \dots, p_n : u_n)$, where u_j stands for associated utility gains. The rational individuals in this case are able to compute the expected utility U of prospect \mathcal{P} as:

Table 1.4.: Expected Utility (EU) Theory notation
[H]

| Notation | Definition |
|---|---|
| E_j | Possible <i>event</i> j , for $j = 1, \dots, n$ |
| $\mathcal{P} = (E_1 : x_1, \dots, E_n : x_n)$ | A <i>prospect</i> |
| u_j | Utility (ex: money) being the <i>outcome</i> of the prospect if E_j is true |
| U | Expected utility of the prospect |
| p_j | Probability of an event E_j |

Table 1.5.: Prospect Theory (PT) notation
[H]

| Notation | Definition |
|--|---|
| π or $\pi(p)$ | Scale reflecting the impact of p on the value of prospect |
| ν or $\nu(u)$ | Scale reflecting the subjective value of an outcome |
| $V(p_1 : u_1, \dots, p_n : u_n) = \sum_{j=1}^n \pi(p_j)\nu(u_j)$ | The overall value of a given prospect |

$$U_{\mathcal{P}} = \sum_{j=1}^n p(E_j)u_j$$

At this point assume that individual faces a choice of two prospects \mathcal{P}_a and \mathcal{P}_b , for both of which the associated utilities might be calculated. Then \mathcal{P}_a dominates \mathcal{P}_b , or $\mathcal{P}_a \succcurlyeq \mathcal{P}_b$, if and only if:

$$\sum_{j=1}^n p(E_j)[u_{aj} - u_{bj}] \geq 0$$

In most applied cases it is assumed that individual discards the dominated prospects and the choices are made over a subset of the more difficult to compare elements.

B. Prospect Theory (PT)

In the literature we encounter the references to the work of Kahneman and Tversky (1979). It is considered to be one of the first attempts to produce a complete theory of the human decision making (behaviour) under uncertainty, as it deviates from the norms of EU theory (Heukelom 2015).

“Decision Theory (DT), or the theory of risk, or rational choice theory, goes back to the second half of the seventeenth century when scholars started to investigate how to calculate mathematically the optimal decision in uncertain situations. The mathematics that came out of these and similar questions was probability theory and rational choice theory.”
(Heukelom 2015)

Kahneman and Tversky (1979) in their work attempt to address the pitfalls of the dominant at the time EU theory. The underweighting of the probable outcomes in comparison with certain outcomes, risk aversion, as well as isolation effect leading to inconsistent preferences are all addressed in by authors. Starting from the EU theory, the authors extend it with incorporation of *certainty influence, reflection*

Table 1.6.: Decision Field Theory (DFT) notation
[H]

| Notation | Definition |
|-----------------------------|---|
| t | Time |
| ϵ | An arbitrary small point in time |
| $\mathcal{P}(t)$ | A vector of <i>preference states</i> at time t for given number of alternatives |
| $\mathcal{P}(t + \epsilon)$ | An approximation of the <i>diffusion process</i> as $h \rightarrow 0$ |

effect, probabilistic insurance, isolation effect. The final model assumes that both the individual perceptions of probability (p_j) and outcomes (u_j) affect the perceived utility or value of a prospect, for a more detailed information on the changes in notation refer to Table 1.5. Which, assuming π and v , could be formalised as:

$$V(p_1 : x_1, \dots, p_n : x_n) = \sum_{j=1}^n \pi(p_j)v(u_j)$$

Heukelom (2015) offers a rather interesting historical overview of the decision making under uncertainty development. The author traces the development of this behavioural literature branch till the works of Nicholas Bernoulli and the so called *St. Petersburg Paradox*, which presumably lied the theoretical basis for the future development of the entire discipline. The key point at that time was the introduction of individual (*subjective*) assessment of different valuations, while it was more common practice in the literature to directly address the issue in the universal monetary values.

C. Decision Field Theory (DFT)

Also known as Cognitive-Dynamical Approach to decision making and preferential choice, which introduces several new concepts summarised in the Table 1.6. A quite exhaustive overview of this method is offered in the work of Jerome R. Busemeyer and Diederich (2002). As claimed in the review, this theory takes its roots quite far away in history:

The name ‘Decision Field Theory’ reflects the influence of an earlier theory of conflict formulated in Lewin Kurt (1935) dynamic theory of personality called a ‘field theory’ of personality ...

The theory is primarily based on psychological principles, which explains its scarce popularity in economics (contrary to psychology). It was initially proposed as a “*deterministic–dynamic model of approach–avoidance conflict behaviour*” (Townsend, Busemeyer, and Izawa 1989) and then extended as a “*stochastic–dynamic model*” (J. R. Busemeyer and Townsend 1993). By its structure the model is very similar to the traditional RUM type stochastic model, although it is more complex. The model takes into account the eventual stochastic changes in individual behaviour during the decision making procedure. Meaning, that the decision process varies across the time. The equation proposed to describe this process is given as follows:

$$\mathcal{P}(t + h) = S\mathcal{P}(t) + V(t + h)$$

Where $S = (I - \beta T)$ is a matrix, which is symmetric with equal diagonal values. The interpretation of the diagonal values provides memory for previous states of the system, while the off-diagonal values allow for competitive interactions among the available alternatives. And $V(t) = CMW(t)$ is a *valence vector*. The *valences* result from comparing anticipated value of an option on an attribute with the anticipated values of other options on the same attribute. In this case the three composing elements are as follows: (1) C is a *contrast matrix*, providing weighted evaluations produced by the $MW(t)$ product (typically it's chosen so $\sum V(t) = 0$); (2) M regroups all the possible evaluations of each option on each attribute under each state of nature; (3) W contains weight corresponding to each column of M at time t , changing over time following a stationary stochastic process.

An interesting discussion is proposed by Jerome R. Busemeyer and Diederich (2002), describing the links and contrasts among RUM and DFT. Specifically applied to respect of IIA and *regularity* of choice, which are typically viewed as constraints for traditional CM, although bypassed by the DFT. An illustration of how the theory is extended to Multialternative Decision Field Theory (MDFT) as, for example, in Roe, Busemeyer, and Townsend (2001). This extension accounts for several effects, which usually attract attention in the psychology literature: (1) similarity effect (Kahneman and Tversky 1979), (2) attraction effect (Huber and Zwerina 1996), (3) compromise effect (Simonson 1989)

D. Quantum Probability (QP) and Quantum Decision Theory (QDT)

This subsection only outlines the general concepts of QDT framework, without dwelling into any details. For a better understanding of those concepts it is advised to explore the cited works, to avoid any confusions. The class of models similar to QDT was described by Vitetta (2016), simulating the case in which the user has an unclear sequence of decision for his final choice of an alternative. The author in this case speaks about Quantum Utility Model (QUM), which in theory bypasses the limitations of standard RUM based models, as illustrated by authors thorough simulation in their work.

Later, Yukalov and Sornette (2017) illustrate that behavioural probabilities, supposedly used by human decision makers, share many common features with *quantum probabilities*. In this case the connection is made not only for the sequences of choices, but for describing decision-making in the case of composite prospects under uncertainty in general. A more recent work of Hancock et al. (2020) offers an overview of the existing methodology underlining the key features of the quantum probability. One of the major differences between classical and quantum logic reveals that under quantum probability theory is that the law of probability following the distributivity of 'and' and 'or' of proposition $A \wedge (B \vee C) = (A \wedge B) \vee (A \wedge C)$. The discussion proposed by authors is rather captivating:

... quantum models have also since made the transition into choice modelling (Lipovetsky 2018). Furthermore, quantum models can be used to accurately capture the change of decision context and mental state" when moving between choices made under revealed preference and stated preference settings (Yu and Jayakrishnan 2018).

In their work authors operate in terms of *lotteries* (L_j), which by their nature are quite close to *prospects* introduced in EU and Prospect theory frameworks (\mathcal{P}). The notation changes are available in the Table 1.7. In this particular case, every decision involving such choices carries a duality of uncertainty, both (1) objective and (2) subjective in nature. In an objective sense, when opting for a lottery, the decision maker lacks certainty regarding the specific payoff they will receive. Subjectively, uncertainty arises from questions concerning one's understanding of the situation, potential hidden complexities, and

Table 1.7.: Quantum Decision Theory (QDT) notation
[H]

| Notation | Definition |
|---|--|
| L_j | Lottery j , for $j = 1, \dots, n$ |
| A_j | Action of choosing lottery L_j |
| $B = \{B_\alpha : \alpha = 1, 2, \dots\}$ | Uncertain events accompanying the choice |

the ability to make the optimal decision. The resulting vector of uncertain items is denoted $B = \{B_\alpha : \alpha = 1, 2, \dots\}$. The choice event A_n is represented by a vector $|n\rangle$ in Hilbert space²¹.

$$\mathcal{H}_A = \text{span}_n\{|n\rangle\}$$

The uncertain event is defined as $B = \sum_\alpha b_\alpha |\alpha\rangle$, in the Hilbert space $\mathcal{H}_B = \text{span}_\alpha\{|\alpha\rangle\}$ ²².

The lottery choice in this case is perceived as a composite event, consisting of the final choice A_n , followed by *deliberations* in the format of the set of uncertain events B . The choice of a lottery is defined as a composite event, denoted a *prospect*, following the convention established in choice modelling under uncertainty disciplines: $\mathcal{U}_n = A_n \otimes B$, which corresponds to a state:

$$|\mathcal{U}\rangle = |n\rangle \otimes |B\rangle$$

Each prospect \mathcal{U}_n is characterised by its quantum behavioural probabilities. At this point the interested reader is suggested to follow the complete methodological presentation of the theory in the work of Yukalov and Sornette (2017), as the precise mathematical formulation of entangled operators requires complex explanations, which would be problematic with the changes in notation adopted.

1.2.3. Concluding remarks

In summary, this section has undertaken an extensive analysis of the available strategies to approach the DCM task, that exist in the literature. Starting from the different paradigms in the approach to choice modelling in general, several theoretical concepts are addressed, as well as the associated modelling approaches. This exploration of the historical and interdisciplinary landscape of DCM offers the reader all the essential prerequisites for further reading.

For each of the two mainstream approaches to choice modelling, be it theory driven or data driven method, a history backed overview of the most popular models is provided. This sheds light on the evolving methodologies, accentuating the ML and DCM techniques advantages and drawbacks. The flexibility and generality of data driven contrast with interpretability of theoretically grounded classic RUM-compliant DCM techniques. The discourse also introduces alternative theoretical frameworks challenging the assumptions of rational behaviour: Expected Utility (EU), Random Regret Minimization (RRM), Random Attention Model (RAM), and Quantum Choice (QC). These theories, rooted in be-

²¹For those unfamiliar, the *Hilbert space* is a vector space H with an inner product $\langle f, g \rangle$ such that norm defined by $\|f\| = \sqrt{\langle f, f \rangle}$ turns H into a complete metric space. A common example is the \mathbf{R}^n space with $\langle v, u \rangle$ the vector dot product of v and u .

²²The $\text{span}\{S\}$ operation appearing in one of the equations is defined as the set of all linear combinations of the vectors in S .

havioural economics and decision-making psychology, open paths for understanding deviations from traditional rationality assumptions.

The introduction of less popular behavioural frameworks allows to demonstrate the issues behind the correct theory choice for each particular use-case. While the data driven models allow the researcher to best approximate the relationships within the data, the interpretability advantage of the theory driven methods remains conditional on the respect of the theoretical assumptions of a particular behavioural framework. This situation perfectly illustrates the issues behind the model comparison and performance comparison task in particular. The number of elements to take into account during the comparison stage is exorbitant and even the generation of a model taxonomy might be unattainable objective in this context.

1.3. Taxonomy issues

As it becomes apparent from the DCM practices and literature overview, the realm of Choice Modelling is rather complex. In the majority of applications, the choice of modelling techniques is closely influenced by the individual scholar's expertise and background. While the multitude of concurrent behavioural frameworks and statistical models are overwhelming, the research is typically directed by the scientists' familiarity with particular methodology and toolset. Proficiency within a specific discipline dictates the preferred methodology and toolkit for implementation. Such state of the literature makes it a lot more difficult to propose a complete methodology for model taxonomy. While having a complete set of all the potential models to be applied in the given context might be one of the scholars main wishes, it is extremely difficult to provide one. Even though it is potentially possible to find a common notation and visual support for the different models, with the increasing model complexity it becomes more and more difficult to preserve consistency in the chosen naming convention. The same may be said about the used vocabulary, as the terminology differs rather significantly between domains. For example, someone with a background in DCM may have difficulties to understand the vocabulary and terminology used in classification tasks by people with ML background.

This state of the literature accentuates the knowledge systematisation problematic. One can question whether a meaningful taxonomy for the ensemble of models *used or potentially useful for individual choice modelling* can be constructed. In this section this exact problematic will be addressed. The first subsection is offering a general overview of existing taxonomies and conventions in model families presentation. The second part presents an alternative vision to the model construction and taxonomy creation methodology.

1.3.1. Existing taxonomies

Many of the existing handbooks and articles in both DCM and classification fields may be roughly separated into two groups. The first ones propose an historical presentation of the modelling techniques development (Hastie, Tibshirani, and Friedman 2009), constructing their discourse starting with some special cases to more general models. The second type of publications offer taxonomy based on subjective criteria (Kotsiantis, Zaharakis, and Pintelas 2006). Both of the approaches are far from optimal when it comes to making a decision about the model to apply in a particular case. In both cases the navigation among the said models is rather difficult for a non-proficient user, who potentially has a limited

vision of the available models families. On the one hand, the methodological historical approach lacks sometimes the versatility and makes it difficult for the reader to decide among unrelated model types. The researcher should have full knowledge of all the historical predecessors of the applied model. On the other hand, a subjective taxonomy may be more instructive, because it offers a clear logical path (or rather a dendrogram) that relates the various models. In this case the arbitrarily created taxonomy offers an even more narrow choice of the techniques to apply. However, the subjectivity of any given taxonomy renders the model choice procedure conditional on the authors principles and vision.

Many of the Machine Learning models as well as Econometric Models may be divided into “*families*”, united by some common factor. Consequently in the literature it is easy to encounter different vision of the *families*, leading to quite different *model space* partitioning. Among the most common separations we may encounter the divisions: (1) by the modelled variable, leading to the quite popular division onto *classification* and *regression* models in ML; or (2) by learning procedure, resulting in another separation into *supervised* and *unsupervised* learning tasks. In the DCM we encounter theoretical model families grouping: (1) by the underlying theories, for example RRM and RUM; or (2) by the assumed distribution of the error terms, starting with the Logit, Probit and other.

In many theoretical papers we may trace the idea of systematisation of the existing models (Kotsiantis, Zaharakis, and Pintelas 2006; Ayodele 2010). All of such works propose a version of systematic review of several models and dividing them into several classes or rather a dendrogram-based taxonomies. One of the eventual drawbacks of such approach is the dependency on the subjective use-case: the order of *branching* of the taxonomy is defined by the author based on subjective criteria. Effectively, some of these criteria in model classification²³ are more or less approved by the community, although no consensus on this topic exists. The opinions widely differ depending on the individual background and experience.

1.3.1.1. Bottom-up approach

Probably the most familiar case of taxonomy presentation, which was faced by everyone at a certain point in time, attempts to build the taxonomy from the most basic concept upwards. A most prominent example in this case would be any type of handbook or of teaching materials. The user in this is gradually familiarised with the concepts of gradually increasing complexity, from the most simple ones and towards the most sophisticated. Such approach to presentation is typically tightly intertwined with the historical representation of the material.

For example, in Figure 1.8 based on Hastie, Tibshirani, and Friedman (2009) we can see a mostly simplified taxonomy oriented towards learners, exploring ML techniques. The learning problems considered in the book are broadly categorized as either *supervised* or *unsupervised*. In supervised learning, the objective is to predict an outcome based on input measures, while unsupervised learning aims to uncover associations and patterns among input measures. Such taxonomy is simple and concise, simplifying the task for researchers unfamiliar with the field, while becoming less interesting for the advanced users. While some mathematical details are necessary, the primary focus is made on the methods and their conceptual foundations rather than their theoretical properties.

²³It is important to clearly separate the notions of *classification* (1) as a ML term standing for a specific classification task performed on a discrete output, from the model classification (2) understood as a procedure of taxonomy creation.

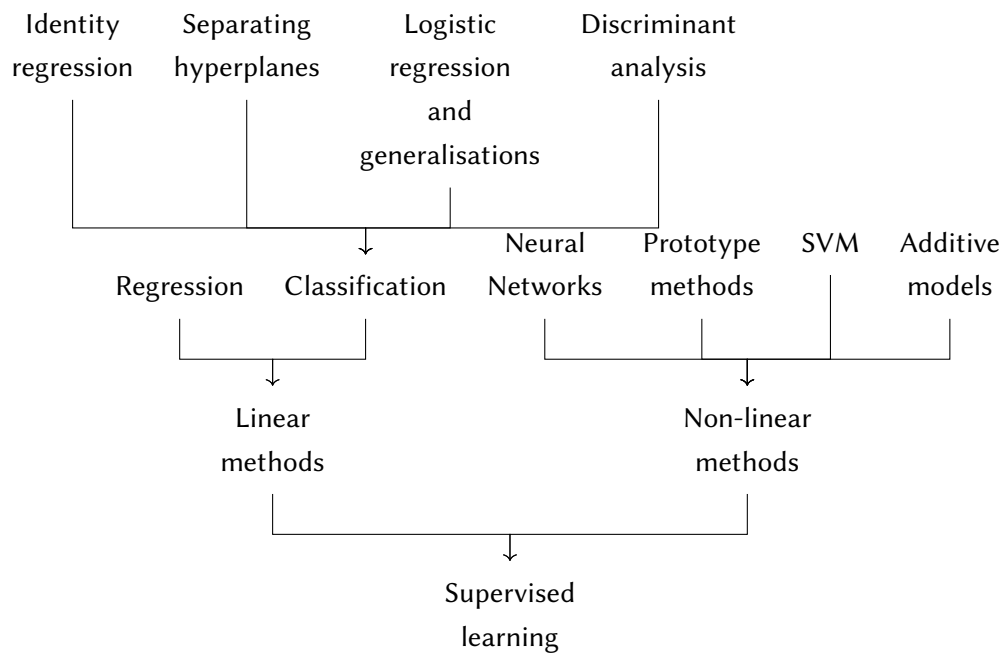


Figure 1.8.: Taxonomy of DCM as proposed by Hastie and Tibshirani (2009), simplified

A similar approach to the taxonomy construction might be encountered in DCM oriented handbooks and guides. For example, among those with econometric background we can distinguish the work of Agresti (2013). The book offers an overview of these methods, including established ones, with a particular focus on GLM techniques, that extend the principles of linear models for continuous variables to categorical responses and multivariate situations. The key principle for DCM distinction is the type of modelled data (Figure 1.9).

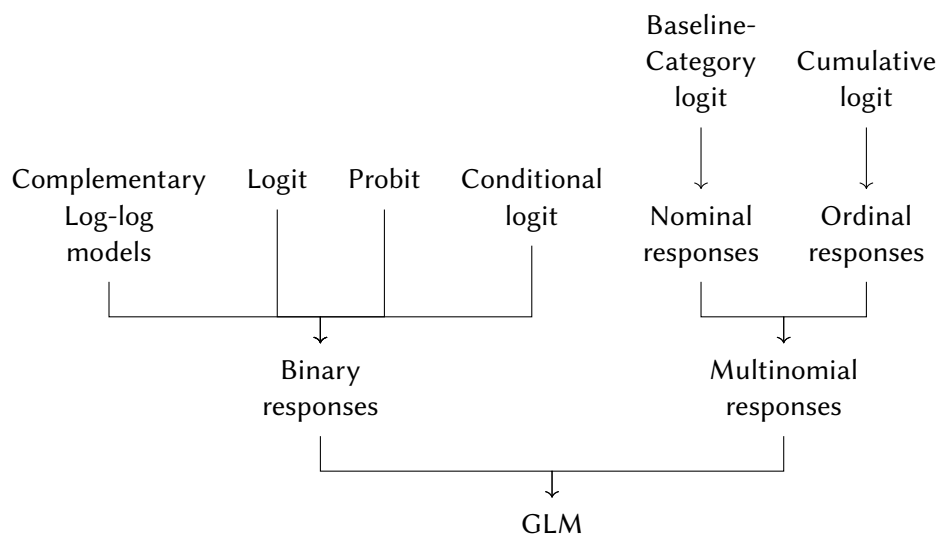


Figure 1.9.: Taxonomy as proposed by Agresti (2013), simplified

The core content of the book covers methods for categorical response variables. It includes sections on distributions and traditional approaches for two-way contingency tables, logistic regression and related models for binary and multiclass²⁴ responses, and explores log-linear models for contingency

²⁴Also denoted as *polychotomous* responses.

tables. Authors also introduce encompassing marginal models and generalized linear mixed models with random effects. The non-model-based techniques for classification and clustering are equally mentioned. The work also offers a historical perspective on the development of all those methods.

1.3.1.2. Top-down approach

Another approach is equally widespread in the scientific community, as it offers much clearer vision of the available elements. In this case the branching structure is inverted, starting with the most general and all encompassing cases or even model families users populate such taxonomies with the particular cases. Such representation facilitates the search for the most appropriate technique to answer a research question at hand, but only provided the user is familiar enough with the proposed techniques.

It is in this section that we encounter the most diverse taxonomies. Personal experience and background of the researchers play a major role in the taxonomy structures they propose. Below we offer two examples.

In Ayodele (2010), Figure 1.10, we can observe a particularly interesting case, where the full complexity and diversity proposed by the fast developing ML field is introduced. Such solution is oriented mostly on the auditory already familiar with ML domain and who searches to get a wider vision of their field.

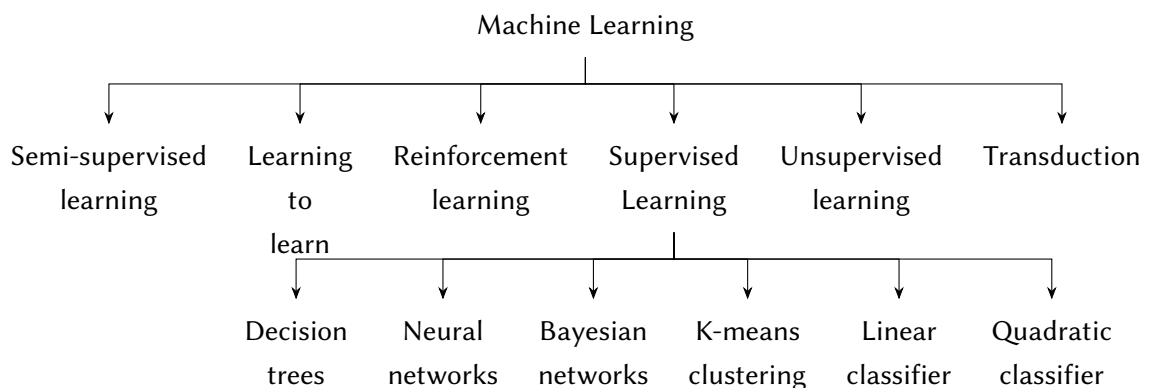


Figure 1.10.: Taxonomy as proposed by Ayodele (2010)

In this particular case the machine learning algorithms are categorized based on the accomplished task. The different learning procedures, or tasks, are presented at this point, which predetermine the The difference from the previously presented work of Agresti (2013) is that *tasks* in this case refer not only to the modelled variable type, as in case of *outcomes*, but rather to the intended use-case.

1.3.1.3. Alternative representation

From the two previous subsections it becomes apparent that most of the existing model taxonomies suffer from the subjectivity bias. Someone may disagree with this statement, arguing that there exist unbiased taxonomies: the historically guided ones. However, even in construction of a historically driven taxonomy the researchers make arbitrary choices. The familiarity with different literature will dictate the same level of subjectivity as in the previously shown taxonomies. Every attempt to construct extensive taxonomy of statistical modelling tools is bound to fail due to the task complexity and the absence of unanimity or even an open dialogue in the scientific community. While some disciplines

manage to sustain some stability in their internal vision on the most widely used model classes and types, the interdisciplinary dialogue remains rather burdened by the incompatibilities of such concept across disciplines.

However, there should exist some alternatives to the basic arborescent taxonomy. With a sufficient input from a large number of practicing scientists it should be theoretically possible to propose a unified framework for modelling technique selection procedure. But will be such a solution sufficiently flexible to reflect all the differences and subtleties of the interdisciplinary study? It is highly probable that not: the created solution will eventually fail following the pitfalls of all the other taxonomies. Logically, it is nearly impossible that one procedure will suit all the use-cases. What is more, even if such solution is able to respond to all the contemporary challenges in the field, it will eventually lack flexibility in future when new needs in research will inevitably arise.

We propose an alternative solution, which aims to overcome all the disadvantages of the previously described one. The main idea in this case is based on the recent understanding of ML models, such as NN, which can be found in S. Wang, Mo, and Zhao (2020) or Welleck et al. (2017). In these recent works we discover the *modular* design of the ML models based on the studies of the loss functions. To be more precise, we may assume that every single model comprises a number of attributes, specifications and transforms, that may be combined and recombined as “*modules*” to obtain new models.

Among the other works expressing similar ideas on the models’ modularity, we encounter the publication of Sokolova and Lapalme (2009), a work written 10 years before. The paper offers a comprehensive examination of twenty-four performance metrics employed across various ML classification tasks, encompassing binary, multi-class, multi-label, and hierarchical scenarios. The study systematically connects alterations in a confusion matrix to specific data characteristics for each classification task. At the end a taxonomy of measure invariance emerges, accounting for all relevant label distribution variations in classification problems.

Among the main features (*modules*) we may identify several recurrent elements, often referenced in applied studies:

- Functional form of the model
- Transformation functions applied to the individual elements during the estimation
- Kernel transformation, also denominated as functional transform in econometrics
- Error term specification
- LOSS function
- Regularisation or penalty transformation
- Modelled variable data type
- Input variables data types
- Input variables transformations prior to estimation
- Dimensionality reduction transformations and data adaptation
- Sampling and resampling strategies
- Boosting implementation
- Variable selection techniques
- Model selection techniques
- Estimation algorithm

The above list includes the most recurrent in literature elements that may potentially affect the model

structure and application use-case. Nevertheless, it may be argued that such partitioning is incomplete, as it potentially lacks some of the most recent and advanced elements. For example, it is rather difficult to include the meta-analysis techniques into this structure, as with each complexity level added the number of potential combinations of the above elements increases. A simple combination of two models in an attempt to perform boosting method puts the number of available element combinations to power of two, which is rather cumbersome to represent. At the same time some may argue that such list is extremely overcomplexified, including the unnecessary elements, belonging to other stages of the modelling task.

Consequently, this representation should be understood as a proof of concept rather than as a direct guideline for model construction and taxonomisation. Further in this work we are going to provide a more in detail analysis of those elements, dividing them into more meaningful groups under statistical modelling task.

1.4. The vocabulary and terminology

The intricacies of vocabulary and terminology when transcribing the theoretical concepts into words pose significant challenge. In economics and choice modelling in particular those challenges become even more accentuated, reflecting the interdisciplinary nature of the field. The intersection of human sciences with mathematics and statistics, as most interdisciplinary studies, often suffer from vocabulary and terminology challenges. Those difficulties are due to the distinct traditions and methodologies employed in concurrent fields. Bridging the gap between the different domains requires a careful consideration of language nuances and the interpretation of shared terms.

For example, in ML, particularly within the realm of data-driven approaches, terms like “*training*” and “*model*” might carry a slightly different connotation compared to their use in economic studies, and DCM studies in general. Another example is the “*bias*” term, as in ML it may refer to the systematic error in predictions, while in econometrics, it could denote the presence of omitted variables leading to estimation bias. Clarifying these terms and their contextual interpretations is crucial for fostering effective communication and collaboration in interdisciplinary research. Establishing a shared understanding of foundational concepts and embracing a common vocabulary becomes paramount to navigating the interdisciplinary intersection of machine learning and econometrics.

Nevertheless, understanding key terms is essential for effective communication and collaboration across various domains, including economics and ML in the case of this study. This section addresses the nuances in terminology, emphasizing the importance of establishing a common understanding to facilitate further reading.

1.4.1. Models and modelling

In the preceding parts we have widely used such terms as *model*, *performance* and *modelling*. Although the concepts might seem self-explanatory and evident for someone with background in ML, statistics or econometrics, there exist some differences between those fields. Those differences come unnoticed while the discussion remains quite *general*, but become drastic as the focus switches to more technical topics, including modelling techniques and their taxonomisation. As an example, illustrating the fact that imprecise definitions could lead to misunderstanding the most general *model* term is chosen. As

is commonly assumed, one of the fundamental notions prevalent in scientific discourse, with certain mathematical connections, is the concept of a *model*. Here are two distinct definitions of a *model* term:

Model (“Model” 2023) a set of ideas and numbers that describe the past, present, or future state of something (such as an economy or a business) ...

Model (“Model” 2022) an informative representation of an object, person or system ...

As one can notice, the first definition remains extremely general regrouping “*ideas and numbers*” describing some concept. The second definition, while still remaining quite general focuses on the model’s purpose: “*an informative representation*” of some object. Those differences may remain unnoticed while the discussion remains generic, but for the purposes of taxonomy creation it may be insufficient. What precisely constitutes a model exactly? Is model an idea, or rather a set of laws and numbers? Should it represent a single object or an entire system? How do those concepts align with the previously introduced ML and econometric techniques? The general definitions provided remain fairly imprecise for taxonomy creation. As there seem to be no consensus on those matters, this subsection attempts to bring some consistency in the terms employed for the remainder of the thesis, laying the ground for subsequent discussions.

Evidently, there exists a lack of uniformity in the terminology employed by scientists across various disciplines. For example, the previously introduced techniques for discrete choice analysis are all denoted as *models* in the respective literature, although these tools are completely different in their nature. This example can be extended to the difference in understanding of *statistical model*’s scope in the different disciplines. While in econometrics the *statistical model* usually means the ensemble of a predefined functional form, excluding sometimes even the estimation technique. The ML community may have a broader vision on the concept including the automated variable selection stage and/or the cross-validation steps. As one can observe, the vocabulary may be extremely ambiguous even within one single discipline, not speaking about their intersection.

Typically, the understanding of the *model* concept remains quite stable and homogeneous within a single discipline. However, even when venturing into interdisciplinary domains, the situation remains obscure sometimes. There, the simple and erroneously obvious concept of *model* indicates distinct entities in dependence of the application field. For example, in economics, or rather in econometrics to be more precise, we immediately face two different concepts of *model*: (1) the idea of *theoretical model* reflecting the assumption of some hidden deterministic relationship between the variables, and (2) the concept of the *statistical or econometrics model*, which attempts to approach the hidden pattern. While such subtleties are quite evident and self-explanatory for the econometricians, this distinction may become quite confusing for someone with a basic Statistical Learning background. Thus, even restricting the scope to a single discipline we may observe inconsistencies. In the works written by economists, outside the context of interdisciplinary studies, we encounter a multitude of references to various model types: *econometric models*, *theoretical models*, *economic models* and many more.

Before proceeding with the discussion of subtleties in *model* term definition in the context of the DCM, behavioural studies and statistics, we should look at the more basic definitions. To do so, we may refer to some of the most general dictionaries (“Model” 2023), basic handbooks on maths, or better, the handbooks about teaching maths (Gardiner 2016). Because, the concept of the *model* is considered to be something extremely basic, one will never encounter its definition or redefinition in advanced literature. It is typically assumed, that readers are familiar with such terminology and

already understand the sense commonly attributed to the term in the given community. The proposed previously definitions are rather simple and intuitive for the beginners. However, the matters become complicated relatively quickly after the switch to the cross-disciplinary sphere. The definitions in this case are not that easily transferable from the disciplines' backgrounds, but represent a set of sometimes mutually-exclusive concepts.

In econometrics, ML and statistics in general, we may define the *model* from different perspectives. This fact roots in the ideology which is behind each of the disciplines. To better illustrate this discrepancy, we may refer to the work of Baltagi (2008), which perfectly contrasts the different approaches. On the one side, we have the Econometrics, attempting to estimate parameters of a presupposed function. This approach allows to extract information from the resulting estimates, assuming the consistency of the primary assumptions. On the other side, the ML approach ignores all the priors, except the basic statistical assumptions essential to make the algorithms work. The ML techniques search to find the functional form that best fits the observed data, or, if we focus on the consequences, the function which produces the best predictions.

Here we face the appearance of both *inter-* and *intra-*discipline paradox of the terminology inconsistency, partially introduced previously. The *intra-*discipline inconsistency appears when inside the single domain (ex: econometrics) we encounter two different terms containing the noun *model*. The first one, theoretical model refers to a priori selected *model*, which is assumed to correctly describe the *behaviour* of agents in the observed system. The second one, econometric model defines the statistical translation of the previously assumed behavioural model. Another inconsistency, the *inter-*disciplinary one, accentuates the differences in the understanding of the same terms among the disciplines. For example, between econometrics and ML. As previously described, the disciplines have different aims and objectives, which induce the different understanding of the terminology. In this particular example, we have already partially defined the econometric model term. The ML model assumes the statistical representation of the *unknown* function, which involves not only the parameter estimation, but may incorporate many more steps in the *model* construction task. Among such tasks one can imagine the functional form family selection.

On this point, we should focus on the specification of the statistical model term. Due to the specificities of *model* construction procedure and the multitude of the available *modelling* techniques, it is extremely complex to clearly delimit the statistical model. For example, in the case of the simplest Ordinary Least Squares regression the statistical model term seems rather self-explaining. But it is not that simple as it seems, one should consider whether or not the *model selection* phase should be included into the *model* definition. Given the intersection of the Econometrics and ML domains, we can imagine that the resolution of this question relies on the familiarity with one or another discipline.

Finally, before proceeding with the topic of *model* construction, we should define the process of *modelling*, which immensely depends on the *model* concept in use. It is quite obvious, that the *modelling* in the context of behavioural model construction will be different from the one implemented in the construction of an econometric model. Among the most used definitions we may cite the one available on internet. Evidently, it is far from the best available source, but in the context of finding the most generic and widely used term it should suffice:

Mathematical model (“**Mathematical Model**” 2021) a description of a system using mathematical concepts and language ...

Mathematical model (“Mathematical Model” 2023) a mathematical representation of reality ...

Both definitions focus on the *mathematical modelling* in particular, but those definitions can be easily altered, substituting the *mathematical* part for any other relevant term. In other words, we can understand the process of *modelling* as the procedure of developing a *model*, regardless of the *model’s* type. Consequently, further in this work the terms *econometric modelling*, *statistical modelling*, *behavioural modelling* and *ML modelling* will be understood according to the above definition. All of those elements being interpreted as field specific descriptions of systems with usage of mathematical concepts and language. In most of those cases, the model assumes the description of underlying, *econometric* or *behavioural* concepts for example, with the help of mathematical concepts.

Focusing on *statistical modelling* we may return to the works of Hastie, Tibshirani, and Friedman (2009) or Breiman et al. (2001). The first book represents the authors’ effort to consolidate and explain many of the significant new ideas in *learning* within a statistical framework, addressing the *Statistical Learning (SL)* topic. The second works contrasts the difference in understanding of *statistical modelling* in different disciplines. However, both of the works do not offer any definition concise enough to be adopted at this point. At the same time the general literature typically simplifies the concept of *statistical model*.

Statistical model (“Statistical Model” 2021) is a mathematical model that embodies a set of statistical assumptions concerning the generation of sample data (and similar data from a larger population) ...

Another extreme is given by a formally mathematical definitions. One of such works addressing the *statistical model* definition is the publication of McCullagh (2002), where the authors attempt to give a more formal mathematical aspect to the *statistical model* definition. In this case the definition appears nearly synonymous to the *probabilistic model* concept, as the focus shifts to the probability distributions and their underlying parameters.

Statistical model (McCullagh 2002) is a set of probability distributions on the sample space \mathcal{S} ...

Returning to the *modelling* concept, such treatment of the theoretical terminology is a result of oversimplification, but it brings more consistency into the work. To explain, why once again we are speaking about oversimplification, we may look at the *econometric modelling* term, which may be understood differently in dependence on the application. Some economists may treat the *econometric modelling* process as the entire framework of *model* construction. This predominant approach means incorporation of both: (1) *behavioural* (or *theoretical*) *modelling* part, as well as (2) the *statistical modelling* counterpart. Other may assume that *econometric model* is a discipline-specific equivalent of the *statistical model*. Finally, a fraction of scientists may argue that *econometric modelling* should also include the information on data collection and treatment in its concept. The same reasoning may be applied to other terms, which in function of the application domain and situation may have different underlying meaning.

1.4.2. Performance

Identically to the differences in the *model* term understanding across different disciplines and application cases, the understanding of the *performance* term vary across the applications. For example, in econometrics, and more traditional DCM applications, the model performances usually refer to the

absence of bias in the estimates, as well as the precision of those estimates in terms of the observed variance minimisation. In the ML field the performances imply the best predictive accuracy of the model in a particular use-case. Finally, in *informatics* and *computer science (CS)* related domains²⁵ the performance may be understood in terms of the computational efficiency and resource consumption minimisation. These variations are due to the specific characteristics and requirements of each field.

In this subsection we are going to explore more in detail each of those use-cases for a better understanding of the conceptual difference in problematic that arises in each of those cases. This will help us to correctly define the performance for the purposes of this work.

1.4.2.1. Estimates and derived indicators

A perfect example of the econometric background work is the publication of Belgiawan et al. (2019) entitled: “*Context-dependent models (CRRM, μ RRM, PRRM, RAM) versus a context-free model (MNL) in transportation studies: a comprehensive comparisons for Swiss and German SP and RP data sets*”. In this work the authors examine the RRM model, which assesses alternatives based on their relative performance, making it context-dependent. Due to the theoretical assumptions differences, in RRM individuals are assumed to make choices aiming to minimize expected regret rather than maximizing utility. This model includes three variations: classical CRRM, μ RRM, and PRRM, along with another approach known as *Relative Advantage Maximization (RAM)*. The performances of the models are then assessed not only in terms of their predictive qualities, but also their capacity to correctly identify the various economic indicators over a number of testing datasets.

The authors conduct a comparison between multinomial logit and these four alternative models using stated choice datasets covering various decisions like mode choice, location choice, parking choice, car-pooling, and car-sharing. The evaluation of these models focuses not only on the model fit, but also on the estimates, calculated values of travel time savings (VTTS), and elasticities. The last two elements representing a class of derived metrics, which are computed with reliance on estimates. Among the obtained results, authors illustrate that the estimates, VTTS and elasticities exhibit substantial variations. Those observations hold significance for cost-benefit analyses and simplified modelling approaches, as the estimates are then reused to support strategic decision making in economic policies.

This case illustrates how the estimates might be an object of model performance evaluation. Many other theoretical econometric works rely on simulation (Haghani and Sarvi 2019; Lorenzo Varela 2018) to test the reliability of the proposed modelling strategies. For the end user in economic applications, the capacity of the model to correctly identify the individual effects play crucial role, as those estimates are then employed in support of strategic decisions.

1.4.2.2. Predictive qualities

As an example of the work with SL and ML background we may point out the work “*A comparative study of machine learning classifiers for modelling travel mode choice*” by Hagenauer and Helbich (2017). There authors emphasize the significance of analysing travel mode choice in the context of transportation

²⁵ *Informatics* is a broader field that encompasses the study of information, its processing, and the interaction between people and technology, often emphasizing applications in various domains. *CS*, on the other hand, specifically focuses on the theory, design, and implementation of computer systems and software, delving into algorithms, programming languages, and hardware architecture.

planning and policy-making, providing an overview of different model performances in mode choice analysis from predictive perspective.

The authors argument this comparison by the assumption that ML powerful classifiers' applicability in modelling travel mode choice has been largely unexplored. Utilizing comprehensive Dutch travel diary data spanning from 2010 to 2012, augmented with factors related to the environment and weather conditions, this study conducts a comparative evaluation of seven distinct ML classifiers for travel mode choice analysis and provides insights for model selection.

The models are primarily compared based on their predictive power and overall goodness of fit. Those elements are assessed using the generalised information criteria²⁶ as well as the general accuracy measures. Additionally, the authors delve into the importance of various variables and their associations with different travel modes. The findings highlight the substantial superiority of random forest over other investigated classifiers, including the widely used multinomial logit model. Trip distance emerges as the most critical variable, but the significance of other factors varies depending on the classifier and travel mode. Meteorological variables are particularly relevant for the support vector machine, while temperature plays a vital role in predicting bicycle and public transport trips. These results underscore the necessity of assessing variable importance concerning diverse classifiers and travel modes, enhancing our comprehension and efficacy in modelling people's travel behaviour.

1.4.2.3. Resource efficiency

Finally, as an example of a CS oriented work we can turn our attention to the work "*xlogit: An open-source Python package for GPU-accelerated estimation of Mixed Logit models*" by Arteaga et al. (2022). Where authors focus their attention on the software implementation of a Mixed Logit model, MMNL being a powerful tool for studying choices, but it involves complex computations due to the need to simulate integrals for estimation.

Because specifying Mixed Logit models involves decisions like selecting explanatory variables and their mixing distributions, which is time-consuming and computationally demanding, the authors introduce *xlogit*, an open-source Python package that harnesses the power of *Graphics Processing Units* (GPU) for efficient Mixed Logit model estimation. They compared *xlogit*'s performance with other Python packages like *PyLogit* and *Biogeme*, as well as R packages like *mlogit*, *Apollo*, *gmm1*, and *mix1*, using both artificial and real data.

In the paper they compare the performances of their software for MMNL estimation with concurrent implementations. The obtained results indicate that, with a mid-range graphics card and a standard desktop computer, *xlogit* is on average faster than concurrent solution as indicated in the Table 1.8 (Arteaga et al. 2022).

Table 1.8.: Performance gains (computation time) by Arteaga et al. (2022)

| Software solution | Performance gains |
|-------------------|-------------------|
| Apollo | 55 times |

²⁶Generalized information criteria usually include the *AIC* and the *BIC*. These criteria balance the goodness of fit of a model with its complexity, providing a quantitative way to compare different models and select the one that best balances explanatory power and complexity.

| Software solution | Performance gains |
|-------------------|-------------------|
| Biogeme | 43 times |
| gml | 74 times |
| mixl | 39 times |
| mlogit | 16 times |
| PyLogit | 27 times |

Moreover, the authors illustrate that `xlogit` manages memory in a more efficient manner than other software. These performance improvements streamline the modelling process, allowing for more efficient testing of various model specifications compared to existing software packages. The `xlogit` package, along with its open-source code, documentation, and usage examples, are made publicly available by authors on the project's GitHub repository.

1.4.2.4. Generalisations

In the general culture, the “*performance*” refers to the effectiveness and quality of a predictive model in accurately capturing and representing patterns within data. Nevertheless, assessment of model performance depends on the specific objectives and requirements of the application.

The diverse understanding of the “*performance*” term across different disciplines and application contexts accentuates the difficulties associated with model performance evaluation. Different disciplines and applications may prioritize distinct aspects of performance, such as: (1) minimizing estimation bias and variance in econometrics, (2) optimizing predictive accuracy in ML, or (3) enhancing computational efficiency in informatics and CS. Recognizing these variations is particularly important in the context of this study. In the subsequent chapters those concepts would be reused for the model performance strategy definition.

1.5. Conclusion

While in economics the DCM techniques are widely used, the same toolset might be encountered in other domains and disciplines as well. Biology and geo-sciences, as well as many other classification applications implement techniques to the ones encountered in economics related DCM studies. Even within economics the different implementations and problematics addressed through DCM toolset fraction scientific community following their very specific research needs. Transportation, health economics, marketing and policy analysis - the DCM has many use-cases focused on individual behaviour analysis. Choice models serve as indispensable tools for economists and policymakers, enabling them to forecast demand, optimize tariffs, design effective marketing strategies, and enhance public policy decisions.

In the context of traditional economic related applications the DCM framework has proven to be an indispensable and versatile approach for studying individual behaviour. However, with the development of more advanced statistical approaches for supervised classification and the integration of MLg techniques, the landscape of choice modelling is expanding. Not only the traditional DCM toolset is expanding through generalisation and extension of classic models, the attempt to combine this method-

ology with more flexible ML tools are made. And while the introduction of new ML base methods encounters criticism for sometimes insufficient alignment with the economic theories, it is undeniable that such tools offer improved predictive capabilities and flexibility.

However, the growing diversity of data analysis strategies poses challenges, particularly for non-experts simply seeking to choose the optimal solution for their choice modelling tasks. The availability of a multitude of novel techniques and the heterogeneous background of these methods, combined with their relative complexity and requirements in terms of prerequisites to end-user, create a complexity that may impact the model selection process. This challenge is further impacted by the data requirements, that differ across all the available methods, imposing restrictions on the research procedures to adopt. This process may be understood as conditioning of the modelling technique usage on the data and associated assumptions about the population and individual properties.

This chapter emphasizes the need for a better understanding of the strengths and weaknesses of various choice modelling approaches, especially in the face of different economic questions. To illustrate this, an attempt is made to familiarise the reader with the various models issued from both DCM and ML background, with a particular focus on a unified notation for those methodologies to simplify the performance comparison task in future. The ability to compare and contrast modelling approaches becomes crucial in navigating the complexities introduced by interdisciplinary studies and the integration of new statistical techniques. Finally, this chapter underlines the importance of unifying terminology and establishing a common vocabulary to facilitate communication and comprehension across diverse backgrounds.

The economics field, and the DCM in particular, continues to evolve with the incorporation of ML techniques and the expansion of interdisciplinary studies. This chapter sets the stage for following discussions on performance comparison and taxonomy construction, setting the common ground for future discussion in focusing on nuances associated with vocabulary and terminology in the context of DCM. A familiarity with those elements is expected from the reader to proceed with the next chapters. The understanding of these foundational elements is essential for researchers and practitioners seeking to navigate the complex landscape of choice modelling, regardless of the exact application environment.

2. A universal performance comparison framework

Performance comparison in the literature emerges at various levels, primarily addressed by theoretical works that provide sufficient information on the subject. However, the understanding of model performance varies across different applications, with a prevalent focus on predictive qualities and goodness of fit. To tackle the issue of inconsistency of theoretical base among the different application fields and knowledge acquisition strategies, we propose a universal approach for the model usage, exploration and performance analysis.

The proposed framework for performance analysis and comparison is based on the standard scientific procedure, with sufficient flexibility to extend it to other fields and disciplines. The adopted procedure may be seen as quite close to many applied and theoretical economic papers. Unfortunately no known to us work approaches the scientific workflow from the same perspective and in such detail as proposed in this thesis.

In this chapter, we offer an overview of the research procedures encountered in the literature, which shape the frameworks structure. The discussion is provided for each of the key elements of the research procedure: (1) performance comparison issues and target metrics choice; (2) data associated limitations; and (3) modelling part of scientific procedure. The data processing related elements are then regrouped into the data analysis stage of the framework, incorporating information about theoretical assumptions, data acquisition and statistical modelling techniques. The theoretical foundations being presented in the previous chapter, a focus is made on data acquisition and processing related issues, or data analysis, as well as the concepts of the research procedure integrity.

At chapter's core the **Section 2.5** serves to introduce the complete framework, putting the puzzle of individual elements together. Once the various elements of data analysis stage united, the complete performance comparison framework is presented. The **Section 2.6** offers a perspective on the framework's alignment with the existing practices, illustrating its capabilities according to the classic literature. A more detailed illustration of framework's capabilities is offered in **Chapter 3**.

2.1. A need for unified methodology

In the preceding chapter 1, the reader was introduced to the fundamental concepts and challenges addressed in this work. This allowed us to demonstrate the convergence point for interdisciplinary discourse, bridging the gap between theory-driven and data-driven studies. Despite various attempts to reconcile these approaches, a consensus on effective gap mitigation strategies has yet to be reached.

Addressing the challenge posed by the inconsistency of common ground across diverse application fields and knowledge acquisition strategies requires the introduction of a user accessible approach for the exploration of model performances. Current literature offers no unified methodology for model performance evaluation, explained by the heterogeneous understanding of performance concepts across different application domains, as described in the section 1.4.2. A new *framework* is designed to bridge the gaps and develop a more unified understanding of model performance across various disciplines and research methodologies. The framework aims to provide a flexible toolset for assessing and comparing the effectiveness of different modelling techniques and their applicability in diverse contexts. This approach attempts to harmonize the perspectives of theory-driven and data-driven studies, facilitating the dialogue within the interdisciplinary community.

The model performance analysis and performance analysis in general serve as a foundation for this chapter. The comprehensive analysis is essential as it encompasses the very essence of both the objectives and methodologies underpinning data analysis and modelling tasks. There exist several studies addressing those issues in a separate manner, but rarely the discussion takes into account all the available dimensions of the problematic. However, the model performance and their analysis cannot be adequately captured in isolation, the integrity of data analysis stage and even the purposes of the scientific task affect the performance perceptions.

Thus proposed framework for performance analysis and comparison is based on the standard scientific procedure, with sufficient flexibility to extend it to other fields and disciplines. The adopted procedure may be seen as quite close to many applied and theoretical economic papers. Unfortunately no known to us work, except for the manuscript of Williams and Ortuzar (1982), does approach the scientific workflow from the same perspective as us. While many research papers implement the ideas similar to the proposed framework, we are not aware of any that make a particular accent on the procedures' systematic part.

In particular, in their work Williams and Ortuzar (1982) address *behavioural theories of dispersion and the mis-specification of travel demand models*. In particular, through introduction of performance comparison framework focused on policy implications, they illustrate how misspecification in choice-set generation can bias model parameters in mode choice models. The ideas presented in their work have a long lasting impact on the DCM analysis landscape. Gonzalez-Valdes, Heydecker, and Ortúzar (2022) reference this fundamental work in the context of dataset simulation for ICLC model performance assessment purposes. Vij and Walker (2016) explore the cases where ICLV models are useful, illustrating their point of view through simulation. Bahamonde-Birke and Ortúzar (2014) use the proposed simulation procedure for HCM models' capabilities exploration and testing. However, most of the works focus solely on the concept of simulation usage for performance analysis, ignoring sometimes the concept of performance evaluation in the context of public policy implications.

In this chapter we are going to describe the proposed approach in detail, element by element, thus

addressing such dimensions as: (1) performance analysis and comparison, (2) model selection and (3) data management. The key concepts for understanding of this chapter's contents were presented previously and should any misunderstandings arise, it is advised to consult the first chapter. All the above elements will then be reunited in a performance comparison framework. The chapter will end on frameworks use-cases description.

2.2. Performance comparison issues

In the general literature on both economics (Costa et al. 2007; K. Train 2002; Andrews and Manrai 1998) and ML (Flach 2019; Hand 2012) the topic of performance comparison arises on many different levels. However, only theoretical works present sufficient information on the performance comparison in their publications (García-García et al. 2022; Belgiawan et al. 2019). What is more, as it was illustrated in Section 1.4.2, the understanding of model performance differs across different applications. The perceptions of performance concepts vary in accordance to the background of the researchers and the addressed research questions. In the theoretical literature one may encounter some rather complex performance metrics (Japkowicz and Shah 2011), incorporating rebalancing and various transformations. Nevertheless, the most common vision involves the direct comparison of the models' performances based on the predictive qualities and the goodness of fit (Liu and Xie 2019). Only a fraction of works, among which are mainly encountered the econometrics oriented theoretical works focused on effects identification (Lewbel 2019), go outside the scope of plain prediction focused metrics and focus on the indicators more suitable for economic questions treatment.

The DCM literature is not an exception. The comprehension of the models' performance and performance in general differs drastically among the authors. While the classic DCM studies (Belgiawan et al. 2019; M. Bierlaire, Bolduc, and McFadden 2008) offer some visibility of models' behaviour in terms of the capacity to correctly identify the effects, most of the community still focuses on the *goodness of fit* oriented metrics (Jong et al. 2019; Askin and Gokalp 2013). According to the interdisciplinary works (Japkowicz and Shah 2011) the performance of competing models may be assessed over several criteria: (1) quality of data adjustments; (2) predictive capacity; (3) quality of the field specific (ex: economic and behavioural) indicators derived from estimates; and (4) algorithmic efficiency and computational costs. We are going to return to this discussion further within this section.

If we were to design a framework which would have the performance evaluation and comparison for the main task, its starting point should be tightly tied to the performance understanding and definition. In the previous chapter various performance perceptions encountered in the economics literature on discrete choice analysis were introduced. In this chapter a closely related concept of *target metrics* is introduced. The *performance* hence could be redefined in relation to those *target metrics* and *research question* in general, which becomes particularly important at this stage of the work. This definition unlinks the performance concept from the plain statistical model and attaches it to the research procedure as a whole. Consequently, a portion of this subsection is devoted to introducing of the *research procedure* concept. As illustrated in the Section 1.2.1, in economics, the explored questions impose target metrics. Those metrics are context dependent and are inseparable from the research question definition. This idea made appearance in the work of Willems and Polderman (1998), in the context of model performance evaluation applied to public policies. However, the previous findings omitted a number of other elements that impact performance perceptions and understanding. This

work contributes to the literature in completing the gaps in providing a more complete vision of the research task.

In this subsection, attention will be directed toward the performance-related aspect of the framework. The initial step involves introducing the concept of the *scientific procedure*, including the importance of the adopted scientific procedure on the performance and results perception. Those will be closely followed by a more detailed exploration of the available performance metrics and performance evaluation approaches. The particular focus will be given to the field specific indicators in the context of the individual behaviour analysis. This groundwork will lay the foundation for the performance analysis and comparison framework.

2.2.1. Scientific procedures

Before proceeding with the construction and description of the framework, the *research procedure* concept should be introduced in more detail. The proposed framework attempt to provide a simple tool-set for discrete choice model testing and comparison. Hence we should start with the traditional scientific procedure description, its *academic* version, in the first place. Those reflections will be extended with a discussion based on the results of a *non-structured interviews*¹ results. A series of 23 non-structured interviews was performed among the practising researchers in France and Canada. The interviews served to explore the common practices in data analysis and discrete choice models application in particular. A more detailed description of the adopted procedure is proposed separately.


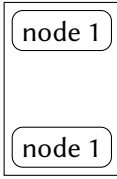



In the literature, regardless of the actual case, all the research takes its root in some problematic: a question to be answered, a barrier to be overcome. Once the task delimited, there are different strategies on how to proceed. Some of them are conventional and described in every practical guide (Agresti 2013) or used in meta-studies (Haghani, Bliemer, and Hensher 2021), while other are more obscure and are sometimes criticised for uncommon practices. As one can imagine, those topics addressed here are mainly discussed in the epistemological works, rather than in more abundant applied studies. Nevertheless, it's extremely important to have the general understanding of the typical procedures and paths implemented in applied research to make the next leap towards framework construction.

To offer an example, an overview is provided of the classical procedure commonly employed in economics, and similarly adopted in various other academic domains. In outlining the components of the scientific procedure, the works of Wooldridge (2012) and Baltagi (2008) might be referenced. Those are the classic works of reference in econometrics and economics: the first one being the general introductory reference to econometric analysis, while the second one is a more advanced guide into the panel data modelling.

First of all, every research starts with a *problematic* identification and *research question* definition. Every study begins with a particular need - a problematic to be addressed. The first steps reflect the transition of the real world problem to be treated into the more restricted context of a research specific question. The next stage in the research requires the researcher to make some assumptions about the nature of the data and the underlying processes. Thus the research question definition may be sometimes dependent on the data analysis procedure, as the question may be altered under the influence of preliminary findings and the results of data exploration. Typically it's during this stage that hypothetical interaction model is defined based on the theoretical assumptions or the preliminary analysis of

¹A more detailed description of the interview conduct procedure is presented in Appendix D.

Table 2.1.: Graphical representation conventions
[H]

| <i>Colors</i> | |
|---|-----------------------------------|
| <i>red</i> | Accented elements |
| <i>black</i> | Regular elements |
| <i>Shapes</i> | |
|  | An element (concept) |
|  | Group of elements (concepts) |
| <i>Line-types</i> | |
|  | Strong (essential) relationship |
|  | Standard relationship |
|  | Weak (complementary) relationship |

the available, in case any is available, data. The general concept of the research question interactions with the *data analysis* procedure may be represented as in Figure 2.1. The research procedure in this case consists of two key elements: (1) the research question itself and (2) the data analysis procedure. Given the context of economic studies, we may restrict the first element to purely economic questions, but in general the addressed problematic might not be discipline specific. Both are tightly linked together with mutual influences, as the research question dictates the requirements for data and analysis methods, while the available data and toolset impose the limitations on the research question.



Figure 2.1.: Research procedure in applied studies

At this point the graphical conventions should be introduced to facilitate further reading. The Table 2.1 enumerates the different graphical elements appearing in this chapter and their intended meanings.

Following the results of the unstructured interviews, the Figure 2.1 may be actually extended one level further, as the researcher has rarely all the desired liberty in the scientific procedure construction. The adequacy of dataset to provide an answer to the research question is essential and typically imposes restrictive limitations on the study. The data is rarely freely available, with some exceptions. What is more the setting of the economic question may usually impose specific requirements on the dataset, which leads to collection of new data. Usually the open access datasets are made available for public once the main research is already performed over the dataset and the results are already published. While for theoretical work an already explored dataset is rather an advantage, as there are already some targets available for the particular application, in the case of applied research the situation is reversed. This means that open datasets represent lesser interest for the applied research due to their lack of novelty. This fact makes the applied research dependent on the influx of the new novel and

unexplored data, which is typically: (1) expensive and/or (2) difficult to collect. In both cases an access to supplementary funds or external assistance is required for the data acquisition. This pushes the researchers to collaborate, or rather depend, on the external entities. Those entities are typically represented by research institutions, public or private research organisations. In either case the financial aid or data availability conditions the research to supplementary limitations or requirements. This may be reflected in schematic format as in Figure 2.2.

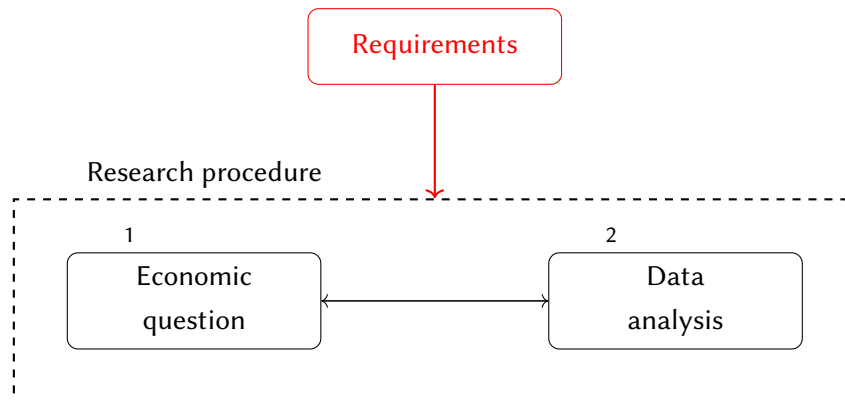


Figure 2.2.: Research procedure, subject to external limitations

The research process is typically influenced by a multitude of external factors that can significantly shape its trajectory, obtained results and societal impact. Externalities in research procedures typically refer to unintended effects or consequences that arise from the research process and extend beyond the intended scope or objectives of the study. As mentioned the financial aspects typically play the most crucial role in the adopted research procedure in many cases. Government policies and regulations also drastically affect the research objectives, methodology, legal and ethical considerations. Those are followed by the technical limitations, access to the required materials, resources and infrastructure. Finally, the societal attitudes and cultural norms contribute to the context in which research is conducted. This last element affects the research topics selections and the public reception of findings, alongside with the reception of the used toolset within the targeted community.

In the case of theoretical studies those externalities (Figure 2.2) may be viewed under different perspective. In fact, there exists a shift in studied object from the applied studies to the theoretical studies. While the applied studies seek to provide an answer to the posed economic question, the theoretical research focuses on the path to this answer, searching to optimise some stages of the procedure, or provide some insight on how the theory affects the obtained answer. This way the object of most applied studies becomes a subject for the theoretical studies, which can be reflected in the similar manner on the research procedure schema, as the influence of the external factors. The researcher in this case will frame and limit the core *research question*, as well as the *data analysis* stage following the objectives of the theoretical study.

Lets now zoom in once again onto the research procedure and the interactions between the research question and the data analysis stage. After the approximative definition of the research question comes a further development of the explored *problematic*, the narrowing and translation into numerical terms: target *metrics* identification. Those *metrics* should allow the researcher to answer the research question (Figure 2.3). For example, one may be interested in causal effect detection, which may be translated into the analysis of particular coefficient significance in an econometric model. Another example is

the prediction task: researchers may be interested to offer the best prediction of consumer behaviour, which may be translated into comparison of various performance metrics for different predictive models.



Figure 2.3.: Target metrics in research procedure

An even more complete representation would include the eventual answer to the economics question. The interconnection in the form of target metrics between the question and the data analysis step assumes mutual influence of the elements through a mediator of target metrics. The target metrics act as a bridge between the research question and the data analysis step. It serves to transcribe the theoretical inquiries into measurable outcomes and at the same time derive a comprehensive solution from a number of estimates. Thus the target metrics not only shape the data analysis strategy, but also influences the way findings are interpreted in the context of economic principles and policy implications. Figure 2.4 adds this second step of the research questioning and data analysis relationship. For better illustration of this relationship a distinction is made between “*target metrics*” and the eventual “*answer*” to the research question. Nevertheless, the reader should keep in mind that both of those elements are tightly interlinked.

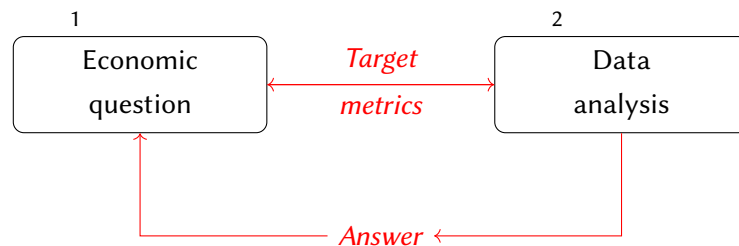


Figure 2.4.: Research procedure, complete

While we separate the target metrics selection stage in our discussion, it should be understood, that in the applied studies the researchers rarely distinguish this stage of the research procedure. Typically the process of target metrics selection passes in such a natural manner for the scientists, that it stays unaccounted for in many discussions. In the theoretical studies the landscape is rather different, as the target metrics constitute the main objective for the research purposes. For example, in their work Newman, Ferguson, and Garrow (2013) explore the precision of the model estimates for the transportation mode choice analysis in the presence of censored data. In this particular case the *effect estimates* are separated as a separate criteria of a model quality. Thus the target metrics play the role of a mediator between the research question and the data analysis elements of a scientific procedure.

The *data analysis* stage of the scientific procedure may be further divided into several major steps. Even with the target metrics assumed to be fixed, which is rarely the case, the research may proceed differently, depending on the available information. Without loss of generality this step may be summarised as *data analysis* process (Figure 2.5). In the discussions with practising researchers three key elements were recurrent: (1) theoretical foundation, incorporating theoretical model structure and assumptions

over the real world state; (2) data, alongside with its collection and pre-processing methods; and (3) statistical models and other analysis methods. The individual aspects present on the Figure 2.5 made their appearance in the Chapter 1. For example, either the actors already have access to some data and build the model using available information, or the model is prebuilt and drives the data collection step. Finally, the data analysis provides the actor with information on the *target metrics* (estimates). Those stage the setting for solution identification to the initial question, providing an *answer* to the initial problematic. All those steps are summarise in the figure 2.5

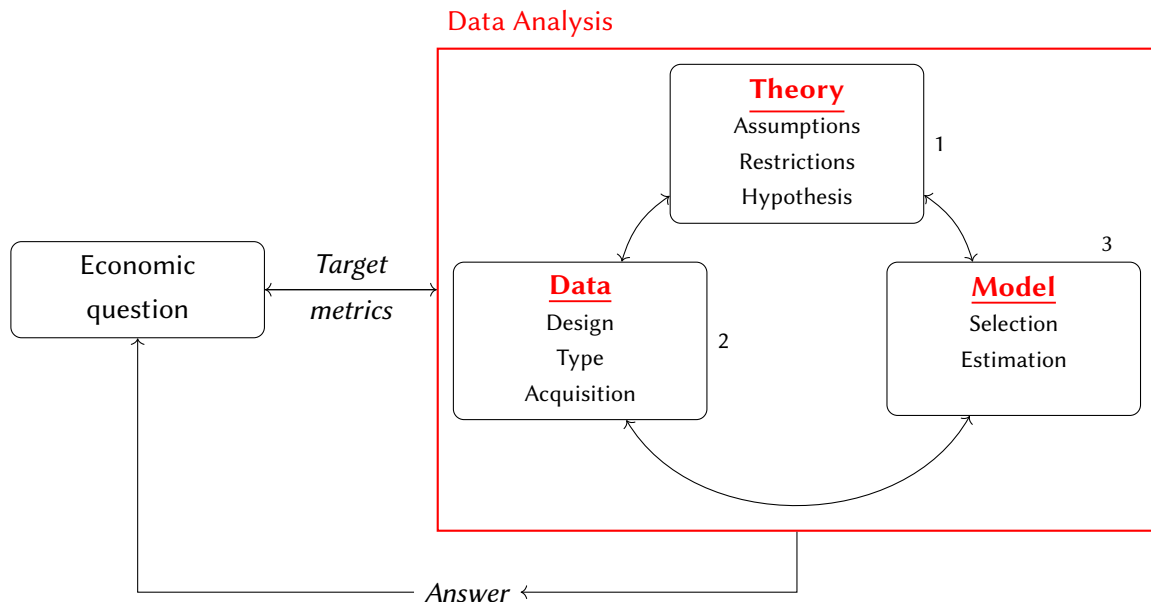


Figure 2.5.: Data analysis

As one can see it's rather difficult to establish an all-encompassing schema of the traditional scientific procedure. Even though at first glance it may appear rather simple, the mixture of all the available possibilities that are open to the scientist is extremely large. What is more, the applied studies evolve differently in dependence of their intermediate findings and eventually uncovered limitations, not foreseen at the start of the research process. Even the minor discrepancies between the available data and theoretical model may require a full revision of the data analysis strategy, thus restructuring the research procedure. Through a series of unstructured interviews, presented in more detail in Appendix D, we managed to identify three radically different strategies. The researchers may start their work with either: (1) theoretical model, (2) data or (3) modelling and statistical analysis. Unfortunately, for the reasons of results anonymisation we are unable to cite the exact works to which the participants were referring during those interviews. Nevertheless, to provide some illustrations we reference the existing literature wherever possible.

The theory grounded studies are predominant among the highly ranked publications as it was illustrated in section 1.2.1. In this case the theoretical assumptions and preliminary hypothesis listing comes prior to the data collection, which respects all the best research practices and guidelines. Unfortunately, this type of studies is among the most resource consuming ones, as they require a rather extensive preparation and prior validation of all the assumptions. Among some examples of those studies we may list the experimental economics or psychometric studies (Blacklow, Corman, and Sibly 2021; Geržinič et al. 2021), where the data is collected under prior assumptions in a restricted environment. As an another example we may face various stated preference based studies (Ojeda-Cabral, Hess, and

Batley 2018; Ben-Akiva, McFadden, and Train 2019), where the data is collected via a specially designed survey. Obviously in both cases we may encounter exceptions, as the factors driving the study may vary and different levels of complexity might be imposed.

Other studies take their root in the data, once its available on the early research stages. It is difficult to provide any examples for this particular case, as such decisions are typically not explicitly stated in the publication version of manuscripts. However, among the works performed by authors, one of the applications was dictated by the data availability (Gusarov, Talebijamalabad, and Joly 2020). This is also the typical case of most enterprise supported research projects related to industrial engineering, where the actors already possess a certain amount of internally collected data and require the scientists to provide some solution based on the available inputs. In this case the preliminary analysis can be performed prior to model selection and even, sometimes, prior to the target metrics choice. Obviously, this approach also has several exceptions, as some studies may start directly with the data collection in the first place: (1) with some prior theoretical assumptions and a model to estimate already present, or (2) without any prior assumptions and only a preliminary data collection strategy, which should meet the requirement to provide an answer to the research question. In econometric studies this situation is rather common, when the research starts with a theoretical model construction, which defines the data collection procedure.

Finally, some studies take their roots directly in the described modelling approach, which is rather rare, but not inexistent. It is due to the limited competence of the personnel or a particular interest in methodological approach implementation that such situations occur. For example, some practising researchers may be reluctant to use some modelling strategies which do not match with their research interests. Another case is when the enterprise has already some personnel available, who are familiar with particular statistical toolset, imposing thus some limitation over the research procedure. Typically theoretical econometrics studies also fall into this category. In this case the desired statistical modelling approach inevitably affects both the underlying theoretical assumptions and the data requirements.

2.2.2. Performance evaluation

Constructing a performance evaluation framework involves developing a structured and systematic approach to assess the effectiveness, efficiency, and impact of a system. In the context of this study the *performance* should be redefined as a metric dependent on the entirety of the scientific procedure, and the data analysis stage in particular. It refers to the accuracy, reliability, and consistency of the scientific method in providing meaningful and trustworthy responses to the research problem. We can thus redefine the performance as following:

Performance: precision with which the scientific procedure answers to the economic question.

It should still be pointed out, that the adopted approach to the treatment of a particular research question inevitably affects the performance perception. While such definition may appear as a perfect choice for interdisciplinary model performance comparison, it has some flaws, being the problem and the solution at the time. In particular the complication arrives from the flexibility of the scientific procedure in research. The previously outlined differences in the research procedure construction, the differences in the sequential order of the key elements within such procedure, accentuate this problem. As previously observed, the various applications and use cases inevitably affect the model performance

perception. The introduced *performance* definition accentuates such difficulties by its flexibility, offering no exact answer as to which performance metrics to use. However, its main advantage consists in allowing to perform the performance comparison respecting the eventual specificities of the treated research question.

The selected performance definition should be interwoven into the framework's structure. For the purposes of the framework construction it might be the best choice to offer some sufficiently rigid structure, instead of the all allowing flexibility. This pushes us to consider the different combinations of the available three key elements of the data analysis procedure. Prior to proceeding with the framework construction we are going to provide the discussion of all its different elements in separate sections. While the theoretical part was more or less presented in the Chapter 1, there is still a discussion to be provided on each of missing elements. Because the previous materials includes all the necessary prerequisites for further reading, this section will cut on the level of details for some basic elements. In this subsection we are going to put the focus on the performance metrics available in the literature.

Even though in statistical modelling, when speaking about model performance assessment and comparison, the focus is typically made on the classification accuracy (Andersson, Davidsson, and Lindén 1999; Hand 2012; Askin and Gokalp 2013) this is not always the best option. In model comparison, whatever the research question is, one will always have some target metrics or criteria in mind, but this might be insufficient. While the performance indicators might serve to compare the different research procedures, in case of uncertainty over the underlying processes it might be insufficient to assess the external validity of such procedures. At this stage, for a performance comparison, one should be able to compare not only the models between themselves, but to validate the results against some externally defined target as well.

The performance metrics may roughly be divided into several groups. For example, Japkowicz and Shah (2011) illustrates how the performance comparison of competing models may be assessed over several criteria: (1) quality of data adjustments; (2) predictive capacity; (3) quality of the field specific (ex: economic and behavioural) indicators derived from estimates; and (4) algorithmic efficiency and computational costs. Those performance metrics are defined from the model exploration perspective, but still can be successfully used in the special case of our broader definition of performance. While assessing the capacity of the particular research procedure to answer to a given research question we may immediately identify two different performance classes. On the one side we have the performance metrics related to the target metrics and potentially specified as the data analysis capacity to provide precise and unbiased estimates for the target metrics. On the other side we have all the *external* performance elements, which might be important for the purposes of research conduct, but have no direct relation to the target metrics. For example, the computational costs are rather important element of the performance, but are external to the eventual economic question explored. While the first class complies with our definition of the models' performance, the second one has no direct links with the research question and target metrics, except for a rather specific case of theoretical study.

For the purposes of this work we will primarily focus on the first class of the available performance metrics. Obviously, the metrics should be defined in the context of the target literature of the economics studies, where the classification and discrete choice models are usually implemented for²: (1) transportation research, (2) health economics, (3) environmental studies and (4) consumer studies. In

²More details are available in Section 1.2.1 and Appendix A.

the previously conducted systematic literature review we have identified several topics addressed in *economics* related studies, which could be potentially reduced to: (1) policy making, (2) attitudes assessment, (3) demand modelling and market analysis³.

This series of tasks offers us a rather interesting opportunity to identify the potential target metrics that could be used for performance assessment task. Thus there exist three main groups: (1) metrics based on the direct model outputs, also denoted predictions; (2) metrics which are based on the direct estimates provided by the model, as for example plain effects; and (3) derived metrics, which are obtained through transformation of the direct outputs. Those three groups only roughly outline the available metrics families, as there are many cases where the metrics will transit from one class to another. For example, the classic RUM model estimates used to compute *Willingness to Pay (WTP)* values seem to belong to the third group, while the same model estimated in the preference space and having WTP as the direct outputs transfers the target metrics to the second group.

The last two groups of metrics are actually quite similar, with a wide gap separating them from the first metrics category. The first metrics group is typically used in the prediction oriented studies, where the possibility to correctly classify the inputs plays the crucial role. Those metrics might be used as supporting evidence in other studies, for example in the variable selection a model fine-tuning task. The second and third groups are mostly used in explicative studies, where the accent shifts towards the understanding of the underlying processes.

2.2.2.1. Output based performance metrics

The direct performance metrics are probably the most common type of metrics that comes to mind when speaking about model performance comparison. Those are the metrics mostly used in the context of the statistical studies, ML and DL (Hastie, Tibshirani, and Friedman 2009) in particular, when the need arises for model performance comparison. In the context of the discrete choice modelling and classification those metrics may be divided into several types: (1) metrics based on the discrete outputs, class or choice predictions; (2) metrics based on the probabilistic model output. While the first types of metrics regroups all the values one can compute based on the confusion matrix, the second one includes more complex indicators. For a better review of conventional model performance metrics we suggest the reader to address his attention to the work of Japkowicz and Shah (2011).

2.2.2.2. Direct estimates metrics

The second group of metrics also relies on the immediate estimates of the model. However in this case the focus shifts from the predictions produced by the model, to the *effect estimates* or *weight estimates* depending on the background. It is important to point out that not all of the model classes are equally good at this task as some of the models do not produce sufficient information for usage of such metrics. While in economics the direct effect estimates might be used as an evidence for policy recommendation or further decision making, other statistical learning communities may be less inclined to use such metrics. What is more, depending on the background of the community behind a particular modelling technique implementation, not all the models are devised to provide such type of information.

³Please note that the numbering is added for better readability and does not imply any direct links between the domains and tasks.

To illustrate this point we may turn to the most basic comparison of the classical DCM toolset and basic NN for example. In this case, the classical discrete choice models the researcher to compute the confidence intervals for the estimated effects. Such models are estimated using quasi-Newton algorithms, which uses information on the second order derivatives for convergence purposes.

The calculated Hessian matrix also allows to compute the standard deviations for individual effect estimates in this case. The NNs are too complex in comparison, which makes the quasi-Newton algorithms less applicable in terms of resources consumption, which leads to the absence of user accessible toolset for such analysis. Moreover, for the more complex models the definition of the second order derivatives is sometimes questionable. This makes rather difficult to analyse the confidence intervals of individual weight estimates, which might be still obtained through bootstrapping the estimates for multiple models with matching structure. Still, the complexity of the model might result in non-concluding confidence intervals. This deprives such task of sensibility as there exist multiple combinations of weights that lead to near identical functional form and equilibrium in such models. Finally, speaking about the marginal effects, while it is technically possible to analyse the individual effects through simulation it becomes far less convenient than in usage of classical theory-backed DCM toolset.

2.2.2.3. Derived metrics

The derived or indirect metrics represent a rather complex case, because they might regroup both metrics requiring transformation for the prediction based and estimates based metrics groups. However, on practice it is more wise to include into this group only the transforms of the effect estimates, as most of the metrics based on the predictive capacities of a model already include some sort of transformation or aggregation of the results, making it impossible to distinguish between those two groups in other case.

Probably the most widely used metrics in this case include: (1) the willingness to pay, requiring to observe the relationship between price and particular attribute of the available alternatives; and (2) the non-linear variable effects, requiring specific transformation to obtain the associated estimates in relation to other variables. Both cases are rather popular and may be often encountered in econometric works.

2.2.3. First framework elements

Given the considerations described in the previous part, we can proceed by carefully establishing the first element for the comparison and hypothesis testing framework. As put in evidence previously the performance evaluation as viewed in this work is reliant on the understanding of the research procedures and the data analysis procedures in particular. The three key data analysis elements will lay ground for the performance comparison framework: (1) theoretical models and assumptions, (2) data collection and treatment, and (3) statistical modelling and analysis. Previously we presented those elements as something interchangeable in the context of the rather flexible scientific procedure presentation. The generalisation of scientific procedures requires the understanding of the interdependence of the above elements and the complexity of their introduction into the framework. The following section will present those elements more in detail, prior to combining them into a framework.

Prior to proceeding with the key elements discussion, as well as the framework presentation, it is essential to list the main properties expected from the framework. The understanding of the purpose

and intended use-cases for the framework might influence its structure and affect its construction procedure. Our framework targets the applied researchers who lack familiarity with the DCM and classification toolset in general. We assume that the proficient users are mostly capable to create tool-chains for their own specific needs and are capable to communicate with sufficient clarity in their works. However, the initiates usually lack the visibility of all the eventual pitfalls and complexities of the discrete choice modelling, making it rather difficult to provide reproducible and replicable results. For the purposes of this work we identify two key properties for the framework.

The first key property of such framework is the **clarity**: the procedure should be sufficiently documented to avoid any ambiguity in understanding and errors in implementation. It should be expected that whomever uses the framework for model performance assessment should be able to implement the toolset for the particular use-case. Nowadays, many of the existing studies use different datasets and models, as well as the modelling algorithms and model specification strategies. All this, alongside with distinct use-case scenarios, performance metrics and research objectives.

The **reproducibility** is another key property of such framework: it is expected that models *perform identically* under *identical circumstances*. Quite a lot of most advanced model rely on simulation or have some random components in their estimation procedures. Moreover, many of models, techniques and algorithms incorporate some *hyperparameters*, which should be defined by the end user. Such elements are often overlooked in typical performance comparison benchmarks.

Both those elements are important in the context of the framework construction, as its intended use is oriented towards the applied scientists lacking proficiency in the particular discipline. We target both the accessibility and ease in understanding of the framework's components and elements, alongside with the simplicity in knowledge transmission in the works constructed with framework implementation.

Thus for the sake of clarity and reproducibility we simplify the framework's structure making it more rigid. While in reality, following the testimonies of the data analysts, the research procedure rarely starts with the theoretical assumptions we prefer to follow the academic vision of the research sequence (Figure 2.6). Such results of interviews may actually be explained by the existence of the endogeneity in the choice of research topics and consent to analyse the applied cases in the research community. Many researchers prefer to participate in the studies where they already have an extensive theoretical knowledge, while the industrial representatives tend to select the research institutions in accordance with already published works which resemble the most to the desired analysis procedure. Thus putting the theoretical assumptions and limitation on the first step makes rather simple to order the two other elements. As in the most exemplar academic works the theory is followed with the data collection or acquisition step and is completed by the statistical analysis stage.

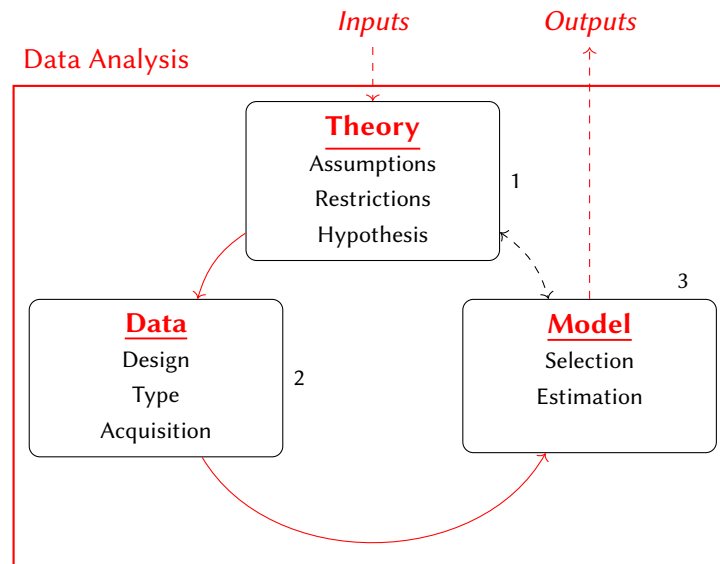


Figure 2.6.: Data analysis

The efficiency and reliability of any scientific procedure relies on all of the presented above elements. Those are equally the elements ensuring the research replicability (Hillel et al. 2021) in most cases: theory, data and modelling process. The theoretical foundation forms the basis upon which scientific procedure relies in the answering to the research question. This encompasses theoretical assumptions, outlining the underlying principles and concepts, as well as hypotheses and restrictions that define the scope and limits of the study. The data related part involves insight into the experimental design, the nature of the variables involved, and the strategy used for data acquisition. The modelling phase translates theory and data into interpretable insights. This involves critical decisions regarding model selection and estimation techniques, the software and algorithms.

Fortunately we have already presented the most of theoretical assumptions and background in the Chapter 1.2.2 of this work. This allows us to proceed directly with the issues of data acquisition in the context of discrete choice modelling studies (Section 2.3), followed by a more in-depth discussion of statistical modelling issues (Section 2.4). Thus all of the elements appearing in the data analysis part of the scientific procedure are covered in this thesis at on point or another.

2.3. Data constraints and simulation

Following the intended framework structure we focus first of all on the data acquisition issues that may arise in both applied and theoretical studies. The second stage in the framework should be left for the dataset choice or dataset generation procedure. The data plays an extremely important role in any type of statistical analysis. There exist requirements to the quality of data, the number of observations, as well as the compatibility with the desired theoretical implications imposed by the adopted theoretical framework. For the purposes of this work we limit ourselves to the data collection and treatment issues that may arise in the choice analysis and classification studies.

There exist two different sources of data that could be encountered in the existing studies: (1) real world data and (2) simulated data. In the majority of cases the scientific articles address some real world problematic using some observed data (*real world data*). The task in this case is to uncover

and interpret the processes behind this data: *causal relationships* and *data generation function*. In other cases, which usually appear in theoretical works, the data is *simulated*, artificially generated under known constraints. This allows to verify the developed theories and modelling techniques, while having full control over the data generation procedure and environment. Finally, there exist various mixes of those two data types. Starting from the simulation used for anonymisation of the data points, with an aim at altering as less as possible the data structure. To the imposition of simulated behaviour over a real world collected data points for population.

Not only the dataset generation strategies may differ, but there also exist many different constraints and limitations over the data in relation to the theoretical assumptions, both from the methodological and statistical perspective. In many cases the research is constructed primarily around the available data, which limits the research questions available for exploration. There exist many various problems related to the data collection and usage. Those may be divided into: (1) *theoretical biases* induced through the inconsistency between the chosen theoretical assumptions and actual human behaviour in the *Choice Experiment (CE)* context; and (2) *statistical biases* associated with experimental design construction and inappropriate modelling strategy choice.

Within this subsection we are going to address both of the above issues in the context of their integration and representation within the performance comparison framework. We are going to address the eventual analysis problems related to data usage, as well as how to use simulation to explore performance differences in a fully controlled and reproducible environment.

2.3.1. Data acquisition

Both *real world* and *simulated* datasets have their own advantages, and both might be criticised depending on their application and use-cases. In this subsection we are going to focus on the data acquisition strategies for the purposes of both applied and theoretical discrete choice modelling studies.

The first data type is issued from the real world and is assumed to bring in itself some valuable information about the world functioning. For example, it may bring some insights about the behavioural patterns within population in certain context. One of the key drawbacks, is that it usually lacks sufficient anchoring for research procedure performance analysis to be conducted. This is typical problem for classic econometric studies, where the target metrics are typically represented by the direct effect estimates, which real values remain unknown in the case of real world dataset. At the same time, this does not affect the studies focusing on the predictive qualities of their models and which seeks to achieve the best adjustment to the data in their research procedure. Yet another disadvantage of this data type are the associated risks and biases. There are scarcely any datasets which are exempt of any measurement errors, missing data. In the context of the behaviour studies this problematic becomes even more accentuated, as human behaviour may be affected by negligible elements, such as the survey structure or the colours used in the surveys' graphical interface. This situation greatly impacts the results validity obtained from such data.

The simulated data in its classical format is exempt of some of the biases, as the behavioural model is imposed on the population by researchers, granting them the full control of the underlying processes. However, its argued in the literature that the evidence on the model performances obtained from such data is negligible, as the data is free from all the real world biases. What is more, it is typically assumed that the imposed behavioural model may not correctly reflect the real world situation, thus depriving

the obtained results from any external validity.

Both of the data collection strategies are briefly summarised in Figure 2.7.

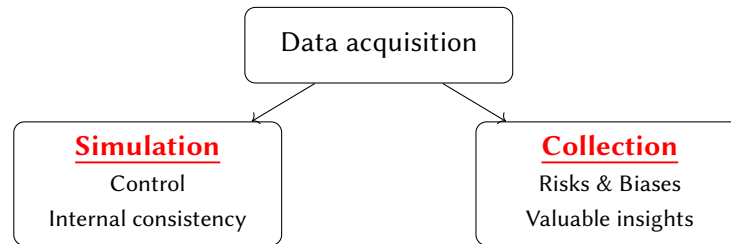


Figure 2.7.: Data acquisition strategies

This data type selection presents researchers with a challenging trade-off between the advantages of maintaining strict control over simulated data and the potential biases from the data collection procedure. Another trade-off concerns the presence of internal consistencies in simulated scenarios, against potential external validity and valuable insights in real-world data. Those trade-offs present a particular challenge for scientists, pushing them to strategically balancing the advantages and limitations of each data source to ensure the match between the data type and the research problematic.

2.3.1.1. Data simulation

For the purposes of theoretical studies the most popular strategy is to artificially generate the required dataset. The simulation in the context of the discrete choice modelling is performed either using simple Monte-Carlo based Simulation (*MCS or simply MC*) (Raychaudhuri 2008; Rubinstein and Kroese 2016) or some more complex simulation approach, such as *Agent Based Simulation (ABS)* (Chan, Son, and Macal 2010).

The MCS is the most widely spread type of simulation procedure which is broadly present in econometric studies. As encountered in the literature, the definition of Monte-Carlo Simulation in general is not focused solely on the simulation, although the technique is widely used for simulation as well.

Monte-Carlo Simulation (Raychaudhuri 2008) is a type of simulation that relies on repeated random sampling and statistical analysis to compute the results

In econometric studies this method typically implies the dataset reconstruction given the input distributions, which are usually taken *as-is* based on some observed real world dataset, or chosen based on the research needs. For example Rose and Bliemer (2013) use arbitrary chosen values for their methodological work. This approach offered them the possibility to provide an extensive analysis of the model performances, depending on the varying factors such as population size and/or number of choice sets observed per individual. In some cases the MCS is reinforced in combination with *Boosting* algorithms for dataset resampling. Such analysis is not very demanding in terms of computation resources and can be easily performed without particular complications in code implementation and adaptation.

The ABS toolset extends the key idea of MCS offering much more flexibility and representing in itself an entire simulation framework. Among the definitions recurrent in the literature we encounter the one of Macal and North (2014).

Agent-Based Simulation (Macal and North 2014) ... (is a) system(s) comprised of individual, autonomous, interacting agents ...

ABS may be understood as a simulation procedure governed by the exact simulation of multiple agents' behaviour. The notion of the *agent* assumes fractioning the population, the unit used in the classic MCS, to the individual level. As every individual becomes a separate agent new opportunities for simulation of complex systems are revealed. It becomes possible to introduce individual specific behavioural patterns within population, creating much more complex heterogeneous artificial populations. It also enables the introduction of interactions between individuals and/or the environment, which makes the choice simulation much more realistic.

However, everything comes with a cost. The ABS simulation is often more demanding in the computational resources and the associated code-base management becomes more complex. The informatics implementation of ABS usually implies usage of *Object Oriented Programming (OOP)* paradigm, which makes the interactions with code-base more structured, but at the same time increasing the code-base volume and the burden for the new users. An implementation of an ABS toolset for choice behaviour simulation created during the PhD is proposed in Appendix E (Gusarov 2022).

Either way the simulation grants the researcher the full control over the data-generative process. It typically includes all the elements that usual dataset would include (Figure 2.8) to achieve the better resemblance with the real world.

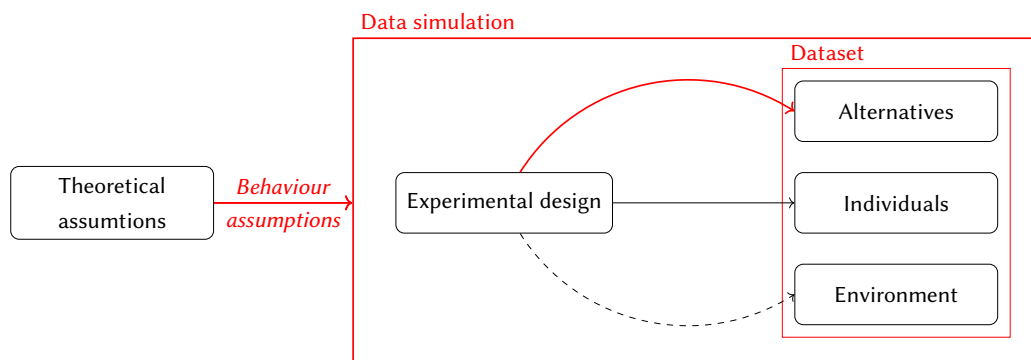


Figure 2.8.: Data simulation

Simulated datasets are often used to validate various theories (Garrow, Bodea, and Lee 2010) or assess the capabilities offered to the researcher (Marcela A. Munizaga and Alvarez-Daziano 2005). Synthetic datasets are equally often used for usability and statistical efficiency testing (Caron et al. 2021). Another objective of such test is the study of the requirements towards dataset sample size (Alwosheel, van Cranenburgh, and Chorus 2018). The effects of alternatives' sampling are equally explored (Nerella and Bhat 2004). There exist even studies focusing on the exploration of statistical properties and procedural implications in synthetic dataset usage (Bodea and Garrow 2006).

However, the usage of simulated data is subject to a lot of criticism. In fact, the researchers relying on the simulated data favour the theoretical models serving to generate the data, which also impacts the results of the following model performance assessment and comparison. For example, the usage of RUM framework for dataset generation will: (1) inevitably favour the RUM compliant models in the performance assessment stage; and (2) reflect potentially non-representative behaviour in comparison with the real world, thus invalidating the research's external validity. Both of those issues have some

solutions, but they may appear as suboptimal.

The most common solution which addresses both issues figures in the work of M. Bierlaire, Bolduc, and McFadden (2008). The idea is to use the real world dataset as a support for simulation. This approach ensures that the individual characteristics within population, as well as the alternative attributes are sufficiently close to reality to remain representative. At the same time the theoretical behaviour restrictions are imposed over the *real world* population, which allows to perform statistical tests of the estimates validity and provide discussion on the model performances.

2.3.1.2. Data collection

The real world data is omnipresent in the applied research, as the key objective in this case is to uncover the underlying model under reasonable theoretical assumptions. Evidently, this approach limits the researcher in the performance assessment task, as there is no way to reference the data generation process in the performance assessment. But combined with analytical model exploration it may not be that limiting for the research. What is more, such approach is assumed to bring more external validity to the results, because of the similarities with other collected datasets. Such decision may be reasonable if we were to compare the model performances in terms of prediction, for example. The comparison of predictive power has more external validity in this case, compared to the simulation approach (Japkowicz and Shah 2011). Such strategy uses real world data, which approaches the modelled situation to the reality. Some researchers imply that this assures at least some degree of external validity, in contrast to simulated data.

However, the data collection task is typically a far more complex task than simulation. First of all, the real world data in the context of behavioural studies and choice analysis may be divided into several categories in function of the choice types that the researcher can observe. The key division is observed between *Revealed Preferences (RP)* data and *Stated Preferences (SP)* data.

The first data type assumes that individual's real choice is observed in the natural or close to natural environment. The train ticket usage information, the shop billing information and many more other similar data sources provide this type of information. Typically for this type of data the information about all the alternatives considered by the individual is not available, which is the key source of criticism for this data exclusive usage. In most cases the information available indicates on the choice or non-choice of a limited subset of analysed alternatives, without any particular control of other available alternatives. In experimental economics this issue is tackled through controlled environment. However, without sufficient effort put into the environment creation the data quality falls. The obtained data in this case approaches by its nature the SP data, as the individuals' choices still bear declarative nature.

The SP data assumes that individual provides some declarative information on the choices made in hypothetical situation. Among the advantages in this case there is the possibility to frame the hypothetical situation to suit the needs of the researcher to the best. However at the same time the individual may not always respond truthfully, or can correctly transcribe the hypothetical situation oral or written description to the real world counterpart. Such biases equally affect the RP studies performed within a controlled environment, but could be evaded in the case of plain observation of the individual behaviour in a uncontrolled setting. This leads to potentially biased data, at the same time allowing to explore the novel alternatives adoption in fictive context.

There also exist some combination of the two data types. In both cases the combined usage of different data types targets the mitigation the individual biases observed in each of the individual cases. The first solution is to combine both types of data in collection RP and SP data for joint estimation (M. Bierlaire, Axhausen, and Abay 2001). This ensures that most effects are realistically estimated over the RP data part, while the SP counterpart allows to introduce non-existent alternatives to the choice set or analyse the changes in behaviour subject to hypothetical situation. The second option is to use the RP data as a support for SP data collection, approaching the hypothetical situation to the real world experience of the individual. This is assumed to increase the external validity of the resulting SP observations. All those data types are summarised in Table 2.2.

Table 2.2.: Data types in choice analysis studies

| Data type | Description |
|-----------|--|
| RP | The subjects provide information about the choices they have made |
| SP | The subjects face some hypothetical situations |
| RP & SP | The subjects provide information about their previous choices and face some hypothetical choice situations |
| RP to SP | The subject provide information about their previous choices, then face a hypothetical situation derived from their previous answers |

2.3.2. Experimental design and sources of bias

This subsection will present the different biases that may affect the data quality and, by consequence, the obtained results (Haghani et al. 2021a, 2021b). Some of the elements presented here will affect solely the data collection step, while other will bear identical importance for both simulated and real world data sources. These biases may originate from a multitude of sources: the inherent limitations of measurement tools (Jang, Rasouli, and Timmermans 2017), the misconceptions in survey or experimental design (Malone and Lusk 2018), or the errors committed in sampling from population (Nerella and Bhat 2004). The biases may be treated differently in dependence of their source, and the research strategy. Nevertheless most of them arise, one way or another, as a result of errors and misconceptions admitted in the data acquisition step.

At this point it is crucial to introduce the notion of *Experimental Design (ED)* to this discussion. This notion may be applied to the most data acquisition strategies and not exclusively to the ones directly related to experimentation, also denoted *Discrete Choice Experiment (DCE)*. While it may seem that DCE implies rather strict SP oriented data acquisition strategy, it is not really so, as the DCE simply assumes targeted data collection in an experimental setting. Thus even the data collection through supermarket billing system may be treated as DCE, provided the data collection strategy was designed to suit particular experimentation needs.

Experimental Design (Kreutz and Timmer 2009) or Design of Experiments (DoE) refers to the process of planning the experiments in a way that allows for an efficient statistical inference. A proper experimental design enables a maximum informative analysis of the experimental data, whereas an improper design cannot be compensated by sophisticated analysis methods.

There exist different approaches to the construction of the experimental designs (Blades, Schaalje, and

Christensen 2015; Reed Johnson et al. 2013) for applied studies: (1) non-optimised ED, and (2) optimised or efficient ED. Non-optimized EDs are characterized by a more straightforward approach, often adhering to conventional methods without emphasis on resource usage optimisation. The optimised EDs are carefully crafted to make the most efficient use of resources, whether it be time, participants, or budget. This second category imposes additional hypothetical assumptions in the data collection step of the procedure. This increases efficiency, but also the complexity and eventually multiplying the potential sources of bias (J. L. Walker et al. 2018).

The first category is rather scarce and it attempts to maximise the quality of collected data without any particular prior assumptions over the data structure, nor setting any priorities over the studied explicative variables and effects. The most popular design in this case is *Full Factorial (FF)* (Bose 1947), or *Randomised Full Factorial (R-FF)* design, popular in health related studies (Bur et al. 2022), for mixed cases where both continuous and categorical variables are present. This design allows to estimate all the possible effects or cross effects that may potentially exist in the collected data. However, among the key disadvantages of this design researchers identify the extreme cognitive burden for the subjects, as the number of choice situations growth nearly exponentially fast with the increase in number of attributes to consider or the attribute levels. Among the solutions to this dimensionality problem one may often encounter the attempts to randomly distribute the FF design elements over the different individuals within population.

Yet another popular solution is the generation of the *Fractional Factorial* (sometimes also denoted as *Partial Factorial*) designs (Louviere and Timmermans 1990), where only the desired effects of interest might be identified. This approach restricts the flexibility for future data analysis and increases the risks associated with the data collection step were something to go wrong.

Finally, the most advanced techniques gaining popularity in recent literature are the *Efficient Experimental Designs* (Scarpa and Rose 2008; Rose et al. 2008; Kuhfeld, Tobias, and Garratt 1994). The ED of this type attempt to maximise the variation within the dataset to be collected, which allows to identify the desired effects in reduced samples. Nevertheless, such solutions usually require prior assumptions on the behavioural model to be identified, thus inevitably favouring the same model structure during the identification stage. The priors are typically extracted from the previous similar research results or obtained through preliminary data collection. In the latter case a small sample is collected using the FF or other similar design, the preliminary statistical model is estimated over this data and the results are used for efficient design construction. The *Bayesian Efficient Designs* are sometimes separated as a standalone category, although the underlying principles in their construction are relatively close to the general strategy in *Efficient Designs* construction (Kessels et al. 2011).

The ED generation usually requires the researcher to identify the following elements:

- The attributes nature and structure (ex: variable scale/levels for each attribute)
- Number of attributes per alternative
- Number of alternatives per choice set/situation
- Number of choice situations

While the above elements are indispensable for the ED and may be encountered in most papers performing data collection through a DCE, there also exist other elements that are not always attributed to ED configuration, but could be viewed as such:

- Number of individuals to observe
- Number of observations (observed choices)
- Individual characteristics to collect

Finally the elements external to the final dataset dimensions or variables composition, but still affecting the data quality and eventually obtained results. Among them we may list:

- Data collection means (ex: survey administration method)
- Data collection setting and format
- Population, from which the data sample is drawn

Each of those elements influences the data quality and reliability. In the literature it is possible to find the methodological papers addressing each of the listed above elements in attempt to provide guidance in research design construction and data collection task simplification.

The interactions between the ED, data acquisition and theoretical assumptions may be represented as in Figure 2.9. The ED is the mediator mapping the theoretical requirements of the research procedure to the data acquisition step, while relaying back the eventual compatibility restrictions.

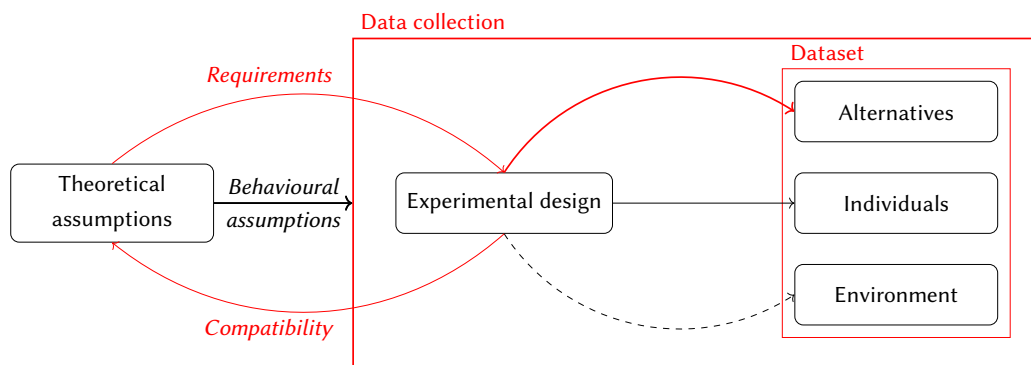


Figure 2.9.: Data collection

2.4. Models and their capabilities

The next step in the procedure concerns the models. By the term *models* the ensemble of statistical analysis models available for the researcher in the context of discrete choice analysis is understood. As previously pointed out, in economics, consumer choices data are mainly studied through classification tools from machine learning techniques or regression tools like discrete choice models from econometric techniques. These two practices in particular illustrate two distinct approaches to applying statistical learning. As described by Breiman et al. (2001) and later by Athey and Imbens (2019): the ML focus on the predictive qualities and Econometrics attempts to decipher the underlying properties of the data. Engineering sciences and Computer sciences focuses mainly on ML techniques, whereas in Economics and other applied Social sciences, the scientific community prefers to implement the traditional econometrics techniques to explore hidden patterns (Athey 2018). The understanding of appropriate model families and their place in the framework's context are shown in this part.

DCM, and individual choice modelling in particular, may be a topic sufficiently narrow to have the control and possibility to present the different underlying theories. At the same time remaining sufficiently large to contrast the different approaches to different research questions in terms of theoretical

assumptions and modelling procedures. This decision equally limits the number of the estimation and statistical modelling approaches we can introduce and exploit. This limitation arises due to the fact that not all statistical modelling techniques are compatible with underlying theoretical assumptions. For example, the random forests may be used to solve some of the questions in the domain of application, but they rarely comply with the theory.

In this section we are going to explore all the elements one may encounter within the scope of the statistical model or closely related to it. The statistical modelling task involves some operations over data, including resampling and cross-validation methods among other. It also includes the complex array of methods, algorithms and informatics implementation of a given statistical analysis toolset. Here we are going to separate those different element composing the model and offer a comprehensive perspective on their interactions for a better understanding of the complexity of performance comparison task in the context of discrete data analysis.

First of all, the statistical analysis step in the traditional academic vision of the data analysis procedure follows both theoretical analysis and data acquisition stages of the research work. Thus, the statistical model selection process is assumed to be impacted by those two prior elements (Figure 2.10). While it is not always true in the real world as the researchers may have their own reasons to switch the order of elements within their research procedures, for the purposes of our framework we adopt the sequence which will potentially the best suit the novices. What is more, this sequence does not forbid to make the returns along the frameworks paths.

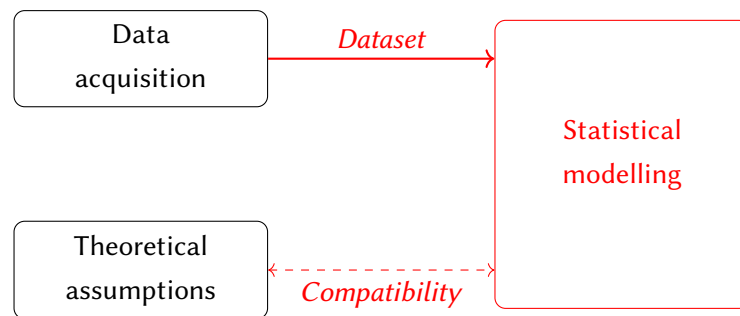


Figure 2.10.: Statistical modelling

The data acquisition procedure returns the dataset which should be used for statistical modelling and may potentially require some adjustments and prior treatment, such as class balancing, resampling or partitioning for cross-validation procedures. Except for those transformations, a practice which is far more predominant in ML community than among econometricians is to partition the dataset into 2 or 3 samples depending on the particular research procedure needs. This partitioning assumes dataset separation into learning (training) and validation sets, to which, depending on the estimation or learning algorithm used, may be added the testing sample.

The theoretical assumptions influence the statistical model selection process, as it was illustrated in the Section 1.2.2. One of the simplest examples includes the implementation of the classic RUM framework based data analysis, where the RUM theory dictates the possible model structure, which should inevitably include the stage with softmax transformation at the output using deterministic utility latent constructs as inputs. Such structural restriction to achieve the RUM-compliance drastically reduce the number of available models nearly barring out all the non-softmax classifiers (ex: decision trees), or any meta-models relying on boosting methods (ex: decision forests). Even the softmax re-

liant models might not completely respect all the conditions of RUM-compliance and might require some adjustments in their architecture. While softmax transformation is present in a predominant part of RUM-compliant models it does not guaranty the RUM-compliance of a model. And inversely, non-softmax models might potentially satisfy the conditions of RUM-compliance, although such case might be extremely rare.

In combining those elements into *statistical modelling* step we add some more elements to this concept in accordance with the modular model concept introduced in the Chapter 1, while introducing the taxonomisation task complexity. As precised in the previous section, the statistical modelling may be fractioned in multiple elements among which: (1) data transformations playing crucial role in model estimation; (2) statistical model itself; (3) estimation algorithm; and (4) the algorithm's software implementation. Those are 4 major elements on which we may divide the statistical modelling stage of data analysis. While those steps are tightly interconnected it is rather difficult to introduce them in a fixed order into the frameworks' integrity, although we attempt to follow the same logic as with previous elements in following the most academic approach to data analysis.

As presented in Figure 2.11 we assume that the key drivers of the statistical modelling are the desired statistical model alongside with the eventual data transforms of the original dataset. For a simple example the first element might be given by a simple Logistic Regression, while the second one, data transformation, may involve a separation of the training and testing subsamples from the original dataset. Those two elements are tightly interconnected, as the usage of particular data transformation technique may have direct repercussions over the model nature, as happens in the case of *Boosting* technique implementation.

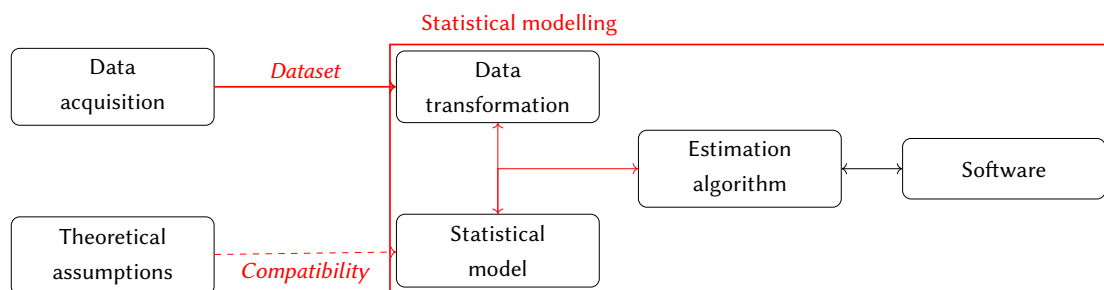


Figure 2.11.: Statistical modelling in detail

The combination of both those elements more or less dictates the following choices. The resulting model class and complexity typically dictate the estimation procedure to be implemented, alongside with the transferred requirements in terms of the required model outputs to satisfy the target metrics prerequisites. At the same time, the software choice is determined nearly simultaneously. While for the most simple models the software choice is rather vast, more specific cases have scarce choice of software with correct technique implementation. For example, while the MNL model has many implementations, the much more advanced and flexible HCM model are available only in a few packages (biogeme in case of Python and apo11o in R). There exist other emerging packages that offer rather targeted solutions for such complex models orienting their efforts on speed and efficiency (Arteaga et al. 2022), but it is important to consider the popularity of the software within scientific community. At this point we may switch to the discussion of each of those four elements separately.

2.4.1. Statistical models

The most prominent families of the statistical models were already introduced in the first chapter. Yet, we should shed some light onto the statistical model interpretation in the context of the scientific data analysis procedure. For the purposes of this section and the following stages of this work we regroup under *statistical model* term the ensemble of modelling techniques having at their heart some statistical toolset. The reference work with nearly identical perception of statistical models or rather *statistical learning* is given in Hastie, Tibshirani, and Friedman (2009).

First of all, as it was previously specified, the statistical model perception differs from one discipline to another, as the models undergo field-specific changes. The model structure and purpose changes in accordance with the particular field specific task and both model interpretation and usage may drastically differ across disciplines. Even such simple models as basic Logit models widely implemented in choice modelling undergo modifications in the marketing and preference learning studies.

Previously we have introduced the idea of modular database-like approach to the model taxonomisation. As it can be understood from the Figure 2.11 we separate several of the elements previously presented as a part of model properties. In particular we set apart the *estimation algorithm*, *software implementation* and *data transformation*. The previously identified elements of the modelling procedure are fractioned in this part across the four dimensions. In particular the ones remaining as a part of *statistical model* part are:

- Functional form of the model. For example, the linearity or non-linearity in parameters or non-parametric model type.
- Transformation functions applied to the individual elements during the estimation. Here we speak about the transformation functions accounted for during the estimation procedure as in GAMs or the transition functions inside the hidden layer of NNs and DNNs.
- Kernel transformation, also denominated as functional transform in econometrics. Regardless of the terminology differences, this element includes information on the transformation of the model. For example, in the case of Logit model it is the Logistic transformation of the linear functional part. Another example will be the more general GLM implementation, where the linear in parameters output is fed to a transformation function prior to the comparison with the observed output.
- Error term specification. Seemingly identically structured models may have different error terms, resulting in the different results at the estimation step. A work of Bouscasse, Joly, and Peyhardi (2019) offers an overview of reference models performance difference depending on cumulative distribution function choice.
- LOSS function. Even though the LOSS function is a part of the model to be optimised during the estimation procedure, we consider it to be a part of the statistical model. It includes information required for further optimisation during estimation stage, but is tightly tied to the model structure.
- Regularisation or penalty transformation. The penalisation argument is specified as a supplementary part of the LOSS function, which ensures that both of those elements are grouped together.

2.4.2. Data transformation

The data transformation stage is a rather complex concept to be presented as a part of the statistical modelling stage of the data analysis procedure. For inexperienced reader it may be difficult to understand the reasons behind the separation of the data transformation stage as a part of the statistical modelling part of the procedure, as there already exists a data acquisition step in the data analysis procedure. However, there are strong reasons to do so. Some statistical models, estimation algorithms or even software may require specific transformations to function correctly. However, those transformations are potentially negligible during the data collection step.

For example, in the case of software differences, we may offer an example from the R language background for estimation of the MNL model types. A rather simple and widely used by novices package `mlogit` requires the data to be presented in a specific *long* format, where each line within the dataset corresponds to a single match between individual, choice set and alternative. In contrast, other software, such as `apollo` or `biogeme` (the later being a Python package) is designed specifically for usage with *wide* data format. This alternative format assumes that each line contains full information on a single choice situation, corresponding to a single match between individual and choice set. In this case each line contains all the information about the attributes of several alternatives at once.

Another example concerns the transformation of the input variables and their type. In the applied field studies the researchers, according to interview results, tend to collect as much data as possible due to the temporal and financial limitations. It is typically much easier to collect more data in a single attempt, than conducting several unrelated studies to collect supplementary or concurrent data. In this case, for modelling purposes the dataset may be simplified: some variables may be omitted and some may undergo transformations including simplification. The later may include creation of classes from the continuous variables or simplification of the multiple choice problems to a less sophisticated binary choice cases. What is more, typically during the data acquisition stage the researchers rarely have full visibility of the resulting dataset statistical properties to correctly account for them. Some of the classes within population may be under- or overrepresented within the obtained sample. Other undesired properties may be observed following the data acquisition stage, which might be potentially corrected through data transformation and adaptation.

Consequently the *data transformation* step of the statistical modelling procedure includes the following elements:

- Modelled variable data type. Here we speak about the output variable which is used as optimisation target by the statistical model. This criteria appears in quite a number of other taxonomies as it inevitably affects the possible models to be implemented over the resulting data (Agresti 2013).
- Input variables data types. Equally important for the purposes of statistical model selection the input variables properties may drastically affect the choice of the analysis methods. Some of the models misbehave in presence of particular variable types (ex: unbalanced binary variables) or require a particular treatment of various variables.
- Input variables transformations prior to estimation. A consequence of the previous point. It may include normalisation of the continuous variables before estimation of the NNs or transformations to ensure the positive or negative values are fed into the model.
- Dimensionality reduction transformation and data adaptation. For example, the Principal Com-

ponent Analysis (PCA) dimension reduction where only several first components are considered in the later steps. The Auto Encoding (AE) techniques may have identical repercussions over the results.

- Sampling and resampling strategies. This point involves both subsampling for the purposes of statistical property exploration and learning algorithm implementation. The first includes for example, bootstrapping of the MMNL or HCM models in order to derive statistical properties of the estimates⁴. The second element assumes the partitioning of the dataset into learning and testing and/or validation subsamples.
- Boosting implementation. Although quite similar to simple bootstrapping, this technique is better understood as a generation of multiple subsamples for estimation of potentially non-identical models over them, in contrast to simple bootstrapping or model bootstrapping.

2.4.3. Algorithms

The estimation stage is nonetheless important for the purposes of statistical modelling procedure part. The statistical modelling results may be greatly impacted by the chosen estimation or *learning* algorithm. While in the simple cases the algorithm selection does not have such great impacts on the obtained results, for example, a fully identified Logistic Regression will produce more or less stable results regardless of the implemented estimation algorithm and even the *learning algorithms* optimised for big data may potentially yield reliable results in this case. The estimates for more complex models may drastically vary not only depending on the algorithm choice, but also depending on the algorithm initialisation and hyperparameter selection.

The comprehension of modelling algorithms might be extended outside the boundaries of simple model estimation (Ortelli et al. 2021). Usually variable and model selection play equally important role in the procedure. And while in many econometric applications those are selected based on expert knowledge and reliance on previous studies, the algorithmic approach may be implemented for this part of research procedure. In the case of ML algorithms, in particular NNs, the flexibility in functional form identifications ensures the feature selection at a certain step. Transcription of those elements into the classic DCM applications equivalents to the iterative model selection approach (Lancsar, Fiebig, and Hole 2017). Those elements might be summarised as:

- Variable selection techniques. Those include some tightly intertwined with the regularisation techniques and dimensions reduction techniques presented previously as parts of *statistical model* and *data transformation* elements. The distinction between the three elements resides in the actual target use case of the applied technique. For example, a regularisation over the loss function may be used *as is*, or it might be used as a preliminary stage for variable selection in which case the model is re-estimated with a reduced support of input variables.
- Model selection techniques. Those include the various meta-algorithms and procedures for model selection based on their information criteria and relative predictive performance.
- Estimation algorithm. The optimisation and search for parameter estimates often may be performed in various ways. This part includes the key estimation algorithm elements, for example the choice of a quasi-Newton optimisation algorithm against the stochastic gradient descent.

⁴Here we present the case of *model bootstrapping*, although the more simple bootstrapping techniques may also be considered.

As further opening for this discussion, one may consider the combinatory methods. The ensemble techniques like bagging and boosting, aggregate the predictive power of multiple models to improve overall performance. By combining diverse models, these methods mitigate the limitations of individual models and contribute to robust predictions. However, as this approach renders the results less interpretable they are rarely implemented in the context of DCM studies. In this thesis those methods will not be explored.

2.4.4. Software choice

The last element remaining from the presented taxonomy concerns the estimation algorithm implementation. Each of the estimation algorithms, when transcribed into machine language, may yield different results depending on the software implementation and hardware configuration requirements. There exist a multitude of software solutions for choice modelling purposes. They vary across different criteria and possess different properties, starting with the ease of usage and user-friendliness and ending with the efficiency and resource requirements.

In this subsection we are going to present the different available software for choice modelling, starting with the most powerful and conventionally used for choice modelling tool-sets, with which the authors have familiarity, and ending with some emerging software solutions. The presentation of those solutions will be organised by software⁵.

2.4.4.1. R language

R is a programming language utilized for statistical computing and graphics. Developed and supported by the R Core Team and the R Foundation for Statistical Computing under a GNU General Public License (GPL) license. It serves as a free software environment designed specifically for tasks related to statistical analysis and data visualization. R is compatible with a diverse range of operating systems, including GNU Linux platforms, Windows, and MacOS.

In R software repositories CRAN there are many packages suitable for classification tasks or choice modelling in particular. However, probably the most popular and tailored specifically for DCM tasks are: (1) `apollo` and (2) `mlogit` packages.

The `apollo` package is a software toolkit designed to facilitate the estimation and application of choice models within the R programming language. It is among the most feature rich tools existing at the time. It offers a versatile range of tools for users to create custom model functions or utilize existing ones. With `apollo`, users can incorporate random heterogeneity, both continuous and discrete, at the individual and observational levels across various model types, including standalone and hybrid structures. The package supports both classical and Bayesian estimation methods and covers a wide array of models, including discrete choice and discrete continuous models. It also provides multi-threading capabilities for faster estimation processes and offers a plethora of pre and post-estimation functions, including the computation of individual-level posterior distributions, enhancing the efficiency and flexibility of choice modelling workflows.

A much more simple and easy to use alternative is represented by `mlogit` package. This tool is designed for conducting maximum likelihood estimation of random utility discrete choice models. It enhances

⁵Here we speak about the programming language with which the individual libraries are implemented.

the modelling process with its intuitive model description interface, which utilizes an enriched formula-data structure. `mlogit` offers a highly versatile estimation function and a robust testing infrastructure specifically designed for handling random utility models, making it a valuable resource for researchers and analysts working with discrete choice modelling in R.

2.4.4.2. Python language

Python is a high-level, interpreted programming language known for its simplicity and readability. Python was designed to be easy to understand and write, with a clean and concise syntax that emphasizes code readability. It is a versatile language that can be used for a wide range of applications, from web development and data analysis to artificial intelligence and scientific computing. Python is known for its large standard library, which provides modules and packages for various tasks, making it a popular choice among developers for its productivity and ease of use. Contrary to R, Python is developed under *Python Software Foundation (PSF)* license, an *Open Source Initiative (OSI)* approved open source license. This permissive license makes it freely usable and distributable, even for commercial use, as it does not enforce the requirement that any derivative work must also be open source. While the R's GPL enforces the openness of the codebase, the Python's license does not enforce such requirements on the commercial actors.

The `Biogeme` package is the closest analog to `apollo` within Python ecosystem. It provides tools for estimating the parameters of discrete choice models, such as multinomial logit, nested logit, mixed logit, and more. While `Biogeme` itself is not written in Python, it has a Python interface that allows users to work with the software using Python scripts for tasks like data preparation, model estimation, and result analysis. It is often used by researchers and practitioners to analyse and model choice behaviour in various contexts.

Among alternative packages created specifically for choice modelling, we can mention `xlogit` package. It is an open-source Python package designed for a GPU-accelerated estimation of Mixed Logit models.

2.4.4.3. SAS software

Statistical Analysis System (SAS) is a software suite widely used for advanced analytics, statistical modelling, and data management. For discrete choice analysis, SAS offers a variety of tools to explore choices made by individuals among a set of discrete alternatives. The specific procedures and features are present within its statistical suite, `SAS/STAT`, for analysing discrete choice models. This includes capabilities for estimating choice models, predicting probabilities, and conducting various tests related to the model. However, in classic publication on DCM the SAS software is rarely cited in relation to economics and transportation applications.

2.4.4.4. Stata software

Stata is yet another popular software used for statistical analysis and data management. It gained its popularity among economists grace to an immense number of numerical implementations of the advanced statistical functions. It provides a wide range of tools and features for researchers, statisticians, and data analysts to perform various tasks related to data analysis, data manipulation, and statistical modelling. Stata is commonly used in academic research, social sciences, economics, epidemiology,

and other fields where data analysis is essential. Some of its key features include data visualization, regression analysis, time series analysis, survey data analysis, and support for custom programming and scripting. Stata has both a command-line interface and a graphical user interface (GUI), making it accessible to users with different levels of programming expertise.

2.4.4.5. Julia language

The final element in the list is Julia. It is a high-level, high-performance programming language primarily designed for technical and scientific computing. It was created to address the need for a language that combines the ease of use and productivity of languages like Python and MATLAB with the performance of lower-level languages like C and Fortran.

Recently Julia has gained popularity in fields where computational performance is critical, such as scientific research, data analysis, and machine learning. It provides a powerful and versatile environment for developing high-performance applications while maintaining a user-friendly and expressive syntax. However, for now there are no powerful implementation designed for choice modelling in Julia yet. Among the available options we may cite `DiscreteChoiceModels`, which still lacks functionality and should probably be compared to `mlogit` package in R.

2.5. Framework presentation

Once all the previously discussed elements come together, the framework's structure begins to take shape. This section involves assembling all the components to create a unified framework, allowing to compare and contrast the performance of different scientific approaches, be that changes in models, theoretical assumptions or the data analysis procedure as a whole. At this point in the work, it becomes crucial to effectively organize these elements into a straightforward and comprehensive structure. The efficient organization of the elements into a coherent and comprehensive structure is the critical point of this thesis, as the failure to do so may give rise to practical implementation challenges.

For the purposes of framework construction the structure presented in Figure 2.6 is adopted as the core element. In the preceding sections, the various challenges associated with choice modelling and data analysis in a broader context have been covered. The Figure 2.12 represents an updated version of the Figure 2.6, which will serve as starting point for the presented performance comparison framework. It will serve as a frame for individual elements consolidation into a easily understandable structure.

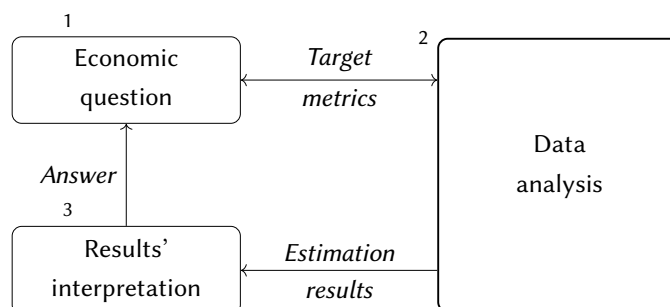


Figure 2.12.: Introduction to framework

This section unites the available insights, combining them into a structure that is both transparent and

easily comprehensible. The Subsection 2.5.1 combines all the elements of the data analysis part of the scientific procedure. The following Subsection 2.5.2 simplifies the resulting structure and incorporates it into the final version of the performance comparison framework. A discussion on the differences in framework's presentation in the context of applied and theoretical studies is provided.

2.5.1. Data analysis

Data analysis is a crucial component of scientific procedures, playing crucial role in data collection, transformation and modelling. In the proposed of scientific procedure, data analysis extends outside the scope of simple application of statistical and computational techniques to interpret the data. It involves the theoretical assumptions employed for data collection, transformation and results interpretation, as well as the complete scope of the toolset associated with modelling procedure. Thus data analysis should be understood as a multifaceted process that involves various techniques and theories, aimed at extracting meaningful information to answer the research question in alignment with the selected target metrics.

First of all we are going to reunite all the elements of data analysis procedure together. In this subsection the ideas expressed in Sections 2.3 and 2.4, as well as the Section 1.2.2, are combined to provide a macro-vision on the data analysis stage of the scientific procedure. The key complexity arises from the difficult to grasp interconnections between the various elements of the data analysis procedure. The emerging structure is presented in Figure 2.13. All the individual elements are regrouped according to the layouts presented in preceding sections.

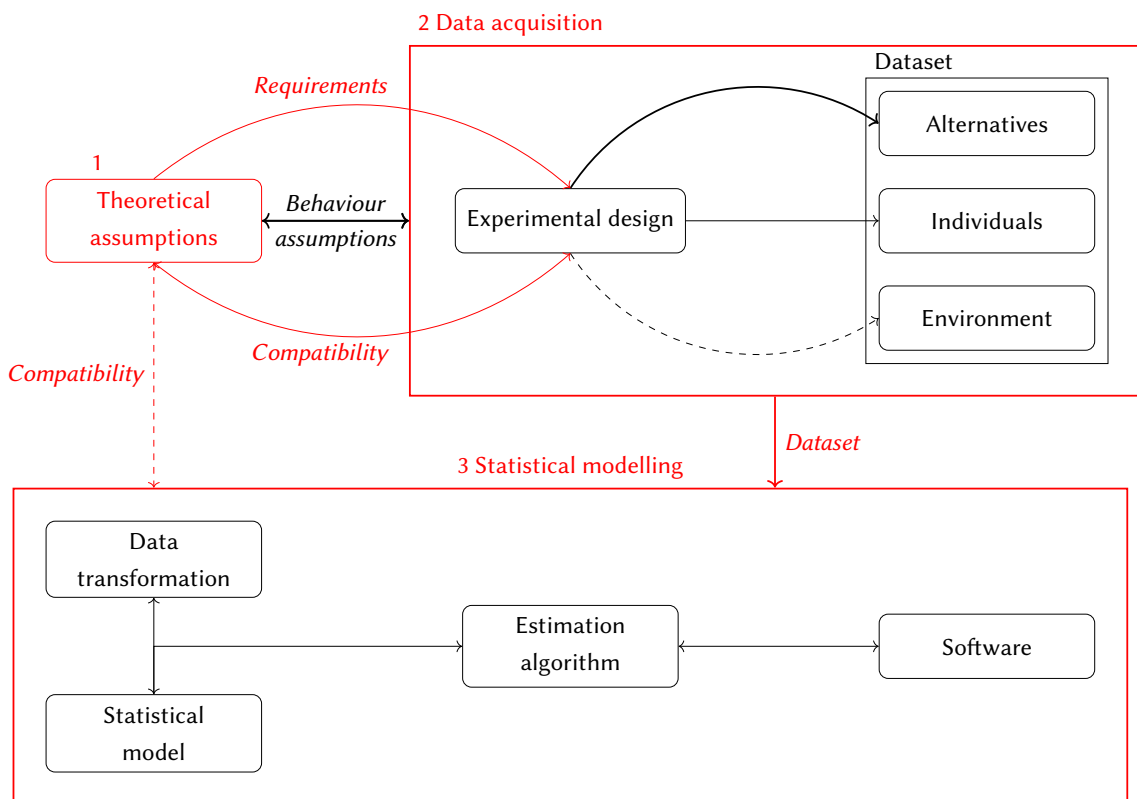


Figure 2.13.: All data analysis elements

The complexity of data analysis arises from the intricate interconnections between its various elements,

as illustrated. The resulting structure is complex and multilevel, taking into account as many data analysis elements as possible. Such structure risks only to confuse the non-proficient users who seek some simple guidelines for the applied case studies. The primary objective shifts at this point to simplification of the resulting structure for further combination with other elements of the performance comparison framework.

For the purposes of simplification of the data analysis procedure elements we should keep focus on the key elements of the data analysis procedure stage. As stated previously in the Figure 2.6 the key elements are: (1) theoretical assumptions, (2) data acquisition and (3) statistical modelling. We highlight them in red in the Figure 2.13, as they will be re-employed in a simplified version. The obvious simplification steps at this point involve the reduction of the framework elements to the previously identified big groups, while preserving the nature of assumed relationships among those elements. The Figure 2.14 offers this exact vision, which should lay at the heart of the resulting performance comparison framework.

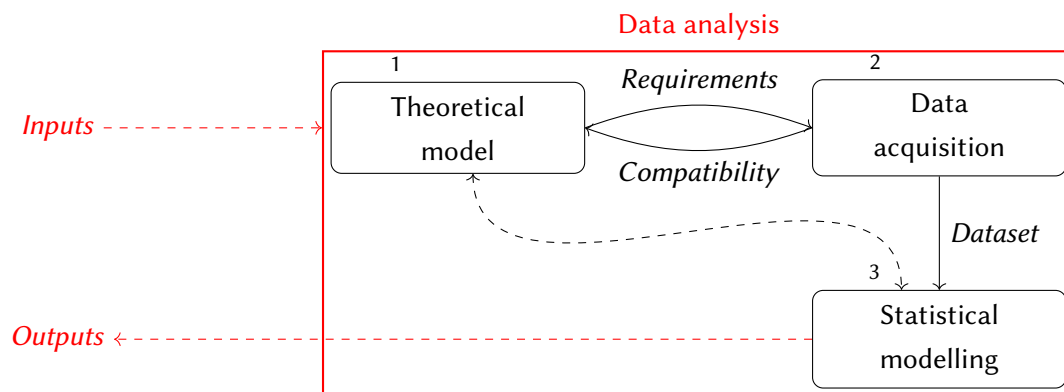


Figure 2.14.: Data analysis stage simplified

In this reduced format the data analysis part of the research procedure encompasses three essential components. First and foremost is the theoretical model, imposing the requirements on second component, the data acquisition process. The latter imposes limitations on the set of feasible theoretical models, through compatibility checks. Those involve various tests and hypotheses verification to validate the data adequacy with the chosen theoretical framework. The final critical element of the data analysis stage involves the application of statistical modelling techniques to the acquired data, employing methods tailored to the context of the chosen theoretical framework.

The data analysis stage in this case operates as a transformation process. It takes input requirements in terms of target metrics⁶ and generating output estimates that are suited to address the associated *research question*, for example an *economic question* in the context of economics studies.

2.5.2. Complete framework

Constructing a performance evaluation framework involves developing a structured and systematic approach to assess the effectiveness, efficiency, and impact of a scientific procedure, particularly emphasizing the data analysis stage. Previously the term “*performance*” was redefined as a metric depen-

⁶In fact the relationships with the eventual inputs are slightly more complex, which will be discussed in the Subsection 2.5.2.

dent on the entirety of the scientific procedure, referring specifically to the accuracy, reliability, and consistency in providing meaningful and trustworthy responses to the research question.

The next stage introduces the simplified data analysis stage vision into the integrity of the research procedure. We complete the framework with the remaining elements including the *economic question* defining the target metrics requirements for the rest of the procedure. The stage which implies interpretation of the statistical modelling results is also separated on Figure 2.15, this step is crucial as not all the data analysis stages produce meaningful and directly interpretable results.

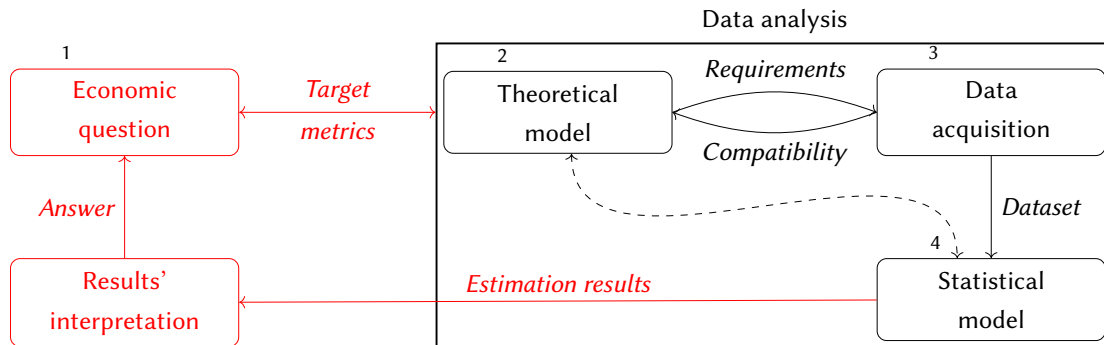


Figure 2.15.: Research procedure in detail

The introduction of an *economic question* definition into the research procedure implies the introduction of the target metrics to drive the data analysis stage.

Precision in addressing the economic question becomes the focal point of performance assessment. In which case there should also exist a step for backward transformation: the extraction of target metrics from the statistical model estimation results and their interpretation. Those elements correspond to *target metrics* and *answer* in the Figure 2.15. The same elements refer to the *inputs* and *outputs* for the data analysis part, presented in the Figure 2.14.

The *estimation results* have briefly made their appearance on the Figure 2.12, although at that time we did not provide any discussion on their inclusion into the framework structure. It is time to rectify this situation. In the discussion on the model performance perceptions provided in Sections 1.4.2 and 2.2.2, we have outline three key conceptually different ensembles of possible target metrics. Those include: (1) metrics relying on the model's direct outputs, computed choice probabilities or predictions; (2) metrics derived from the model's direct estimates, such as direct effects; and (3) derived metrics obtained by transforming the model's outputs or effects. While the first two elements are given by the *estimation results*, the third category requires further transformation applied to the results. This involves the process of *results' interpretation*, which is separated into an independent stage in the frameworks context. The *answer* thus refers to the final form of the refined estimation results, suitable to answer the research question and matching the requirements imposed by the chosen target metrics.

Alongside those eventual clarifications, it should be pointed out that the double edged arrow associated with the *target metrics* definition, assumes the possibility of inverse relationship at this point of the research procedure. From the conducted interviews (Appendix D) it became apparent that in some cases the research question arises from the already available data, or is adjusted to align with the familiar modelling techniques. Moreover, the target metrics and research question might be altered according to the preliminary results and research feasibility. In the theoretical studies, the research question might be selected to allow for better illustration of the theoretical concepts introduced in the

publication (Gusarov, Talebijamalabad, and July 2020). Thus the two-sided relationship at this point of framework should not confuse the reader.

While such simplification makes it much easier to understand the framework and graphically represent the research procedure, it also adds some inconveniences to the more in depth analysis of the various stage individual elements. While attempting to map existing studies according to our emerging frameworks we will see the eventual drawbacks of such approach. This drawback is not that prominent in frameworks application to the rather simple cases, although in the context of more theoretical studies the difficulties become more prominent.

Theoretical studies are typically focus on an external research question, matching the externalities ideas introduced in Section 2.2.1. In the context of theoretical studies, these externalities reflect the theoretical *research question*, which explores the behaviour of the research procedure under certain changes and modifications. The shift from applied studies to theoretical studies involves a focus on the path to answering *economic questions*.

This means that focus shifts to the optimisation of the various stages of scientific procedure, seeking answer as to how the changes in research procedure impact the obtained answer to economic question.

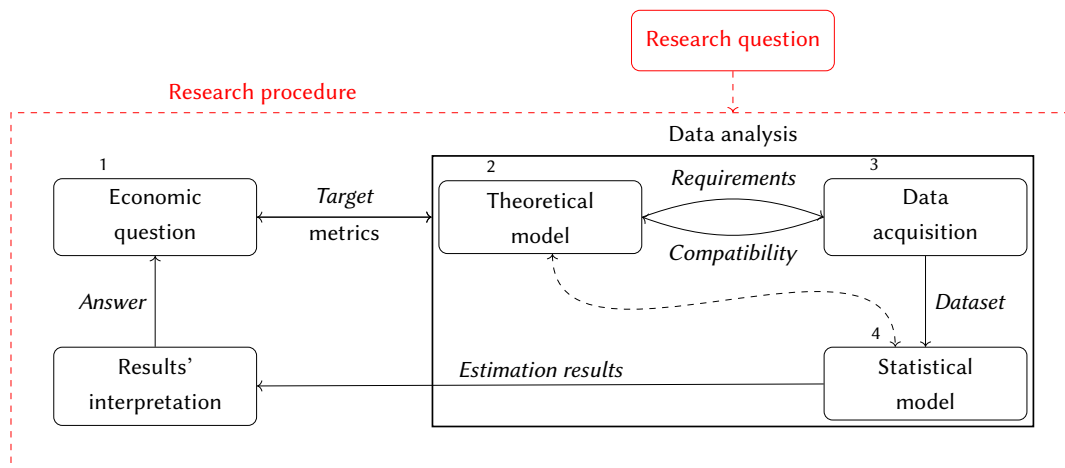


Figure 2.16.: Proposed performance comparison framework

The performance analysis framework, valuable in the structure it brings to the performance assessment and comparison task, comes with certain limitations and potential inconveniences. Those limitations are tightly linked to the rigidity of the simplified structure. To conclude this part of the limitations should be outlined to avoid any misuse of the proposed toolset.

The major strength and limitation at the same time is the generality of the resulting structure. While it is relatively simple to integrate any scientific publication to the proposed structure it is relatively complex to account for all the minor changes, modifications and changes in a relatively concise graphical format. A part of this limitation may be negated by usage of the simplified representation only as a set of recommendations for the research procedure construction and subsequent performance evaluation, rather than step-by-step guide.

Another issue is equally related to the high level of simplification allowed in the final framework's version. The generalised format omits a lot of information and dimensions that should be accounted for at each stage of the research procedure. This problematic has much more dangerous repercussions for the end users, as guided solely by framework, they may easily forget to incorporate in their works

information on one or multiple crucial elements of the scientific procedure.

2.6. Existing studies in framework context

Once the framework structure is well defined we may proceed with presentation of its potential in procedure performance evaluation. We are going to demonstrate how existing case studies enter the framework structure. This presentation will be divided into two separate parts.

First of all, within this section we are going to inscribe the existing reference studies available in the literature into the resulting framework structure. In the Chapter 3 a series of novel studies performed by us will be presented, those studies were designed relying on the performance comparison framework. It will be possible to observe the evolution of the framework from the early stages of the thesis to its final form as presented above.

Following the order of elements within the data analysis stage of the framework, and scientific procedure, we explore several works, each addressing different elements of this structure. The work of M. Bierlaire, Bolduc, and McFadden (2008) represents a classical example of an econometrics oriented methodological work proposing a novel estimation strategy for the traditional choice models. The publication of Rose and Bliemer (2013) addresses the issues of data collection. Through a controlled simulated experience they explore the dataset requirements in the context of choice modelling. Finally, the work of Balbontin, Hensher, and Beck (2022) contrasts the first two examples, representing a traditional applied study deprived of simulation and methodological issues exploration.

This offers us a support of three works different in their nature. One representing a classical example of a fully applied study. Another being a perfect illustration of a fully methodological analysis. And finally an example of innovation oriented theoretical study offering novel statistical approach to choice model estimation.

2.6.1. Applied study procedure

As an example of the applied study procedure we take the relatively recent work of Balbontin, Hensher, and Beck (2022), entitled “*Advanced modelling of commuter choice model and work from home during COVID-19 restrictions in Australia*”. As it could be seen from the title the authors perform an analysis of commute choice situation in a relatively restricted geographical and social context. Even though a relatively complex HCM statical technique is implemented for the choice modelling purpose, the study does not offer any particular meta questions except for the classic economic questions conventional for such case-study.

The economic question in this case is identical to the general research question and is dictated by rather simple reasoning. The decision to *work from home (WFH)* or commute during COVID-19 has had a significant structural impact on individuals’ travel, work, and lifestyles, according to the authors. This non-marginal change is influenced by various factors, some of which are quantifiable through objective variables, while others are best understood through underlying latent traits, which justifies the implementation of a HCM model. Those latent traits include attitudes towards WFH and the use of specific modes of transport for the commute, especially those associated with bio-security risks like *public transport (PT)*. The research attempt to “*identify the nature and role of underlying attitudes,*

perceptions and beliefs that influence the decision to work from home for a specified number of days per week, and how this relates to the incidence of commuting by day of the week and time of day” (Balbontin, Hensher, and Beck 2022).

The theoretical stage of the work included the adoption of classic RUM framework for the purposes of choice modelling, enhanced with such advanced elements as latent psychometric concepts. This framework dictated the selection of both. The authors conducted research in which they developed and implemented a HCM to investigate the sources of influence, accounting for the endogenous nature of latent soft variables for workers in metropolitan areas in New South Wales and Queensland. Data for this study was collected between September and October 2020, a period characterized by minimal lockdowns and relatively minor restrictions on workplaces and public gatherings.

According to the authors’ findings, one of the most significant factors contributing to a favourable attitude towards WFH is the workplace’s WFH policy. Workers who have the flexibility to decide where they work are more likely to embrace WFH, and those who are directed to do so also exhibit a higher inclination compared to individuals under other workplace policies. Additionally, the authors noted that bio-security concerns, particularly in relation to shared transportation modes such as public transport, play a pivotal role in influencing individuals to opt for WFH or choose to commute using the safer option of private cars.

The resulting research procedure in the context of performance analysis framework is introduced in the Figure 2.17. This case serves to illustrate the research procedure in the context of applied studies. Any changes within this procedure represent a perfect ground for further performance comparison.

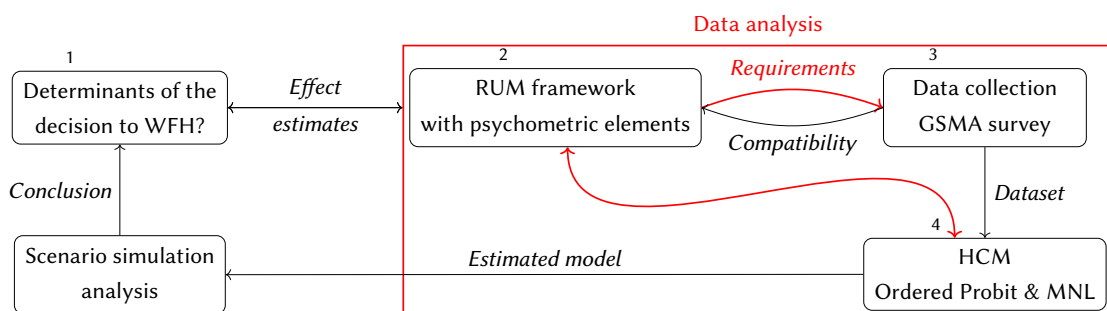


Figure 2.17.: Commuter choice model and WFH analysis research procedure

For example, addition of another concurrent statistical model in the data analysis procedure stage immediately lays ground for the performance comparison task. The same applies to any other singular or multiple changes of other elements of the procedure, be it another data collection method or theoretical assumptions about the individual behavioural patterns.

2.6.2. Theoretical innovation introduction procedure

Following the previous illustration of the frameworks consistency on a simple applied study case we proceed with more complex studies. For the second example we select a study proposing a new estimation technique for rather common at the time NL models performed by M. Bierlaire, Bolduc, and McFadden (2008). The study is entitled “*The estimation of generalized extreme value models from choice-based samples*” and addresses the issue of estimation difficulties of the classic choice models in the presence of choice-based samples.

In the context of data collection employing choice-based sampling strategies, the authors point out that the property of MNL models, which typically allows for consistent parameter estimates except for the constants through *Exogenous Sample Maximum Likelihood (ESML)* estimation, does not generally extend to *Generalised Extreme Value (GEV)* models. The obtained results might be erroneous and lead to erroneous conclusions on later research stages and more importantly in the policy repercussions. To tackle this issue the authors propose a consistent ESML estimator tailored for GEV models in this scenario.

The authors start their paper with the identification of a specific subset of GEV models that share a desired property with MNL models, where the constants are capable of absorbing potential biases. In applied part of the paper the accent is put on the NL models. Subsequently, they introduce a novel and straightforward *Weighted Conditional Maximum Likelihood (WCML)* estimator suitable for the broader context of GEV models. Unlike the *Weighted Exogenous Sample Maximum Likelihood (WESML)* estimator by Manski and Lerman (1977), the new WCML estimator does not necessitate external knowledge of market shares for correct functioning.

The authors emphasize that this approach remains applicable even when alternatives are drawn from an extensive choice set. They provide practical demonstrations of the estimator's utility using both synthetic and real data. Nevertheless, the synthetic data plays a major role in the research presentation, as it sets up an observable target for the concurrent estimation algorithms to compete for. The performance of statistical models estimated by both WCML and ESML algorithms are explored in the paper. For this purpose the simulated population is subsampled multiple times and the statistical properties of the estimates' reliability are compared.

The research procedure for this particular case study may be interpreted in the format as represented in Figure 2.18. Although in this case it becomes rather complex to present the whole procedure without descending onto a lower level of data analysis procedure elements, as was proposed previously, the presentation still remains rather synthetic and clear.

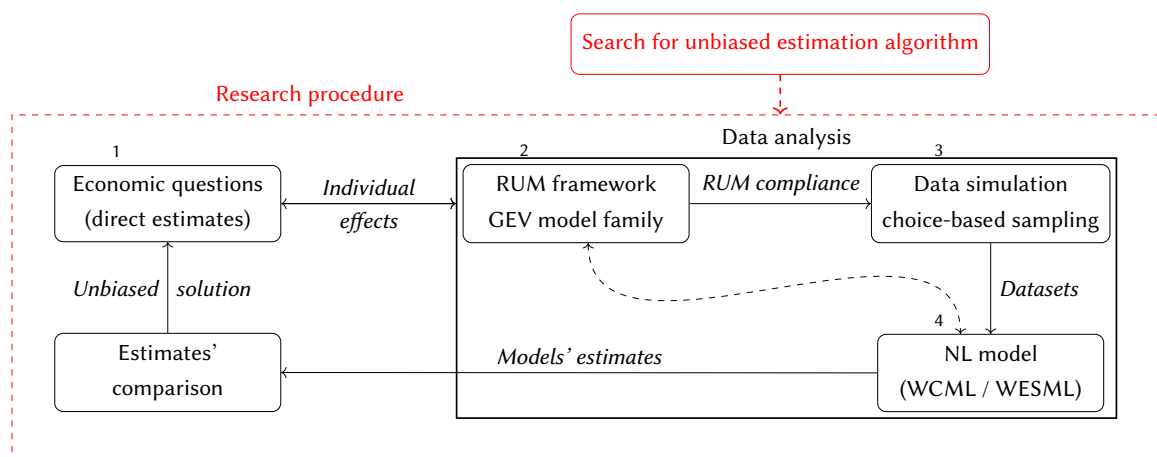


Figure 2.18.: WCML algorithm for GEV models in choice-based sample presence

This example illustrates us how the different questions are related within a single research procedure. While the baseline research question regroups an ensemble of the potential research questions and cases reliant on the estimates precision, the *meta*-question addresses the issue of algorithm efficiency in the estimates production. While the first element thus sets up a performance metric target, the

second question direct the entirety of the research procedure.

This particular case study also demonstrates that the ideas similar to the presented as a part of the framework are already present in literature, although lacking a systematic part. The framework aggregates those existing practices, providing a systematic and structured approach for evaluating the performances of different models within the context of scientific procedures.

2.6.3. Theoretical study procedure

The final example in this sequence represents a state of the art methodological work. The publication by Rose and Bliemer (2013) bearing the title of “*Sample size requirements for stated choice experiments*” explores the effects of sample size variation on the stated choice experimental data analysis results. While in the literature there exist some basic statistical tools to identify the desired sample size in the context of some baseline model application, for more complex models, such as multiple choice models, there are no evident mathematically justified solutions. And while it is rather simple to determine the limits of models’ identifiability, it is much more complex to determine the optimal sample size, given that typically the *Stated Choice (Sc)* are subject to budgetary constraints. While the budgetary constraints apply to the most cases of data collection, it is in experimentation that the cost of an additional data-point is relatively costly in comparison with other disciplines.

As pointed out in the paper, in the study of behavioural responses among individuals, households, and various organizations, SC experiments serve as the predominant data paradigm. Despite this, there is an historical lack of understanding regarding the sample size prerequisites for models derived from such data, or at least was at the time of the study conduct. Traditional orthogonal designs and existing sampling theories have proven inadequate in addressing this issue. Consequently, researchers have resorted to simplistic rules of thumb or, at times, overlooked the matter altogether, collecting samples of arbitrary sizes in the hope that they would be sufficiently large to yield dependable parameter estimates. In some cases, researchers have been compelled to make assumptions about the data, assumptions that are unlikely to hold true in practical situations. This background determines the research question for this theoretical study.

In this paper, the authors illustrate how a previously proposed sample size calculation method can be leveraged to create what are known as S-efficient designs[^] [For information, here is a short note on the different *Efficient ED* types (van den Broek-Altburg and Atherly 2020):

1. *D-efficient Designs* minimize the determinant of the *Asymptotic Variance-Covariance (AVC)* matrix under the assumption of a vector of prior coefficients β
2. *A-efficient Designs* minimize the trace of the AVC matrix
3. *S-efficient Designs* minimize the maximum sample size required for statistically significant parameter estimates]. These designs use prior parameter values for the estimation of panel mixed multinomial logit models. The authors delve into the sample size requirements essential for such designs in the context of SC studies. The theoretical background imposes the restrictions of classic RUM compliant framework as authors focus their attention on the MMNL becoming a baseline statistical model at the time.

In a numerical, simulated case study, they demonstrate that a D-efficient design (J. L. Walker et al. 2018), and even more so, an S-efficient design, necessitate a notably smaller sample size than a ran-

domly orthogonal design to estimate all parameters with statistical significance. Furthermore, they highlight the positive impact of a wide range of levels on the efficiency of the design and, consequently, the reliability of parameter estimates.

The research procedure structure may be summarised up in the Figure 2.19. The task of representing all the choices made by authors in the research procedure construction remains challenging at this point. While the complete version of the performance comparison framework would have had the possibility to account for all the changes and alterations brought into the procedure, the reduced version should be more accessible for the end-user.

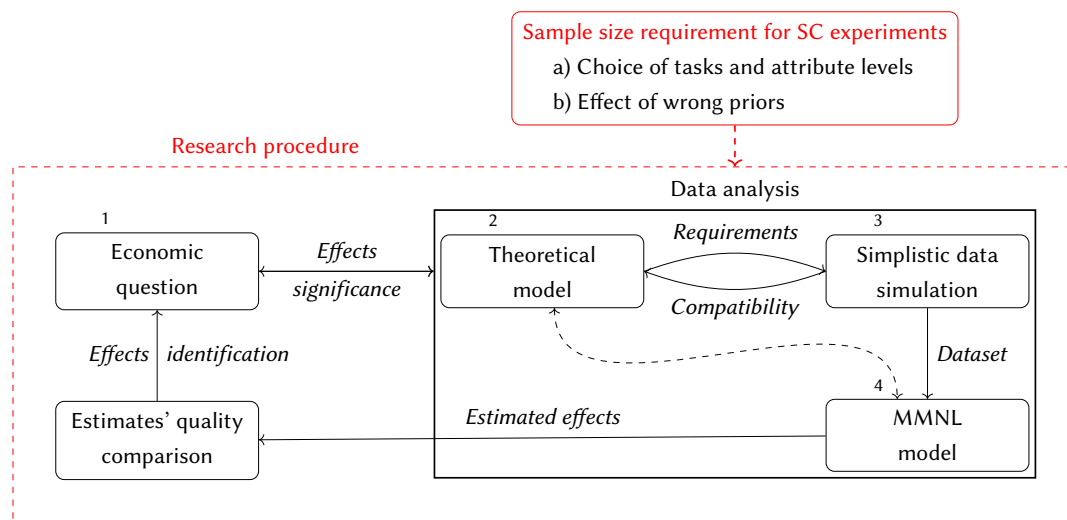


Figure 2.19.: Sample size requirements exploration for SC experiments

In this particular case the difficulties in graphical representation arise from the complexity of analysed changes and alterations in the *data acquisition* step of the procedure. Not only three different experimental designs are contrasted in the work, but the simulated dataset is also subject to changes in sample size.

Nevertheless, for the purposes of procedure design and framework implementation on the future research such limitation bear nearly no restrictions. Considering that for methodological analysis it might be more convenient to use the full framework schema without any particular simplifications, in contrast to the more simple applied research cases.

2.7. Concluding remarks

In conclusion, this chapter presents a comprehensive framework for performance analysis and comparison in the context of DCM use-case. The motivation behind the framework lies in the inconsistencies in the literature across different application fields. The model performance understanding, the terminology and different research problematic requirements explain the diversity of strategies for model performance assessment in the literature. In order to tackle this issue a shift in focus is made from the model comparison to the comparison of research procedures. In this chapter we outline a universal approach that aligns with the standard scientific procedure: the framework aims to bridge the interdisciplinary gap in existing literature. This framework offers a flexible toolset for assessing and comparing modelling and data analysis techniques in DCM related research.

The chapter separates the research procedure into a set of key elements, each being introduced and discussed separately. It emphasizes the importance of considering theoretical foundations, data acquisition, and statistical modelling techniques in a combined manner. This lies ground for the performance assessment and comparison in the context of interdisciplinary studies, where the compared methodologies are extremely heterogeneous for standard comparison methods. The proposed framework is positioned as a valuable tool for researchers to perform model performance evaluation, providing a systematic and detailed approach, for both experts and non-proficient users.

The framework's presentation in Section 2.5 serves as a crucial step in consolidating various elements into a cohesive structure, ensuring clarity and ease of implementation. It is designed to be accessible to users, whether they are novices or experienced practitioners, offering a step-by-step guide through the intricacies of discrete choice analysis. The subsequent section, 2.6, showcases the application of the framework through existing case studies. It illustrates the flexibility and adaptability of the proposed methodology for different types of studies, from theoretical studies to applied research.

The framework consolidates the established methods applicable to performance comparison task, offering a systematic and structured method for assessing the performances of various models in the context of DCM related scientific work and data analysis in general. Not only the framework help in understanding the strengths and weaknesses of different models, but also simplifies the replication of future works and results by other researchers. It enhances transparency and reproducibility in research, identifying the key elements of research procedure in the DCM context. As a bonus, the framework can easily be extended to non-economic related research questions and disciplines. Its flexible nature incorporates variations in theoretical assumptions, data acquisition, and modelling procedures, making it applicable to a wide range of scientific inquiries.

This chapter lays the ground for the subsequent exploration of the framework's capabilities in new scenarios, available in Chapter 3. The proposed framework not only addresses the challenges associated with DCM analysis, but also contributes to the broader discourse on performance evaluation methodologies. The proposed performance comparison toolset might be easily extended to other disciplines and use-cases, outside the scope of DCM. It offers a harmonized perspective that united the concepts of theory, data, and modelling. Through this work, an attempt is made to communicate a more unified understanding of model performance, facilitating interdisciplinary dialogue and advancing the DCM analysis in context of interdisciplinary studies.

3. Framework in action: Case studies

To support and illustrate the performance comparison framework proposition we offer several applied studies making use of the framework. The reader will encounter 3 examples of its implementations serving to answer different research objectives and requirements. Presented case studies focus on the different element of the framework: (1) theoretical assumptions and target metrics selection effects, (2) data acquisition and related issues, and (3) model comparisons. Each section of this chapter presents one study in original format, as it was presented or submitted for conferences. While the original layout for those works could not be preserved due to the restrictions on formatting of the thesis, the articles remain unchanged in other aspects, preserving the original syntax and punctuation. In some cases this leads to usage of American English, instead of British syntax employed for the rest of the work.

The first paper addresses the issue of how the incorrect incorporation of theoretical assumption into the model may alter the target metrics. The later are explored not only from the perspective of plain predictive qualities, but also in term of resources usage efficiency and direct effect estimates derivation. The second work explores and analyses the effects of the dataset configuration changes on the complex target metrics, such as willingness to pay and value of time in particular. The third and final work puts in concurrence the different statistical modelling approaches available to the researcher. It explores the differences between classic RUM oriented models, such as MNL and NL, alongside with the emerging RUM-compliant ML techniques, including the *Alternative Specific Utility Deep Neural Network (ASUDNN)*.

Each article is accompanied with an explicative note, adding eventual clarifications and corrections to the original works. Those studies allow to illustrate the changes undergone by the performance comparison framework throughout the thesis project.

3.1. Introduction

In the previous chapter reader was introduced to the performance comparison framework, which brings novelty and consistency unifying the existing approaches to methodological and applied studies based on the choice analysis. Several examples were offered on how the existing studies relate to the framework's structure. In this chapter a series of case studies follows, offering a more in-depth understanding of the frameworks functions and use-cases.

Each of the case studies is focusing on the different element of the framework: (1) modelling stage relationships with the theoretical assumptions and target metrics, (2) data acquisition issues, addressing in particular the dataset configuration; and (3) statistical modelling. Obviously the framework incorporates much more steps and stages, though given limited time we had a chance to focus only on several of the key elements. All of those elements are part of the data analysis procedure stage, completing the framework use-cases presentation.

The first study, Section 3.2, combines econometrics and ML models for consumer choice preference modelling, addressing interdisciplinary challenges. It is the first to introduces the simulation and theory-testing framework. The framework's adaptability across economics and statistical indicators is exemplified using Michaud, Llerena, and Joly (2012) work. Three models from econometrics and ML are estimated and compared over two synthetic datasets with predefined utility functions, simulating homogeneous and heterogeneous preferences.

The second study, Section 3.3, focuses on WTP elicitation task, which is a widely used metric to assess individuals' preferences for attributes in economic choices. The study extends the performance comparison framework to organize previous research and systematically evaluate model performance in the WTP elicitation task, considering potential misspecifications, changes in sample size, and dataset balance. A synthetic dataset is employed for practical application, with simulations altering sample size and configuration for model estimation and WTP elicitation. The findings demonstrate the variability in WTP estimates across different configurations.

The third case study, Section 3.4, explores the comparison between econometric and ML models in the context of commute mode choice modelling. It evaluates traditional discrete choice models against emerging ML approaches within the context of economic indicator elicitation, specifically WTP. The study utilizes the well-known *swissmetro* dataset and generates synthetic samples. It then contrasts conventional discrete choice models (MNL and NL) with emerging ML alternatives, among which ASUDNN (S. Wang, Wang, and Zhao 2020), in the WTP estimation task. As a work in progress, this case study does not offer new evidence on the topic, but serves to illustrate the eventual issues associated with model performance analysis and comparison for the purposes of model selection.

Because all of the individual papers were produced at the different stages of maturity of the thesis, there might be some inconsistencies in the framework vision and presentation. An evolution of the framework may be traced over those works, as their order matches the chronological order of their production. We equally preserve the original format of those papers, including wording, definitions and spelling. Each section within this chapter starts with an abstract for the respective article, followed directly by the contents. At the end of each work we offer a short discussion addressing the eventual insights obtained from the work. The differences in between the elements figuring in the paper and the final version of the framework are equally put in evidence. Such structure should help the reader

with comprehension of the works' limitations, as well as with its positioning in the context of the final framework version.

3.2. Case 1: Theoretical assumptions

DA2PL 2020 paper: “*Exploration of model performances in the presence of heterogeneous preferences and random effects utilities*”

Abstract

This work is a cross-disciplinary study of econometrics and machine learning (ML) models applied to consumer choice preference modelling. To bridge the interdisciplinary gap, a simulation and theory-testing framework is proposed. It incorporates all essential steps from hypothetical setting generation to the comparison of various performance metrics. The flexibility of the framework in theory-testing and models comparison over economics and statistical indicators is illustrated based on the work of Michaud, Llerena, and Joly (2012). Two datasets are generated using the predefined utility functions simulating the presence of homogeneous and heterogeneous individual preferences for alternatives' attributes. Then, three models issued from econometrics and ML disciplines are estimated and compared. The study demonstrates the proposed methodological approach's efficiency, successfully capturing the differences between the models issued from different fields given the homogeneous or heterogeneous consumer preferences.

3.2.1. Introduction

Consumer choices data are mainly modelled through classification tools from *Machine Learning* (ML) or econometric techniques. Economists and demand analysts deepen these analyses by studying consumers' willingness to pay (WTP). These economic measures are traditionally deduced from the assumed consumer behavioural theory underlying the estimated econometric model. In applications, the WTP are directly analysed or deduced from ML tools, for example from recommendation system (Scholz et al. 2015). These approaches of economic and behavioural indicators illustrate one of the differences between the two disciplines applying statistical learning. As described by Breiman et al. (2001) and later by Athey and Imbens (2019): the ML focuses on the predictive qualities and econometrics attempts to decipher the underlying properties of the data. The hypothetico-deductive approach of econometrics allows the production of economic indicators under validity conditions of the model hypotheses, on which ML tools do not depend. This difference between approaches can then be viewed as a constraint or as an opportunity of the methods.

A specific focus of economists is the estimation of WTP of consumer for goods or goods' attributes. Multinomial regressions have been proposed in the literature to manage several behavioural assumptions, among which the heterogeneity across individuals by allowing taste parameters to vary in the population. Many choice experiments collect consumption choice data and analyse them with a mixed logit model which provides the advantage to consider such heterogeneity. Nevertheless many questions arise from the assumptions surrounding the introduction of the taste heterogeneity in the behavioural model distribution choice of the parametric form, inter or intra-consumers heterogeneity

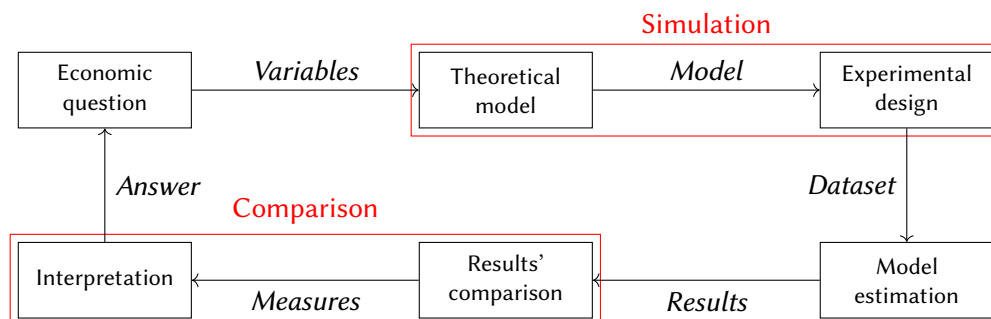
(Hess and Rose 2008; Danaf, Atasoy, and Ben-Akiva 2020), leading to many multinomial model competing specifications.

Switching focus to explanation of the findings clarifies why many of the advanced ML techniques rarely appear in economics publications. This is because of their believed lack of interpretability. Nevertheless, some pluri-disciplinary scientists make attempts to breach this wall between ML and econometrics: Athey and Imbens (2019), Mullainathan and Spiess (2017), Varian (2014). Their advances are mostly focused on the general interdisciplinary question, without entering into the application specific details.

Objective of the paper is to evaluate and contrast performances of the two approaches, ML and econometrics, facing consumers preference heterogeneity. There have already been a multitude of studies comparing the performances of different econometric and ML models in various real world scenarios: the study of ML methods to model the car ownership demand estimation of Paredes et al. (2017); or the use of decision trees in microeconomics of Brathwaite, Vij, and Walker (2017). The performance of competing models are studied according to several different criteria: (1) in terms of the quality of data adjustments; (2) in terms of predictive capacity; (3) in terms of the quality of the economic and behavioural indicators derived from estimates; and (4) according to their algorithmic efficiency and computational costs. However, to the best of our knowledge, there is no work providing a complete and comprehensive methodology for assessing and comparing model performances given the context of consumer preference studies.

The study proposes a theory-testing framework (Figure 3.1) exploring the performances of different econometric and ML models in presence of preference heterogeneity among individuals. More specifically, we assume the data are coming from choice experiment designed for value elicitation (David A. Hensher, Rose, and Greene 2005; Louviere, Hensher, and Swait 2000). In order to produce reliable results we construct two simulated datasets with homogeneous and heterogeneous preferences structures respectively. We start with the general methodological presentation of the undertaken procedure taken within the limits of the proposed theory testing framework. The second section on the work presents the simulation and estimation results, ending with a comparison of the performance metrics for the selected models. The last section concludes.

Figure 3.1.: Proposed framework



3.2.2. Methodology

We aim to observe how some minor changes in theoretical decision model (specifically taste heterogeneity) may affect the results of value elicitation in choice experiment. We generate artificial datasets

based on predefined behaviors and predetermined statistical properties for individual characteristics and alternatives' attributes. Such a set-up ensures that we know the exact data generation process and have all the control over the parameters and experimental design.

We use the work of Michaud, Llerena, and Joly (2012) to settle a realistic economic context of the empirical part of our work and to provide reasonable behavioural target values for the choice rules and tastes parameters. The work in reference investigates the impacts of the environmental attributes in the context of a consumer choice of non-alimentary agricultural goods. The Valentine's Day red rose is the good of the experiment. Subjects are faced with choice situation composed of two identical red roses by aspect with different specifications and an opt-out option. Product specifications are different among choice situations following random design technique. Among products attributes, *price* and two environmental aspects of roses' production are retained. The *eco-labelling* indicates the cultivation environment and conditions. The *carbon footprint* two-level factor measures the greenhouse gases emissions during the cultivation and transportation. The consumers are at the same time observed by four main socio-economic characteristics: *age*, *gender*, *income* and individual *habit* to acquire eco-labelled goods.

3.2.2.1. Artificial dataset

Generation of synthetic datasets is a common practice in many research areas. Such data is often generated to meet specific needs or certain conditions that may not be easily found in the real world data. The nature of the data varies according to the application field and includes text, graphs, social or weather data, among many others. The common process to create such synthetic datasets is to implement small scripts or programs, restricted to limited problems or to a specific application. In this work the simulation of the two datasets involves: (1) generation of an artificial population with characteristics issued from a set of predefined distributions; (2) creation of an experimental set-up based on a specific choice set ensemble; (3) simulation of the individual choices for given population and alternative sets using an arbitrary defined decision rule.

Following the reference paper (Michaud, Llerena, and Joly 2012), we consider the same four socio-economic characteristics. These four characteristics define the generated artificial population. For simplicity, we assume that these characteristics are not correlated. Sex and purchase habit are both binary variables generated separately with random draws from a Bernoulli distribution. To generate the class variables, age and income, we convert to the discrete-continuous multilevel scale draws from normal distribution defined by the mean and standard deviation parameters from the reference paper.

Stated choice (SC) experiments face sampled respondents with several different choice situations, each consisting of a comprehensive, and yet finite set of alternatives defined on a number of attribute levels. Based on this, respondents are asked to select their preferred alternatives given a specific hypothetical choice context. The experiment is designed in advance by assigning attribute levels to the attributes that define each of the alternatives which respondents are asked to consider (Rose et al. 2008). In this research, we have implemented modified full factorial (FF) design following the ideas of the original paper, where the concern of reducing the number of choice situations' number was addressed. To make complete FF design taking into account the prices of the alternatives, we would have been faced with the nearly infinite number of distinct alternatives. To tackle this, we generate initial choice sets based on two binary variables using the FF design. We assume that individuals are presented with two

unlabelled alternatives, roses A and B , as well as a no choice alternative (denoted C). The two attributes, eco-label and carbon footprint, have two levels which make four possible combinations for one alternative and 16 possible combinations in the case of multiple choice set-up (the no choice alternative has the levels fixed to zero). The prices are then randomly assigned to the predefined alternatives guiding the learning by adding potentially non-existent alternatives. Our simulated experimental design finally ‘ask’ the subjects to repeat 10 times their choices on new random designs in order to capture individual specific elements and achieve better statistical convergence.

Consumers’ decisions are analysed with the discrete choice framework based on the utility maximisation assumption. This framework assumes that consumers associate each alternative in a choice set with a utility level and choose the option, which maximises this utility. The general estimation framework of the Random Utility Model (RUM) proposed by McFadden (1974) provides the opportunity to estimate the effects of product attributes (denoted as γ) and individual characteristics (β) and to compute WTP indicators. The deterministic part of utility function is given as follows¹:

$$V_{ij} = \alpha_{i,Buy} + \beta_{Buy,Sex}Sex_i + \beta_{Buy,Age}Age_i + \beta_{Buy,Income}Income_i + \beta_{Buy,Habit}Habit_i + \gamma_{Price}Price_{ij} + \gamma_{i,Label}Label_{ij} + \gamma_{i,Carbon}Carbon_{ij} + \gamma_{i,LC}LC_{ij} \quad (3.1)$$

For different datasets the individuals are assumed to have homogeneous or heterogeneous preferences for the environmental attributes of alternatives. Each individual had his personal attitude to the eco-label and carbon footprint of the roses, determined by their awareness of the environmental issues.

In order to calculate utilities, we took parameters from the paper of reference (*a priori*). We started with calculating the relative deterministic utilities respectively for each individual and alternative, assuming that no choice option has zero utility for everyone. After adding some random noise, following the Gumble distribution parametrised with $(0, 1)$ we select the alternative with highest utility per each individual per each choice set. We took no-choice as reference alternative. This procedure is described in detail during obtained dataset presentation.

3.2.2.2. Modelling consumer choices

Adopted econometric models are multinomial logistic regression (MNL) and mixed multinomial logistic regression (MMNL), the later being of the possible generalisations of the former. The third model, a simplistic version of convolutional neural network (CNN), comes from the ML disciplines. Such models are rarely implemented by the economists in their studies since this family is usually perceived not to offer enough insight when it comes to the effects estimation. The ML techniques are usually viewed by economists as some black boxes, which do not provide any information about the underlying process. It is quite easy to comply with their position. Although the most advanced techniques perform better in terms of predictive power, they rarely offer any insight into the modelling process. The chosen CNN is adjusted to answer the economic question through modelling of the relative deterministic utility functions.

The two econometric models are perfectly adapted to model one of the two generated datasets respec-

¹Where $LC = Label \times Carbon$.

tively². The MNL model should yield the best performance results on a dataset assuming fixed effects, while its counterpart, MMNL, should be the most performant in the presence of random effects in the utility function. The MNL model assumes that the decision makers view the available alternatives to be independent and that attribute impacts are fixed for the whole population across all alternatives (McFadden 1974). This assumption is relaxed in the MMNL, where coefficients (or some of them) vary for each individual (Agresti 2013). The logistic regression models are derived from GLM specifications (Agresti 2007). This class of models relies on the hypothesis, that each individual maximises his perceived utility over a closed set of alternatives. His utility is assumed to be determined by a fixed and a random parts. The probability structure incorporates the theoretical assumptions of the finite choice set, the uniqueness of the chosen alternative and the idea of utility maximisation. Many of the existing applied econometrics papers use the most simple MNL model, which may lead to erroneous results and conclusions in the presence of random taste coefficients in the utility.

The model issued from the ML field focuses on more advanced and atypical modelling techniques. The neural network (NN) models can be viewed as an even wider generalisations of the generalised additive models (GAM), and are capable to imitate more simple models similar to MNL. The resulting CNN comprises two layers: (1) convolutional layer and (2) *softmax* transformation layer. The convolutional layer transforms the linear combination of individuals characteristics and alternatives' attributes into the relative deterministic utilities. Then, the utilities are passed to the *softmax* layer with fixed weights to derive the resulting choice probabilities. This choice was made since the seemingly identical models by their structure may produce different results depending on the implemented estimation techniques. The NN's offer us a great number of different algorithms which are more advanced than the algorithms traditionally implemented in econometrics, which make us wonder whether the changes in the estimation algorithm will allow us to achieve better results. In this study we use *Adam* algorithm (Kingma and Ba 2017) for CNN estimation, which is parametrized according to *Keras*³ standards, with increased learning rate (fixed at $1e - 1$).

3.2.2.3. Performance measures

In the first place we are interested by the overall goodness in estimation of the utility function components. In this task we should compare the obtained estimates with the target values we have settled into the utility functions. The best model should produce the mean estimates, which are equal to the targets, with the minimal variance possible.

Secondly, we are attracted by the WTP for roses and the premiums associated with particular alternative specific attributes. These were the only target metrics present in the article of Michaud, Llerena, and Joly (2012). The WTP could be read as the value the consumers are willing to pay for a rose. At the same time, the premiums may be translated as how much consumers are ready to pay for a unit change of a given attribute of the product. Both the WTP for a product and the premiums can be computed as the marginal rates of substitution between the quantity expressed by the attributes and the price (Louviere, Hensher, and Swait 2000). These theoretical values could be easily derived for all the three explored models, calculated as ratios of the corresponding coefficient (or weights). They will allow us to compare how close the derived values are to the theoretical input values, which were defined on the dataset generation step.

²For model estimation we use *mlogit* package, version 1.1-0 from CRAN

³Version 2.3.0.0 from CRAN

Thirdly, it is important to assess the overall goodness of fit over the whole dataset for the selected models. To address this issue, the best suited measure is the *accuracy*, describing the part of correctly classified instances in a given set and is by its nature a complement to the empirical error-rate measure (Japkowicz and Shah 2011). Doing so, we will be able to observe the ratio of the overall correctly modelled choices. We may as well implement the Kullback–Leibler Divergence (KL or KLD) estimator for overall goodness of fit. This will allow us to quantify the difference between the estimated posterior distribution and the true underlying distribution of the choices.

Finally, we observe the performance of these different models in terms of computational efficiency in resources consumption. For this task we will observe the computation times for given models. This measure is one of the most complex, because it accounts at the same time for different models, different estimation algorithms, different numerical implementations in the statistical software and different PC configurations. It is valid in this particular case, because all models were estimated using the same hardware and software set-up.

3.2.3. Results

We present the obtained results in several steps. First of all a discussion on the simulated datasets is provided. Then we present the estimation results and present to the reader the goodness of relative deterministic utility function coefficients estimates for the different models. Finally, we provide an extensive discussion of the performance results.

3.2.3.1. Data

Each artificial datasets regroups 1000 artificial individuals, each of them faced with 16 different choices 10 times with random prices allocation (160 choice situations in total), hence, 160000 observations per dataset. In both situations the utility functions are determined as in paper: we use the exact means for the coefficients estimates, assuming they are correct (Table 3.1a). The variance-covariance matrix for RUM individual coefficients is supposed to be a matrix of zeros for the homogeneous preferences case and to be as in the reference paper for the heterogeneous preferences dataset (Table 3.1b). These coefficients are then randomly assigned within population with draws from a multivariate normal distribution.

It is interesting to explore the statistical properties of the two resulting artificial datasets and the original one, gathered by Michaud, Llerena, and Joly (2012)⁴. ANOVA and χ^2 tests (table 3.2) show no significant means difference between the simulated datasets and the original one, except for the *Income* variable. This is explained by the implemented dataset generation procedure based on transformation of draws from symmetric normal distribution. The distributions of *Carbon* footprint and *Eco-Label* attributes follow the ones inside the original dataset, while the distribution of price differs (table 3.3).

This particular divergence, may be explained by the procedure implemented to assign prices to the alternatives inside choice sets, because the random generator algorithms differ across statistical programs and potentially the procedures implemented in different softwares⁵ are not identical.

⁴To save some space, these summary statistics are available upon request

⁵In this work the *R* version 3.5.2 (2018-12-20) – “Eggshell Igloo” was used.

Table 3.1.: The assumed relative utility function parameters

| (a) Mean effects | | (b) Variance-covariance | | |
|---|--------|-------------------------|--------|-------|
| <i>Effects</i> | | <i>Effects</i> | | |
| <i>Means</i> | | Fixed | Random | |
| Characteristics (β) | | Variance | | |
| Sex | 1.420 | Buy | 0 | 3.202 |
| Age | 0.009 | Label | 0 | 2.654 |
| Income | 0.057 | Carbon | 0 | 3.535 |
| Habit | 1.027 | LC | 0 | 2.711 |
| Attributes (γ) | | Covariance | | |
| Price | -1.631 | Buy:Label | 0 | -0.54 |
| Buy | 2.285 | Buy:Carbon | 0 | -4.39 |
| Label | 2.824 | Buy:LC | 0 | 6.17 |
| Carbon | 6.665 | Label:Carbon | 0 | 8.77 |
| LC | -2.785 | Label:LC | 0 | -2.33 |
| | | Carbon:LC | 0 | -4.82 |

Table 3.2.: Individuals' descriptive statistics by dataset

| | Fixed Effects | Random Effects | Target | p value |
|---------------|---------------|----------------|----------|---------|
| Sex | | | | 0.851 |
| Mean | 0.506 | 0.515 | 0.490 | |
| SD | (0.500) | (0.500) | (0.502) | |
| Range | 0 - 1 | 0 - 1 | 0 - 1 | |
| Habit | | | | 0.182 |
| Mean | 0.683 | 0.657 | 0.604 | |
| SD | (0.466) | (0.475) | (0.492) | |
| Range | 0 - 1 | 0 - 1 | 0 - 1 | |
| Income | | | | < 0.001 |
| Mean | 2.750 | 2.671 | 2.147 | |
| SD | (1.476) | (1.438) | (1.222) | |
| Range | 1 - 6 | 1 - 6 | 1 - 6 | |
| Age | | | | 0.255 |
| Mean | 41.862 | 42.161 | 39.755 | |
| SD | (13.685) | (13.820) | (18.895) | |
| Range | 18 - 84 | 18 - 84 | 18 - 85 | |

Table 3.3.: Alternatives' descriptive statistics by dataset

| | Fixed Effects | Random Effects | Target | p value |
|---------------|---------------|----------------|-----------|---------|
| Price | | | | 0.002 |
| Mean | 2.936 | 2.936 | 3.005 | |
| SD | (0.958) | (0.958) | (0.887) | |
| Range | 1.5 - 4.5 | 1.5 - 4.5 | 1.5 - 4.5 | |
| Carbon | | | | 0.999 |
| Mean | 0.5 | 0.5 | 0.5 | |
| SD | (0.5) | (0.5) | (0.5) | |
| Range | 0 - 1 | 0 - 1 | 0 - 1 | |
| Label | | | | 0.999 |
| Mean | 0.5 | 0.5 | 0.5 | |
| SD | (0.5) | (0.5) | (0.5) | |
| Range | 0 - 1 | 0 - 1 | 0 - 1 | |

Differences in the *Choice* proportions appears interestingly. There is an important work in comparing the statistics for different classes in our sample to ensure that they are not biased in favour of label “A” or label “B”, as in this case, the estimates are prone to be biased. For the artificial dataset the ratio of choices per “Buy” alternative is higher than 40% and reaches 47.3% for the fixed effect utility. At the same time, for the random effects specification, the numbers are lower, reaching only 42% in mean for two classes. This particular observation is rather interesting as it demonstrates how the heterogeneous tastes for alternatives' characteristics affect the consumer decisions.

3.2.3.2. Estimation results

The comparison of the utility function coefficient estimates obtained by the different models over different datasets can be done in several steps. First of all, we are interested in the observed mean effects over the datasets, because the possibility to correctly identify the means for the coefficients is of outmost importance for the analysis, regardless of the assumption about homogeneity or heterogeneity of these effects. Then we explore the additional dimension provided by the MMNL estimates, which comprises the estimates for the variance-covariance matrix of the correlated random effects. Finally, we will give some comments on the CNN model estimates.

In the case of homogeneous preferences structure the MNL model obtains the exact estimates with fast a convergence rate and relative simplicity of the problem (table 3.4). The estimates obtained with the MMNL model for the fixed effects dataset demonstrate quasi-identical estimates to the MNL model. The only disadvantage of the MMNL models misspecification in this case resides in the significantly increased estimation time, which requires significantly more iterations in order to estimate correctly the variance-covariance matrix elements and, consequently, the estimation complexity.

In the case of heterogeneous preferences as estimates are significantly biased for the MNL model (table 3.4). The MNL model tends to significantly underestimate the effects of all of the characteristics and attributes for the choice situation. This can potentially lead to a significant bias in case we were using incorrect model specification during a field experiment data exploration. The estimates obtained with the MMNL model are slightly biased as well in this case.

Even as the estimates of the means obtained with MMNL in the presence of the random effects are

Table 3.4.: Estimation results
[!htbp]

| | <i>Fixed effects</i> | | | <i>Random effects</i> | | | <i>Target</i> |
|------------------------|----------------------|----------------------|--------|-----------------------|----------------------|--------|---------------|
| | MNL | MMNL | CNN | MNL | MMNL | CNN | |
| Characteristics | | | | | | | |
| Sex | 1.401*** (0.031) | 1.400*** (0.031) | 1.369 | 0.712*** (0.016) | 1.297*** (0.024) | 0.719 | 1.420 |
| Age | 0.009*** (0.001) | 0.009*** (0.001) | 0.010 | 0.007*** (0.001) | 0.010*** (0.001) | 0.005 | 0.009 |
| Salary | 0.048*** (0.010) | 0.048*** (0.010) | 0.060 | 0.066*** (0.005) | 0.120*** (0.008) | 0.062 | 0.057 |
| Habit | 1.070*** (0.030) | 1.071*** (0.030) | 1.056 | 0.361*** (0.016) | 0.641*** (0.024) | 0.343 | 1.027 |
| Attributes | | | | | | | |
| Price | -1.626*** (0.010) | -1.628*** (0.010) | -1.618 | -0.886*** (0.006) | -1.586*** (0.010) | -0.886 | -1.631 |
| Buy | 2.311*** (0.065) | 2.313*** (0.066) | 2.228 | 0.662*** (0.036) | 2.180*** (0.054) | 0.665 | 2.285 |
| Label | 2.815*** (0.022) | 2.817*** (0.022) | 2.810 | 1.279*** (0.015) | 1.922*** (0.023) | 1.277 | 2.824 |
| Carbon | 6.654*** (0.032) | 6.662*** (0.033) | 6.634 | 3.259*** (0.016) | 5.430*** (0.030) | 3.250 | 6.665 |
| LC | -2.781*** (0.028) | -2.782*** (0.028) | -2.765 | -1.546*** (0.019) | -2.663*** (0.030) | -1.558 | -2.785 |

Note:

* p<0.1; ** p<0.05; *** p<0.01

close to the theoretical ones, the estimates of the variance-covariance matrix elements are rather close, but not perfectly measured. This situation demonstrates the existing trade-off between the need to correctly specify the model from the start and the potential computation inconveniences in the case of implementation of a more complex model under uncertainty. In other words, there is always a choice either to simply use more complex model, which requires more data, calculation time and resources, or to perform an extensive theoretical study beforehand in order to correctly specify and delimit the model from the start.

Our CNN model is identical in structure to the MNL model, estimated with *Adam* algorithm. Because of the nature of the constructed CNN model, the obtained estimates in the presence of fixed effects are technically identical to the estimates obtained with the MNL model. These results demonstrate the flexibility of the NN models and the hypothetical possibility to implement them in place of traditional econometric models with only inconvenience being the relative complexity to obtain the variances for the weights estimates, as no method known to us allows this, or to estimate variances through a cross-validated training of the NN. In the presence of random effects, the proposed CNN algorithm is, identically to MNL model, unable to correctly identify parameters and consequently derive the true means for the underlying coefficients of the relative utility function in the presence of heterogeneous preferences among individuals.

3.2.3.3. Performance comparison

Performance in terms of utility function estimation was presented in the previous section. Three complementary performance metrics are described: (1) the overall fit quality, (2) the computational efficiency and (3) the economic indicator precision.

First of all we focus our attention on the general performance metrics, describing how well the estimated models fit the predicted outcomes over an original dataset. We can observe the values of accuracy and KL divergence, describing overall performance of a given model, in Table 3.5. The table gathers the metrics' values for all the estimated models over both datasets. We observe quite natural situation: the best model in terms of overall performance is the model which is based on the choice rule used in the data generation step. The MNL and MMNL models perform equally well on the fixed effects dataset. This fact supports our initial hypothesis that an implementation of a more complex model is preferred when the real effects are unknown to the researcher. CNN model did not outperform the MNL. This observation may be explained by the data-generation set-up, where the generative algorithm favoured the MNL model with Gumbel error term rather than more general NN framework.

Table 3.5 presents the resources efficiency indicator: CPU time spent for execution by the system on behalf of the calling process. The more advanced *Adam* algorithm implementation with *Keras* easily outperforms the algorithms available in the *mlogit* package, although this boost in efficiency goes at the cost of lower overall performance and goodness of fit. At the same time, the MMNL implementation is far less efficient and takes 128 times more time, than CNN model. This situation clearly illustrates us how the precision and flexibility come at higher costs.

Finally we focus on the case specific metrics, WTP and premiums estimates present in the Table 3.9, that the consumers are eager to pay for particular environmental attributes. Comparing the estimates with the input values, we notice that the variances of the WTP and Premiums estimates (presented in brackets), estimated over a fixed effects dataset, do not potentially affect the conclusion one can

Table 3.5.: General performance measures

| | MNL | MMNL | CNN |
|-----------------|--------|----------|--------|
| Accuracy | | | |
| FE | 0.863 | 0.863 | 0.723 |
| RE | 0.725 | 0.863 | 0.721 |
| KL | | | |
| FE | 0.623 | 0.623 | 0.328 |
| RE | 0.349 | 0.625 | 0.317 |
| Time | | | |
| FE | 20.910 | 452.414 | 17.433 |
| RE | 18.722 | 2066.934 | 16.806 |

Note: FE - fixed; RE - random effects

Table 3.6.: Performance in terms of WTP and premiums

| | <i>Fixed effects</i> | | | <i>Random effects</i> | | | <i>Target</i> |
|---------------|----------------------|---------|-------|-----------------------|---------|-------|---------------|
| | MNL | MMNL | CNN | MNL | MMNL | CNN | |
| WTP | | | | | | | |
| Mean | 1.421 | 1.416 | 1.377 | 0.747 | 1.360 | 0.751 | 1.401 |
| SD | | (0.058) | | | (1.887) | | (1.973) |
| Label | | | | | | | |
| Mean | 1.731 | 1.732 | 1.737 | 1.445 | 1.243 | 1.442 | 1.731 |
| SD | | (0.019) | | | (1.667) | | (1.611) |
| Carbon | | | | | | | |
| Mean | 4.091 | 4.097 | 4.101 | 3.679 | 3.467 | 3.669 | 4.086 |
| SD | | (0.103) | | | (2.323) | | (2.134) |
| LC | | | | | | | |
| Mean | 4.112 | 4.116 | 4.129 | 3.378 | 3.036 | 3.352 | 4.110 |
| SD | | (0.098) | | | (3.240) | | (3.379) |

derive from the results. We may conclude that given sufficiently large dataset the implementation of a more complex model (MMNL in this particular case) is preferable, because it will allow to control for unknown parameters without adding a risk of obtaining biased results. The simpler models, should be preferred in a more restricted context. They empower us to obtain valid results only in the case of correct theoretical assumptions, biasing the estimates in other conditions. Consequently, in the presence of uncertainty about the presence of heterogeneity in the customer choice modelling questions there is a strong interest to implement a more complex model, readjusting it afterwards if needed.

3.2.4. Conclusion

In this work we have introduced the reader to the problematic of the different modelling paradigms in application to the consumer choice studies. By means of an experimental theory-testing framework we demonstrate the complexity of the model performance evaluation problematic, showing the eventual bottlenecks and the questions to be answered on all the levels of data exploration procedure. The correct specification of the theoretical assumptions, the dataset generation, the model choice as well

as the performance measure choice were studied.

Given the experimental design and selected parameters, the MMNL model proves itself to be preferable. The ability to correctly estimate the target effects in presence of preference structure uncertainty is of great value in the field experiments. The CNN model illustrates the possibility for economists to implement the advanced ML techniques to treat economic questions.

One limitation of this work concerns the external validity of the observations. Arbitrary choices made in the study limit our conclusions to this specific case, and require more extensive experimentations to produce more general conclusions. Metaparameters of the framework will allow to specify sample size and compared tools. The presented results are conditioned with (1) the large sample size leading to highly significant estimates of MNL and MMNL and (2) the CNN design aiming to reproduce the MNL, including its limitations. This work demonstrates only a fraction of the full potential of the theory-testing framework. Many extensions and generalisations should be performed before it could be used at scale. For example, it is particularly interesting to introduce an extension which will provide the possibility to explore and compare how different behavioural theories affect the estimation results. The framework could be complemented with a methodological tool-set for hypothesis testing using the advantages of a controlled experiment data collection as well.

3.2.5. Discussion

This paper focuses on the comparison of ML and econometrics approaches in modelling consumer preferences, with a specific emphasis on the impact of theoretical assumptions. The comparison is made on how the different modelling techniques behave in presence and absence of preference heterogeneity within the explored population. Traditionally, economists have relied on econometric tools to derive WTP and understand consumer behaviour, while ML techniques, have gained traction for their predictive capabilities. The fundamental distinction lies in econometrics' focus on understanding underlying data properties, whereas ML prioritizes prediction. The paper critically examines the performance of these methodologies through a theory-testing framework, in the context of a simulated choice experiment.

The work represents the starting point of this PhD work, giving out the first elements of interdisciplinary studies complexity. However, it bears some misconceptions on the differences among the disciplines. While the first outlines of issues in performance assessment emerge in the discussion, the work remains rather optimistic on the simplicity and feasibility of interdisciplinary model performance comparison.

The initial elements of the performance comparison framework could be observed, although the focus is made on the model, rather than on the scientific procedure as a whole. The accent is not made on the framework, as the schema presented in Figure 3.1 should only be perceived as a rough approximation of the performance comparison task in a controlled environment. What is more, the structure presented in the paper remains incomplete even in comparison with the framework of Williams and Ortuzar (1982), as the authors remained ignorant of this literature at the moment.

In the context of the final framework's version this article should be perceived as a methodological work, staged in a controlled environment (Figure 3.2). The methodology involved generating artificial datasets with *heterogeneous* (H) and *homogeneous* (non-H) preference structures to observe the effects

of taste heterogeneity on model performances. At the time the decision to use datasets of large size, including 160000 observations, was dictated by the decision not to offer a competitive advantage to the econometric models, which could easier outperform the ML methods in small samples.

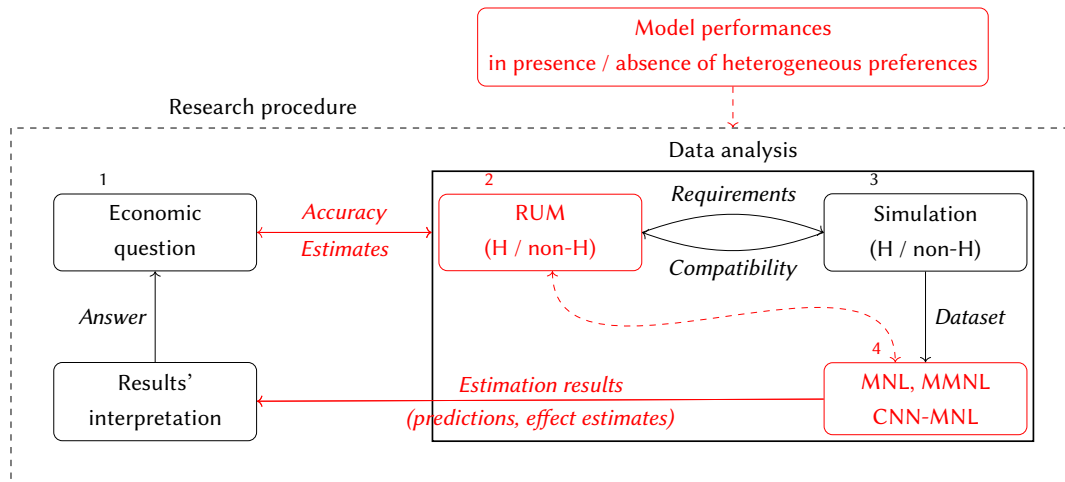


Figure 3.2.: Positioning of DA2PL 2020 paper

The performance comparison part of the work explores overall fit quality, computational efficiency, and economic indicator precision. Those ideas are further developed in the Sections 1.4.2 and 2.2.2 of this thesis. All of the above elements are outputs of the data analysis stage of the scientific procedure, *estimation results*. The exact economic question is not defined in this study, as at the time the understanding of importance of the scientific procedure context in the analysis was not yet acquired. Consequently only the fit metrics and estimates are explored, without providing any particular link to the underlying economic implications. Those errors would be corrected in the following works.

The work reveals that model choice depends on the researcher's assumptions about the true underlying data-generating process. The obtained results were quite expected and did not offer any sufficient breakthrough related to the subject, aligning with the general literature. In the case of homogeneous preferences, simpler models like MNL performed relatively well, but in the presence of heterogeneity, MMNL was winning, highlighting the trade-off between model complexity and estimation accuracy. The CNN-MNL model version should be interpreted more like a proof of concept in the context of this publication, because according to the experimental setting this model was bound to lose. It demonstrates the feasibility of integrating advanced ML techniques into economic inquiries, completing the existing literature on the subject. The controlled simulation setting, while offering advantage in the visibility of targets, favoured the classic DCM models. The paper acknowledges those limitations, emphasizing the need for further experimentation and extensions to enhance the framework's generalizability and practical utility.

Nevertheless, among the obtained insights one was relatively new to the literature, as it was rarely put in evidence by other authors. It appears that in model selection under uncertainty about the underlying theoretical assumptions on the individual profiles within population the preferences should be given to the more complex models. Such approach to model selection reduces the potential biases introduced by the incorrect model specification in more complex cases, equally producing reliable estimates in the more simple settings. Obviously this observation should be taken with a grain of salt, as this recommendation is always subject to the resource availability for more complex models estimation.

3.3. Case 2: Dataset acquisition

ITEA 2023 paper: “*Willingness to pay in commute mode choice: Model performance comparison under sample size and balance impacts*”

Abstract

In economics studies one of the wide-spread target metrics is the *Willingness to Pay (WTP)* of individuals for particular attributes of transportation mode choices. There already exists a vast literature addressing some major issues of the WTP elicitation task. We propose a performance comparison framework, allowing to systematize the previous research. With its help, in this work we explore models perform in WTP elicitation task under potential misspecifications, sample size and dataset balance changes. The ‘swissmetro’ dataset is used for application purposes. We use simulation to vary sample size and configuration, which are used for model estimation and WTP elicitation. The results illustrate the variability in WTP estimates under different configurations.

3.3.1. Research question

In economics studies one of the wide-spread target metrics is the *Willingness to Pay (WTP)* of individuals for particular attributes of goods or services. In transportation studies popular manifestation of WTP are *Value of Time (VOT)* or *Value of Comfort (VOC)*. The WTP elicitation lies at the heart of various tasks in the transportation mode choice analysis: adoption of sustainable transportation modes (Ilahi et al. 2021), perception of the resilient shared transportation modes (Ardeshiri, Safarighouzdi, and Rashidi 2021), consumer preferences for delivery services (Merkert, Bliemer, and Fayyaz 2022), attitudes towards trip attributes (Boto-García et al. 2022).

There exist multiple ways to deduce WTP from the data, most of which rely on the *Random Utility Maximisation (RUM)* framework (McFadden 1974). The obtained results are affected not only by the selected methodology, but by the modelling strategy as well. With time the number of available models and estimation techniques increases, many of them remaining *RUM-compliant*. Following McFadden (1981) the RUM-compliance translates in the independence of the ranking of the choice probabilities of the alternatives by any monotonically increasing transformation of the utility functions of all elemental alternatives (David A. Hensher and Greene 2002). One can also observe a growing number of papers focusing on interpretable *Machine Learning (ML)* techniques in application to choice modelling analysis (Han et al. 2022; Aboutaleb et al. 2021; S. Wang, Wang, and Zhao 2020), some of which address the WTP elicitation (Bergtold and Ramsey 2015; S. Wang, Wang, and Zhao 2020). The multitude of available models and techniques makes it sometimes difficult for the researcher to select the best modelling approach for the particular situation.

There exists several studies in the literature which are dedicated to the exploration of sample size and dataset balance effects for various choice modelling techniques (Huber and Zwerina 1996; Burgess and Street 2006; Rose and Bliemer 2013), exploring model performances in general (Zeng, Zhong, and Hunt 2018; Jong et al. 2019). The majority of researchers focus on the predictive accuracy as the main performance metrics for their sample size requirements calculation. However, according to the interdisciplinary works (Japkowicz and Shah 2011) the performance of competing models may be assessed over several criteria: (1) quality of data adjustments; (2) predictive capacity; (3) quality of the field specific (ex: economic and behavioural) indicators derived from estimates; and (4) algorithmic efficiency

and computational costs. We attempt to complete previous findings with a more extended view on the derived metrics, WTP in particular. **How the various models perform in WTP elicitation task under potential misspecifications? Does the sample size and class balance impact the WTP estimates in various RUM-compliant models?**

3.3.2. Methodology and context

Many of the listed above studies focusing on the WTP metrics rely on *Stated Preference* (SP) data. However, while setting up a DCE little is known about the exact behaviour within the target population. The researchers typically rely on the previous studies in selecting the most plausible theoretical assumptions while conducting a DCE, but there are always some limitations. One of the important elements in the WTP elicitation tasks is tied to the model requirements in terms of sample size and overall dataset configuration. This pushes us to explore empirically the potential consequences of inadequate model usage under changes in data.

Some may say that the research questions were already addressed in the literature and they will not be wrong. There is a number of studies, which in one way or another proposed some insight into the data requirements for particular model families, or explored the data quality impacts on the estimates.

Among the reference works we may encounter, a revision of WTP elicitation approaches performed by Daly, Hess, and Ortúzar (2022). Or a criticised estimation of WTP under utility specification restrictions of Carson and Czajkowski (2019). Paper of Bazzani, Palma, and Nayga (2018) addressing the usage of flexible mixing distributions in WTP space. As well as a rather complete comparison of confidence intervals measures for WTP under sample size changes published by Hole (2007). Among the data focused studies we encounter the mitigation of class balance effects for NL models by M. Bierlaire, Bolduc, and McFadden (2008). The study if impacts of sample size, attribute variance and choice distribution on the accuracy in the paper of Zeng, Zhong, and Hunt (2018). An extensive analysis of ample size requirements for stated choice experiments of Rose and Bliemer (2013).

All of the above works are relatively close to the research questions we have outlined in the introduction. However, as most of the research is focused on the theoretical fundamentals with scarce empirical illustrations, we attempt to complement the existing literature with a more accessible evidence. For this purpose we propose a theoretical performance comparison framework, which should simplify the empirical theory testing procedure.

In this section we are offering a short focus on the WTP elicitation approach, which will be used further on. Then we outline the proposed performance comparison framework that will guide our data-driven study.

3.3.2.1. Willingness to Pay

For the purposes of this study we use the simplest WTP definition. We assume that individual deterministic utility of an alternative j (from a set of available alternatives Ω) is given as function with parameters β_j : $V_j = f(\beta_j)$. The simplest option is then to provide the point analytical estimates of the WTP values, which is justified if V_j is linear in attributes. The total variation of V_j with respect to joint variations in the k -th attribute $x_{k,j}$ and the cost attribute $x_{cost,j}$ is $\Delta V_j = \Delta x_{k,j} + \Delta x_{cost,j}$. Resolving this

equation for the case of $\Delta V_j = 0$ we obtain the change in cost, which keeps the deterministic utility unchanged given a change in k -th attribute:

$$WTP_{k,j} = \frac{\Delta V_j / \Delta x_{k,j}}{\Delta V_j / \Delta x_{cost,j}}$$

The easiest option focuses on confidence interval calculation for WTP values using the less resource heavy **Delta method** (Daly, Hess, and Ortúzar 2022), which avoids simulation step (Scaccia, Marcucci, and Gatta 2023). Such method is usually used to calculate the standard error for a function of the parameter estimates. For simplicity in this study we do not use any alternative WTP confidence intervals identification strategies. This methods add some more prerequisites, as we should assume that WTP_k is given as $\omega_k = h(\beta_k, \beta_{cost}) = \frac{\beta_k}{\beta_{cost}}$ is a differentiable function. The formula for the standard error of ω_k is hence (Daly, Hess, and Train 2012) is:

$$\sigma_{\hat{\omega}_k} = \frac{1}{\beta_{cost}} \sqrt{\sigma_{\hat{\beta}_k}^2 + 2\omega_k \sigma_{\hat{\beta}_k \hat{\beta}_{cost}} + \omega_k^2 \sigma_{\hat{\beta}_{cost}}^2}$$

3.3.2.2. Performance comparison framework

To address the model misspecification and data imbalance issues under a new angle we propose a performance comparison framework. It incorporates all essential steps from the research question definition to the performance comparison in relation to the given context. This framework is based on the concepts described by Williams and Ortuzar (1982), revised and extended.

We believe that the most rational way to construct such framework is to mimic in its structure the traditional scientific **research procedure**. In the literature, regardless of the actual case, all the research takes its root in some problematic: a question to be answered, a barrier to be overcome. Once the task delimited, there are different strategies on how to proceed. Some of them are conventional and described in every practical guide (Wooldridge 2012; Baltagi 2008), while other are more obscure and are sometimes criticized for uncommon practices (Daly, Hess, and Ortúzar 2022). As one can see, those topics we'll rise here are mainly discussed in the epistemology works, rather than in more abundant applied studies. Nevertheless, it's extremely important to have the general understanding of the typical procedures and paths implemented in applied research to make the next leap towards framework construction.

The procedure may be in general divided into several major steps (Figure 3.3). First of all, every research starts with a *problematic* identification and *operational* or *economic question* definition⁶. Every study begins with a particular need - *operational problematic* to be addressed. The first step reflect the transition of the real world problem to be treated into the more restricted context of a research specific question. The next stage in the research requires the researcher to make some assumptions about the nature of the data and the underlying processes. Typically it's during this stage that hypothetical interaction model is defined based on the theoretical assumptions or the preliminary analysis of the available (if available) data. Thus the second step is a further extension of the *problematic* narrowing and translation into numerical terms: target *metrics* identification. Those *metrics* should

⁶Here we avoid speaking about *research question*, as sometimes it may not be directly linked with the *economic question* treated in the study. Moreover, the *question* may be purely *operational*, without production of any particular new knowledge and be purely context specific for particular application.

allow the researcher to answer to the research question. For example, one may be interested in causality exploration, which may be translated into the analysis of particular coefficient significance in an econometric model. Another example is the prediction task: researchers may be interested to offer the best prediction of consumer behaviour (ex: to identify the market shares), which may be translated into comparison of various performance metrics for different predictive models. Once the target defined, the research may proceed differently, depending on the available information. Without loss of generality this step may be summarized as *data collection and analysis* process. Either the actors already have access to some data and build the model using available information. Or the model is prebuilt and drives the data collection step. Finally, the data analysis provides the actor with information on the target *metrics (estimates)*. Those allow to answer the initial question and offer a *solution* to the initial problematic.

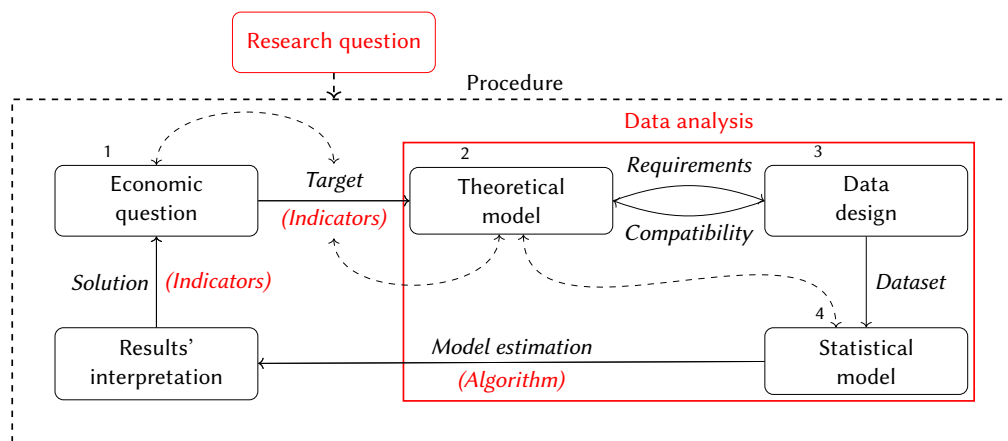


Figure 3.3.: Proposed performance comparison framework

The performance of a model can scarcely be assessed without any particular context. In fact we redefine the **performance** as **the model's capacity to bring a correct answer in the context of explored problematic**. This performance term redefinition brings us to the necessity to step aside from the typical *model performance comparison* and switches our focus to another conceptual unit - *the procedure*. By *the procedure* we understand the entire process starting with the research question definition to the answer to this question. This means that the procedure in this case includes such steps as data collection, processing and analysis. This also includes all the eventual (be it arbitrary or not) choice in terms of model configuration, selection and fine-tuning.

Consequently, the framework should be inevitably dependent on the research question: some models are simply not capable to answer some questions or there are no known or established practices of their usage. The definition of the research question should therefore be considered as the first step in the proposed framework. It will provide the researcher with particular metrics to consider while performing the model comparison.

The second stage in the framework should be left for the dataset choice or dataset generation procedure. This includes all the potential assumptions and *a priori* choices on the assumed individual behaviour within population, external effects and potential biases. At this point the opinions may vary as discussed by Japkowicz and Shah (2011). Even though in statistical modelling when speaking about model performance assessment and comparison the focus is typically made on the classification (prediction) accuracy (Andersson, Davidsson, and Lindén 1999; Hand 2012; Askin and Gokalp 2013) this is not always the best option. On the one hand, in model comparison, whatever is the research

question, one will always have some target metrics or criteria in mind. It means, that for a complete comparison procedure one should be able to compare not only the models between themselves, but compare the results with some externally defined target as well.

The next stage is represented by the *modelling procedure* itself. This includes the choice of the model and its implementation, the configuration of the estimated utility functions, etc. Later it will be equally subject to the numerical specificity: the choice of the estimation algorithm and its implementation, the particular code base and approach to the problem solving.

Finally, comes the *post-treatment* of the obtained estimates. In our particular case this step involves the WTP calculation, assuming the model was not estimated directly in the WTP space. All the essential indicators obtained on this step should be evaluated in the context of the research question and, if possible, compared to the target values used as inputs for the simulation task. Now, once the framework is fully described we can proceed with the application.

3.3.3. Application

For illustration purposes we use the rather popular and publicly available dataset *swissmetro*. The dataset firstly appeared in the paper of M. Bierlaire, Axhausen, and Abay (2001), where it was used to assess the acceptance of the state proposed modal innovation (A. Nash et al. 2007). A more in-depth description of the dataset, as well as the dataset itself are available on the *biogeme* project website. This data was used in many illustration of newly created model capabilities, as well as in several model performance comparison tasks. The most closely related works to our usecase are: M. Bierlaire, Axhausen, and Abay (2001), M. Bierlaire, Bolduc, and McFadden (2008) and Newman, Ferguson, and Garrow (2013).

We rely on preceding works to construct artificial datasets of different sample sizes and configurations. The conventional *Nested Logit* (NL) structure is imposed, which reflects quite common in reality decision rule structure. Several models are then estimated over the resulting datasets. The tests for WTP estimates validity are then performed, through a comparison with expected target results, as well as their overall significance.

3.3.3.1. Dataset description

The original dataset (as presented by M. Bierlaire, Axhausen, and Abay (2001)) is based on a combination of the *revealed preferences* (RP) and *stated preferences* (SP) data collected in Switzerland, during March 1998. At the first stage, the study relied on collection of the initial information (observation) of the trip performed by subject. This step was followed by a SP data collection step where they were proposed a novel *hypothetical* alternative: the *swissmetro*. To ensure that the new *hypothetical* transportation mode was pertinent for the subjects the sampling was performed through approaching the subjects while they travelled on the target routes. 470 observations (435 suitable ones) were collected in the train between St. Gallen and Geneva. Another 770 usable SP surveys were collected among the car user, this part being performed by mail with the support of central Swiss car licence agency. In the SP part of the study authors used *fractional factorial design* offering the following set of alternatives: (1) rail (TRAIN), (2) *swissmetro* (SM) and (3) car (CAR, only for car owners). All the alternatives were designed by *travel time*, *fare/cost* and *headway* (for rail based alternatives only).

For this study we adopt the approach described by M. Bierlaire, Bolduc, and McFadden (2008) and later used by Newman, Ferguson, and Garrow (2013). The original dataset will be used for simulation purposes, which allows us to observe the model performances in a more controlled environment. Prior to simulation the dataset is filtered, excluding the observations for which there is no choice made and limiting our attention to the commute and business purpose trips. The descriptive statistics for the resulting dataset are presented in the table 3.7 (only reused explicative variables are shown).

Table 3.7.: Descriptive statistics

| Variable | N | Mean | St. Dev. | Min | Max |
|--------------------|-------|---------|-----------|-----|-------|
| Cost | | | | | |
| TRAIN_CO | 6,768 | 490.885 | 1,062.594 | 9 | 5,040 |
| CAR_CO | 6,768 | 78.656 | 55.922 | 0 | 520 |
| SM_CO | 6,768 | 641.066 | 1,411.658 | 11 | 6,720 |
| Travel time | | | | | |
| TRAIN_TT | 6,768 | 166.077 | 69.796 | 35 | 1,022 |
| CAR_TT | 6,768 | 123.155 | 91.718 | 0 | 1,560 |
| SM_TT | 6,768 | 84.507 | 47.113 | 12 | 796 |

3.3.3.2. Simulation

We proceed with a simulated dataset, which is based on the original one. The simulation approach adopted is identical to the one performed by M. Bierlaire, Bolduc, and McFadden (2008). Each observation is replicated 100 times to provide us with synthetic observations. The alternative attributes values were overwritten by draws from normal distribution $N(\lambda, \sigma^2)$, where λ is the value of the corresponding attribute in the original dataset, and $\sigma = 0.05\lambda$ (M. Bierlaire, Bolduc, and McFadden 2008).

Speaking about the decision rules, we decide to adopt the identical nested logit structure as in the other studies (M. Bierlaire, Bolduc, and McFadden 2008; M. Bierlaire, Axhausen, and Abay 2001). The choice model specification is given in the Table 3.8.

Table 3.8.: Utility specification

| Utility | Value | TRAIN | SM | CAR |
|-----------------------|---------|-------|------|------|
| Parameter | | | | |
| ASC_{CAR} | -0.1880 | 0 | 0 | 1 |
| ASC_{SM} | 0.1470 | 0 | 1 | 0 |
| β_{TRAIN_TIME} | -0.0107 | TT | 0 | 0 |
| β_{SM_TIME} | -0.0081 | 0 | TT | 0 |
| β_{CAR_TIME} | -0.0071 | 0 | 0 | TT |
| β_{COST} | -0.0083 | COST | COST | COST |
| Nests | | | | |
| $\lambda_{EXISTING}$ | 0.4405 | 1 | 0 | 1 |
| λ_{FUTURE} | 1.0000 | 0 | 1 | 0 |

Nesting structure was introduced through error components following the specification provided by M. Bierlaire, Bolduc, and McFadden (2008)⁷. This structure assumed that alternatives can be separated according to their real availability. Meaning that while error components behave identically for existing transportation modes (car and train), the effects may differ for non-existing (*future*) alternative.

The WTP (VOT in this particular case) true values can be calculated as $\omega_k = \frac{\beta_k}{\beta_{cost}}$ (ex. for TRAIN alternative we would calculate $WTP_{TRAIN_TIME} = \frac{\beta_{TRAIN_TIME}}{\beta_{COST}}$). This computation is justified because we assume, in our simulation, that effects are fixed within population. This gives us the values as presented in Table 3.9.

Table 3.9.: True WTP (VOT) values

| WTP_{CAR_TIME} | WTP_{SM_TIME} | WTP_{TRAIN_TIME} |
|-------------------|------------------|---------------------|
| 0.8554217 | 0.9759036 | 1.289157 |

The final step includes drawing random observations from the resulting database to compose individual datasets of desired size and class-distribution. We vary the sample size from 500 observations, a number quite often encountered in econometric studies, to 10000 observations⁸, which approaches the frontier of the datasets available for some very simple ML tasks. The different configurations are tested for all possible combinations of classes with a step of 0.2⁹, as well as the perfectly balanced class distribution with equally distributed observations. For each pair of sample size and configuration parameters we randomly draw 50 datasets and estimate selected model over them.

This approach to simulation allows us not only to obtain a consistent baseline for performance assessment, but also the possibility to compare our results with similar papers, where identical simulation strategy was implemented.

3.3.3.3. Estimation

For the purposes of this study we implement three closely related econometric models, which might be potentially used by novices in choice modelling. Among them: (1) the optimal NL model, (2) the misspecified Multinomial Logit (MNL) model and (3) the Mixed MNL (MMNL) model, which still allows to capture non-uniform error structure. For easier results interpretation we use scaling during the estimation step for all the models.

The NL model follows the specification used during the simulation step and is expected to perform the best on the available data. The MNL model differs from it only by the absence of the nests (Table 3.10), meaning the nesting parameter α is omitted.

⁷For this purpose we used the `evd::rmvevd()` function in R

⁸Those values may vary by ± 1 for the datasets with balanced shares.

⁹This results in a plain defined as $SHARE_TRAIN + SHARE_SM + SHARE_CAR = 1$.

Table 3.10.: Utility specification for MNL model

| Utility | TRAIN | SM | CAR |
|-----------------------|-------|------|------|
| Parameter | | | |
| ASC_{CAR} | 0 | 0 | 1 |
| ASC_{SM} | 0 | 1 | 0 |
| β_{TRAIN_TIME} | TT | 0 | 0 |
| β_{SM_TIME} | 0 | TT | 0 |
| β_{CAR_TIME} | 0 | 0 | TT |
| β_{COST} | COST | COST | COST |

With a deterministic alternative specific utility given as:

$$V_j = ASC_j + \beta_{TIME,j}x_{TIME,j} + \beta_{COST}x_{COST,j}$$

The MMNL model (Table 3.11) mimics the NL model structure, although it is a theoretically incorrect way to introduce nesting in the model as it was illustrated by Marcela A. Munizaga and Alvarez-Daziano (2001). We introduce the random term with zero mean and variance σ_v for the alternatives within a single nest. This allows to address the differences in errors variances, but also introduces a biased covariance structure to the estimated model.

3.3.3.4. WTP and model performance

As we have previously shown, in the literature there is no known consensus on the performance metrics and the “*model performance*” definition. As our study focuses on the WTP estimates, we assume that the objective of a model can be viewed as correct estimation of the target metrics. The WTP in its turn relies on the correct estimation of the effects within the model, assuming that the functional form is known and true.

Hence we are interested to observe the shares of estimation routines which manage to correctly identify the effects. Here, the term “*correctly estimate*” means a production of human readable results, which are not contradictory with real world (simulated in our case) scenario. To properly analyse this information, we are going to explore two different shares: (1) a share of models reporting estimates significantly different from 0, meaning that in the real world application the researcher would take the estimates into account; and (2) a share of models reporting estimates not significantly different from target values (the true values used for simulation of individual behaviour). One of the main advantages for this approach is that we can use basic *t*-test for hypothesis verification in each of the estimations and report the results in a convenient human readable form.

The same reasoning may be applied to the WTP estimates directly (Daly, Hess, and Ortúzar 2022; Hole 2007). For WTP estimates we set $\alpha/2$ to 0.125 for confidence interval specification, as in the work of M. Bierlaire, Bolduc, and McFadden (2008). The WTP variance estimates are obtained using the *Delta* method, as suggested in the manuscript of Daly, Hess, and Ortúzar (2022).

Table 3.11.: Utility specification for MMNL model

| Utility | TRAIN | SM | CAR |
|-----------------------|-------|------|------|
| Parameter | | | |
| ASC_{CAR} | 0 | 0 | 1 |
| ASC_{SM} | 0 | 1 | 0 |
| β_{TRAIN_TIME} | TT | 0 | 0 |
| β_{SM_TIME} | 0 | TT | 0 |
| β_{CAR_TIME} | 0 | 0 | TT |
| β_{COST} | COST | COST | COST |
| σ_v | 1 | 0 | 1 |

With a deterministic alternative specific utility given as:

$$V_j = ASC_j + \beta_{TIME,j}x_{TIME,j} + \beta_{COST}x_{COST,j} + v_j(0, \sigma_v)$$

Performing similar test in over our simulated dataset estimates results in the following shares (Table 3.12). Here we observe the shares of models in dependence of the sample size. Each entry relies on $10 \times 50 = 500$ estimated models, mixing all available class balance configurations within sample. The WTP estimates are considered as appropriate if the desired condition (test) is fulfilled across all three alternatives, as facing three alternative mode choices makes us compute three distinct WTP values. The results presented in this part are for traditional estimation method, without any transformations (Carson and Czajkowski 2019) nor transitions into the WTP space (K. Train and Weeks 2005). We can observe that the number of estimates different from zero increases with sample size.

However, the same cannot be said about the shares of results not different from analytical targets (Table 3.13). Obviously, the simple *difference from zero* test is not the only one interesting for us. We might be interested with an additional test - the exploration of whether or not the obtained WTP estimates are significantly different from zero. Obviously, it's important that the estimator is unbiased, but from operational point of view it's equally important to obtain a meaningful result, which correctly reflects the reality. Which is extremely important in the context of potential strategic decision making based on the estimated values. We can see that those shares decrease with sample size, which has two potential explanations. Assuming the simulation procedure and random sampling has no apparent flaws, we may imply that such behaviour might be explained by the changes in class balance within the dataset.

Table 3.12.: Shares of WTP estimates not different from target, by sample size.

| Observations | MMNL | MNL | NL |
|--------------|-------|-------|-------|
| 500 | 80.20 | 78.80 | 87.50 |
| 1000 | 80.60 | 72.00 | 86.40 |
| 5000 | 52.60 | 50.20 | 70.80 |
| 10000 | 37.40 | 34.00 | 58.60 |

Table 3.13.: Shares of WTP estimates different from zero, by sample size.

| Observations | MMNL | MNL | NL |
|--------------|--------|--------|-------|
| 500 | 89.80 | 91.00 | 41.68 |
| 1000 | 99.00 | 99.00 | 72.60 |
| 5000 | 100.00 | 100.00 | 95.60 |
| 10000 | 100.00 | 100.00 | 99.40 |

Finally we explore the shares of WTP completing both of the above conditions, as presented in Table 3.14. While WTP estimates non-distinguishable from zero may be discarded by researcher leading to non-concluding results, the biased estimates are not so easy to detect in the field.

Table 3.14.: Shares of all correctly estimated WTP by sample size.

| Observations | MMNL | MNL | NL |
|--------------|-------|-------|-------|
| 500 | 77.20 | 76.20 | 41.08 |
| 1000 | 80.60 | 72.00 | 67.20 |
| 5000 | 52.60 | 50.20 | 70.80 |
| 10000 | 37.40 | 34.00 | 58.60 |

A similar analysis can be applied to the results aggregator by class balance (Table 3.15). In this case each *shares combination* regroups $4 \times 50 = 200$ entries with all available class balances within sample.

Table 3.15.: Shares of all correctly estimated WTP by dataset balance.

| Share TRAIN | Share SM | Share CAR | MMNL | MNL | NL |
|-------------|----------|-----------|-------|-------|-------|
| 0.10 | 0.10 | 0.80 | 71.50 | 69.00 | 83.00 |
| 0.10 | 0.80 | 0.10 | 64.00 | 57.50 | 70.50 |
| 0.20 | 0.20 | 0.60 | 73.00 | 68.50 | 85.00 |
| 0.20 | 0.40 | 0.40 | 67.00 | 56.00 | 85.50 |
| 0.20 | 0.60 | 0.20 | 63.00 | 57.00 | 72.50 |
| 0.33 | 0.33 | 0.33 | 59.00 | 60.50 | 67.00 |
| 0.40 | 0.20 | 0.40 | 59.00 | 57.00 | 60.50 |
| 0.40 | 0.40 | 0.20 | 60.00 | 63.50 | 38.50 |
| 0.60 | 0.20 | 0.20 | 51.00 | 52.00 | 19.19 |
| 0.80 | 0.10 | 0.10 | 52.00 | 40.00 | 12.63 |

3.3.4. Conclusion

In this paper we have empirically explored the effects of the model misspecification and changes in sample size and class balance within dataset on the WTP estimates. This study offers primarily a case dependent evidence, which is intended to raise the awareness of the perverse effects of the modelling strategy choice and data selection in empirical work.

In the particular application we have demonstrated that the increase of the sample size may not always be the best solution. In particular the attention should be paid to the modelling technique implemented and the reliability of the underlying assumptions. Those observations underline the problematic of model performance assessment and toolset selection in the empirical work.

Finally, but not less importantly, we have outlined the baseline of a model performance comparison framework, which can be extended to the other domains. The proposed toolset allows to efficiently contrast the performance impacts of the changes in the research procedure, which is invaluable for the empirical studies. Such toolset may allow to reduce studies' costs and time through prior experimentation.

3.3.5. Discussion

This second conference paper represents a much more mature work. In this case the framework presented in the paper (Figure 3.3), although still not in its complete form, reflects better the concept of orientation towards the scientific procedures. This shift becomes clear by how the problematic is presented, as this time the focus shifts to the economic question, staging the context for the scientific procedure analysis. The study focuses on the challenges associated with deducing WTP from data under restrictions in data availability, considering both traditional DCM models and interpretable ML techniques. The complexity residing still in the number of available models and estimation techniques, which makes it challenging for researchers to choose the most suitable approach for specific situations.

For better visibility of the differences between the two framework versions the work should be positioned according to the more recent version of the framework. In the Figure 3.4 the minor changes become more apparent. Those are mostly due to slight changes in understanding of the relationships

between various elements of the framework, such as, for example, the role of the *target metrics* in the scientific procedure. At the time it was assumed that target metrics might be quasi-separated into a standalone stage, instead of a mediator between the *research question* and *data analysis* stages. What is more, the target metrics were assumed to impact the choice of the theoretical framework, instead of impacting the entirety of the data analysis stage.

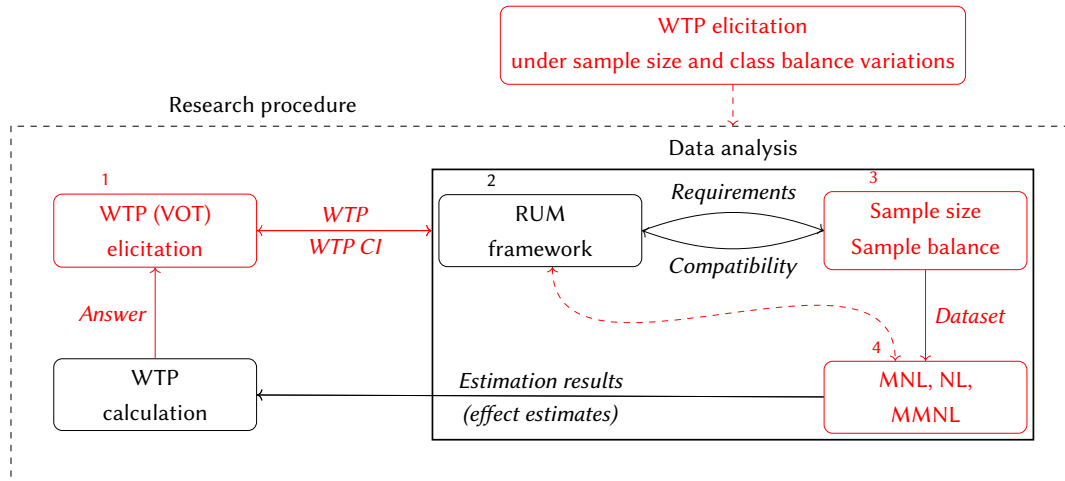


Figure 3.4.: Positioning of ITEA 2023 paper

This research investigates the critical aspect of WTP elicitation in economic studies, particularly in transportation mode choice analysis. Thus, the WTP elicitation task fills in the gap present in the previous study (Section 3.2), completing the framework with an *economic question* and setting *target metrics* at the same time. The methodology involves a comprehensive exploration of sample size and dataset balance effects on WTP estimates for RUM-compliant models. This part reflects the *theoretical question* part in the context of the proposed framework. While existing literature has already extensively explored those topics, this study aims to provide more accessible empirical evidence, introducing the unified performance analysis framework for the first time. To illustrate the application of the framework, the study once again relies on simulated data. The *swissmetro* dataset is used following the existing practices in synthetic data generation available in the literature. The performances of various models are then assessed, among which: MNL, NL, and MMNL. The conclusion emphasizes the empirical exploration of model misspecification effects, changes in sample size, and class balance within datasets on WTP estimates.

However, this work still has several drawbacks. For example, one of the most evident misconceptions is directly linked to the obtained results, which drastically differ from what could be expected. It should be natural that the share of correctly estimated effects increases at least for the optimal model, represented by the NL in this case, with the increase of sample size. However in the Table 3.14 we observe a contradictory situation, which should be explored further. The share of correctly identified WTP values increases at first, but then falls several percentage points back, which could be potentially explained by the low number of samples included into analysis. Another disturbing finding relates to the share of correctly identified WTP elements in dependence from the class balance in dataset. The results observed for NL model might be expected, as presented in the work of M. Bierlaire, Bolduc, and McFadden (2008), although this does not fully explain the observed values. For example, in several cases for MMNL and MNL models the best results are obtained for imbalanced datasets, which points out some potential inconsistency.

3.4. Case 3: Statistical modelling

IATBR 2024¹⁰: “Issues of model selection: insights from a failed experiment”

Abstract

This work in progress addresses the popular issue of econometrics and Machine Learning (ML) models comparison in the context of the commute mode choice modelling. Contrary to the previous known to us works the focus shifts from the presentation of the novel techniques and their comparison to the baseline models, to a systematic performance analysis through a novel framework. This model performance comparison framework is inspired by the work of Williams and Ortuzar (1982) for modelling technique selection purposes. The traditional Discrete Choice Modelling (DCM) techniques are put in competition with the emerging and not yet fully accepted ML and Deep Learning (DL) driven approaches to choice modelling (S. Wang, Wang, and Zhao 2018; L. Cheng et al. 2019). In contrast to the ML oriented literature this task is performed in the context of economic indicators elicitation task, WTP in particular. For application purposes we use the popular ‘swissmetro’ dataset based on which the synthetic samples are generated. The conventional DCM techniques (MNL and NL) are then contrasted with the ML emerging alternatives in the WTP elicitation task.

3.4.1. Introduction

There exist many concurrent ways to perform the mode choice analysis, most of which rely on the *Random Utility Maximisation* (RUM) framework (McFadden 1974). With time the number of available models and estimation techniques grows, but many of them are still *RUM-compliant*, as the roots in economic theory make such models easy to interpret. Following (McFadden 1981) the RUM-compliance translates in the independence of the ranking of the choice probabilities of the alternatives by any monotonically increasing transformation of the utility functions of all elemental alternatives (David A. Hensher and Greene 2002). The family of RUM-compliant models was recently extended by a number of more advanced models, as the mode choice modelling is on the verge of transition to incorporation of novel *Machine Learning* (ML) based analysis methodologies. The transportation studies and mode choice modelling applications in particular were among the first to attempt the implementation of ML methodology for typically *theory driven* tasks. One can observe a growing number of papers focusing on interpretable ML techniques in application to choice modelling analysis (Han et al. 2022; Aboutaleb et al. 2021), some of which address the WTP elicitation (Bergtold and Ramsey 2015; S. Wang, Wang, and Zhao 2020). The multitude of available models and techniques makes it sometimes difficult for the researcher to select the best modelling approach for the particular situation. Nowadays, there seems to be no unanimity on the exact performance comparison procedure to adopt. To the best of our knowledge, there is no work providing a complete and comprehensive methodology for assessing and comparing model performances given the context of mode choice analysis studies. To tackle this issue we propose a performance comparison framework, which incorporates all essential steps of the *data analysis procedure*: from the research question definition to the performance comparison in relation to the given context.

¹⁰A short abstract was submitted for review on 30 September 2023, the answer is due by the end of January. The version presented in this thesis does not represent a scientific product ready for presentation, but rather a collection of intermediary insights and by-products related to an ongoing work.

The proposed framework covers the key elements of the performance comparison task through identification of key points of the data analysis differences in the research procedures. The performance of competing models may be studied according to several different criteria (Japkowicz and Shah 2011; Hand 2012; Athey and Imbens 2015; S. Wang, Wang, and Zhao 2020): (1) in terms of the quality of data adjustments; (2) in terms of predictive capacity; (3) in terms of the quality of the field specific (economic and behavioural in mode choice modelling) indicators derived from estimates; and (4) according to their algorithmic efficiency and computational costs. In the context of the choice mode analysis there exist a number of studies focusing on performance comparison of different models (Han et al. 2022; S. Wang et al. 2021; Belgiawan et al. 2019). However, there is no established procedure for the model performance analysis and testing. While a number of studies (Gonzalez-Valdes, Heydecker, and Ortúzar 2022; Vij and Walker 2016) rely on the framework proposed by Williams and Ortuzar (1982) for model performance comparison, they usually reference the simulation related part of the framework. The authors avoid or provide no explicit commentary on the public policy implications of the original work in the context of their research. The proposed general performance analysis framework covers all of the above elements. Taking inspiration in the work of Williams and Ortuzar (1982) the framework is extended to cover the diversity of potential use-cases, focusing primarily on the *Discrete Choice Modelling* (DCM) context.

This work opens with a revision of approaches to model performance comparison task, underlying the decisions in model selection. The various modelling strategies are discussed, putting in evidence the emerging disparities in the *Discrete Choice Modelling* (DCM) literature. The concurrence between ML and classic theory-driven approach to DCM establish an entry point for the performance comparison framework construction. Both classic and emerging approaches to DCM are briefly described, offering a short literature overview on the topic. Among the classic approaches both MNL and NL are included as baseline models. For the more complex methods the *Alternative Specific Utility Deep Neural Networks* (ASUDNN) and *Convolutional Neural Networks* (CNN) are employed (S. Wang, Wang, and Zhao 2020; Bentz and Merunka 2000). The performance comparison framework presentation completes the first part of the article. A simulation driven application is proposed, allowing to better contrast the differences among the modelling approaches in the context of economics and transportation related research question.

Among the expected results the NN, ASUDNN- in this particular case, should demonstrate better accuracy related to capturing complex, non-linear relationships within data. At the same time classic DCM techniques should be able to better capture the individual effects of the explicative variables, offering a set of interpretable estimates. This interpretability can be valuable in understanding the drivers behind choices. Finally, we might expected to observe faster estimation times for the ML based models, at least in large samples.

3.4.2. Performance comparison and model selection

In the literature the topic of performance comparison arises on many different levels, although only theoretical works present sufficient information on the performance comparison in their publications. What is more, the understanding of model performance differs across different applications. For the purposes of performance comparison we propose a novel framework, constructed under assumption that *model performance analysis* cannot be conducted without taking into account the context of the research procedure. Hence we switch the focus from the basic *model performance comparison* to the

research procedure performance analysis. The framework relies at its core on the concepts described by (Williams and Ortuzar 1982), revised and extended to a more general use-case. The *research procedure* in this case consists of two key elements: (1) the *research question* itself and (2) the *data analysis procedure*. Both being tightly linked together with mutual influences, as the research question dictates the requirements for data and analysis methods, while the available data and toolset impose the limitations on the research question.

Since much of the research concentrates on theoretical foundations with limited empirical illustrations, we aim to contribute to the existing literature with more accessible evidence. To achieve this, we introduce a theoretical performance comparison framework designed to streamline the empirical theory testing process. This section provides a brief overview of the WTP elicitation approach to be employed subsequently, followed by an outline of the proposed performance comparison framework that will serve as a guide for our data-driven study.

3.4.2.1. Model differences

The choice modelling tasks are more and more often approached from the Machine Learning perspective in transportation research. In general, the trends in ML techniques introduction might be separated into several major groups. Among which we encounter: (1) usage of ML models to perform DCM tasks, substituting classic DCM models with ML toolset; (2) implementation of ML learning algorithms for the baseline econometric models; and (3) combination of the ML and classic DCM for the increased accuracy purposes.

First of all there is a number of studies that simply use the ML techniques to treat the typical choice modelling problems, such as demand prediction. Among those works we may focus on several recent works. Omrani et al. (2015) discusses land-use modelling, introducing the multi-label (ML) concept for more accurate representation of mixed land use through the k nearest neighbours (kNN) method. Hagenauer and Helbich (2017) explores travel mode choice analysis using Dutch travel diary data, highlighting the superiority of the random forest classifier and emphasizing variable importance. S. Wang, Wang, and Zhao (2020) examines the interpretability of deep neural networks (DNNs) in choice analysis, indicating their comparable provision of economic information to classical discrete choice models (DCMs). Challenges include small sample sizes, emphasizing the need for robust training methods.

Other studies attempt to implement the ML estimation algorithms in the context of classical choice analysis models. Several studies aim to apply machine learning (ML) estimation algorithms to classical choice analysis models, primarily focusing on enhancing computational efficiency. In the work by Lederrey, Lurkin, and Bierlaire (2018), the discussion revolves around the infrequent mention of optimization algorithms in discrete choice literature, attributing it to the historical success of traditional Newton-Raphson methods. However, with the rise of abundant data in choice situations, computational challenges arise, leading to the introduction of the *Stochastic Newton Method* (SNM) for parameter estimation. Preliminary findings suggest its superiority over stochastic first-order and quasi-Newton methods. Building on this, Lederrey et al. (2021) explore the application of ML models to large datasets in the context of statistical Discrete Choice Models (DCMs). The article introduces efficient stochastic optimization algorithms, incorporating a stochastic Hessian, batch size adjustments, and dynamic algorithm selection based on batch size. Experimental comparisons reveal the HAMABS

algorithm's superior performance, significantly reducing optimization time and offering potential for innovative choice model specification. Integrating these algorithms into DCM estimation software holds promise for expedited model estimation and encourages exploration of new approaches by researchers and practitioners.

Finally there exist studies that attempt to merge the two approaches with addition of some transitional steps. Several studies aim to integrate machine learning (ML) techniques with traditional choice analysis models through transitional steps. Sifringer, Lurkin, and Alahi (2020) address inaccuracies in discrete choice modelling (DCM) by dividing the systematic utility specification into knowledge-driven and data-driven components, enhancing predictive capabilities without compromising interpretability. Their approach introduces Learning Multinomial Logit (L-MNL) and Learning Nested Logit (L-NL) models, outperforming traditional models in predictive performance and parameter estimation accuracy. S. Wang et al. (2021) employs statistical learning theory to examine trade-offs in deep neural networks (DNNs), addressing challenges of overfitting and interpretability. The study establishes a framework to measure estimation and approximation errors, highlighting DNN's superior performance over the binary logit model in both prediction and interpretation. Sfeir et al. (2021) introduce a Latent Class Choice Model (LCCM) utilizing Gaussian-Bernoulli mixture models for latent classes, enhancing out-of-sample prediction accuracy and heterogeneity representation. Han et al. (2022) tackle utility misspecification in discrete choice models by incorporating a neural network (TasteNet) to learn taste representation. The TasteNet-MNL model demonstrates predictability and accuracy in recovering taste functions, outperforming benchmark models on a *swissmetro* dataset. This last category may be also divided into several subcategories, as there is no unique way to merge the tool-sets of two rather differently inclined scientific communities. Most of the methods implement some adjustment to the existing ML techniques to comply with the classical choice theory and thus be *explicative* at least in some degree.

In this work we will illustrate the differences between those approaches with the assistance of a performance comparison framework. The case study will allow us to demonstrate the performance differences across several dimensions, serving at the same time to accentuate the differences in the proposed toolset functionality.

For the purposes of this study we are going to introduce a relatively simple example for each of the above categories of modelling approaches. In the first place, we are going to use the state of art MNL and NL models, representing the DCM classic toolset. The same models will be implemented within the TensorFlow framework, representing the case of ML issued learning algorithms implementation. In particular, the usage of Adam algorithm will be put in place for treatment of this task. The third category regroups those models will be extended with a Deep Neural Network stages prior to computation of alternative specific utilities. This corresponds to ASUDNN (S. Wang, Mo, and Zhao 2020) -MNL and -NL models, depicting integration of ML toolset with basic DCM models. Finally, a baseline DNN model is implemented, representing the contrast of state of the art ML toolset.

A. MNL and NL model translation to NN graph

The Neural Networks may be interpreted as interconnected graph of computational neurons performing transformation operations on the received inputs. This allows us to translate nearly any imaginable functional transformation into a NN format. In the case of the MNL model this operation was first

described around year 2000 (Bentz and Merunka 2000; Hruschka, Fettes, and Probst 2001). The model later reappeared in the work of Sifringer, Lurkin, and Alahi (2020) as a part of the Learning Multinomial Logit (L-MNL) and as an intermediary part in the works of S. Wang, Mo, and Zhao (2020) and S. Wang et al. (2021).

Nevertheless, for a better understanding we are obliged to start with the most basic model bundling together the concepts of CNN and MNL models. The transition from MNL to CNN is rather straightforward in this case. The MNL model already has the latent components integrated into its structure, which are the deterministic utilities V . Those deterministic utility elements may be viewed as elements of a hidden layer. However, in order to respect the structure of the classical MNL model with attribute and characteristic specific effects there is a need to impose further restrictions on the hidden layer generation functions. It is unreasonable in this case to use a fully connected network structure, because it will lead to the situation where all the inputs are involved in construction of each of the hidden layer elements.

There are different options to impose those restrictions over the model. One of the simplest ones is to use the convolutions to calculate the alternative specific utilities. And while the focus is made on the conditional Logit, where only alternatives' attributes play some role in utility computation, this is rather simple. In the case of complex mixes of individual characteristics, attributes and environmental effects the NN internal design might appear rather cumbersome.

Here is a relatively short example formalising the resulting NN structure. The designed CNN consists of two transformation layers. The first one is 1D convolutional layer¹¹ with a linear activation function each. It takes as input the dataset in a *wide* format¹², and produces a single value as an output for each alternative. This is effectively an equivalent of computation of deterministic utilities V_j in the context of classic MNL model. Thus we can interpret the elements of the last network layer, preceding the softmax transformation (z_k), as deterministic utilities V_j , assuming $j = k$. All this assuming the probability of choosing alternative ω_j from among those available $\{\omega_1, \dots, \omega_k\} \in \Omega$ by individual i , can be expressed in closed form as:

$$P_{ij} = \frac{e^{V_{ij}/\sigma}}{\sum_{l=1}^k e^{V_{il}/\sigma}}$$

To translate this into a NN format, let's assume that each alternative is described by a set of d attributes, and no individual characteristics are simplified. This means that convolutional layer have a size and stride both set to d , assuming that all the alternative specific attributes are grouped by alternative. The second layer is a restricted softmax transformation layer, which directly applies softmax transformation over the inputs, without any supplementary permutations. The only restriction imposed for this layer is the absence of the weights to calculate for the algorithm. In the baseline MNL model the V_j are typically not subject to further transformations for the purposes of choice probability calculation.

The vector of inputs issued from the dataset transformed into the "wide" format can be represented as¹³: $\mathbf{X}_i = \{\mathbf{x}_{i,j=1}, \mathbf{x}_{i,j=2}, \dots, \mathbf{x}_{i,j=K}\}$. Where $j = k \in \{1, \dots, K\}$ and each element $\mathbf{x}_{i,j}$ is a vector of d

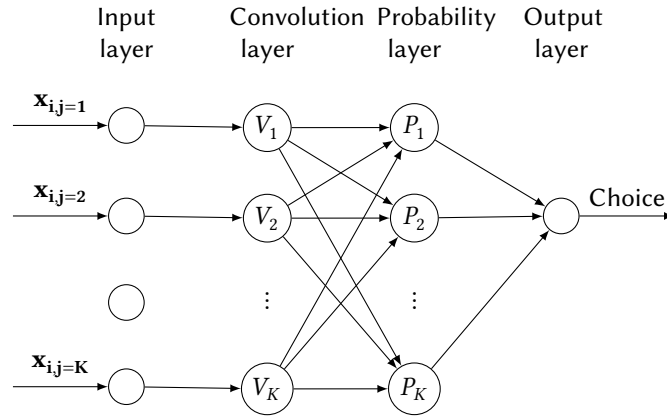
¹¹Meaning that each convolution produces a unique value.

¹²Sometimes this data format is referred as *flattened* format. It assumes that all alternative specific attributes are input into the model in a single vector format.

¹³Although here we focus on a linear version of the model representation with a 1D convolution layer, it can as well take form of a 2D convolution layer. A matrix of $\mathbf{x}_{i,j}$ with K rows is constructed in this case, where each line corresponds to

attribute values corresponding to the alternative j . This simple transformation allows us to use 1D convolutions alongside this vector with both size and stride equal to d , Ensuring that the input vector \mathbf{X}_i with dimensions of $d \times K$ will be transformed into a vector of relative deterministic utilities \mathbf{V} of size K . Each element of this vector is thus composed of a linear combination of single 1D convolution across $\mathbf{x}_{i,j}$. The graphical representation may be depicted in figure 3.5.

Figure 3.5.: Convolution Neural Network design



However, such simple interpretation poses some difficulties once there are individual characteristics involved, or once there are effects varying across alternatives (Conditional Multinomial Logit). To incorporate the variation between the alternative specific effects, as well as their interactions, we may adopt a more complex approach. Within the convolution on per-alternative basis we should include the effects structure information. This means that instead of simple concatenation of the alternative specific vectors $\mathbf{x}_{i,j}$, those should be remapped into a higher dimension space. Hence we introduce the new element $\hat{\mathbf{x}}_{i,j}$, which contains additional information on the coefficients varying between the alternatives. Let assume that an element $x_{i,j,r} \in \mathbf{x}_{i,j}$ is an attribute which has per-alternative effects. Assuming \mathbf{v}_j is a vector of zeroes and ones corresponding to the alternative index, such that for alternative j :

$$\mathbf{v}_j : v_k = \begin{cases} 1 & k = j \\ 0 & k \neq j \end{cases}$$

Thus to a vector $\mathbf{x}_{i,j} = \{x_{i,j,1}, \dots, x_{i,j,r}, \dots, x_{i,j,R}\}$ corresponds a new vector: $\hat{\mathbf{x}}_{i,j} = \{x_{i,j,1}, \dots, \mathbf{v}_j \times x_{i,j,r}, \dots, x_{i,j,R}\}$. This procedure may be performed for multiple elements of $\mathbf{x}_{i,j}$ and performs more complex structures, tying together the values of several alternatives, while leaving some other out. This leaves us with the final vector of input variables of type: $\hat{\mathbf{X}}_i = \{\hat{\mathbf{x}}_{i,j=1}, \hat{\mathbf{x}}_{i,j=2}, \dots, \hat{\mathbf{x}}_{i,j=K}\}$ The size and strides of convolution window should be adjusted accordingly. As this method ensures that size of $\hat{\mathbf{x}}_{i,j} \forall j$ is identical, the size of convolution window corresponds precisely to the the length of the $\hat{\mathbf{x}}_{i,j}$ vector.

In this paper we extend the previous findings by extending this transformation to accommodate the NL model. This includes a rather complex transformation of the deterministic utilities layer output prior to obtaining the choice probabilities. Assuming the λ_m is a scaling parameter for the nest m , the choice probability for an alternative j in NL model might be expressed as:

the alternative specific utility part.

$$P_{ij} = \frac{e^{V_{ij}/\lambda_l}}{\sum_{k \in \Omega_l} e^{V_{ik}/\lambda_l}} \times \frac{(\sum_{k \in \Omega_l} e^{V_{ik}/\lambda_l})^{\lambda_l}}{\sum_{m=1}^M (\sum_{k \in \Omega_m} e^{V_{ik}/\lambda_m})^{\lambda_m}}$$

Where the first term expresses the probability of choosing the alternative j within the nest l . And the second one corresponds to the probability of choice of the nest l among all the available nests. Unfortunately this transformation might be rather unstable due to the bound weights for *multiplications* (scaling) and *power* layers. An alternative approach discussed in the work of Sifringer, Lurkin, and Alahi (2020) involves the usage of logarithmic transformation, which in fact makes the estimation far more complex as the TensorFlow backend does not have inbuilt functionality for gradient calculation for suggested logarithmic transformation.

B. ASUDNN-MNL and -NL versions

The *Alternative Specific Utility Deep Neural Network* was initially described by S. Wang, Wang, and Zhao (2020) as one of the natural extensions for the works of Bentz and Merunka (2000) and Hruschka, Fettes, and Probst (2001). The resulting model remains RUM compliant, while adding the alternative-specific DNN layers for each alternative, which allows for better detection of the non-linear interactions between the attributes.

A logical extension of this baseline model was introduced by S. Wang, Wang, and Zhao (2020) and further S. Wang et al. (2021). The shortest description of the new method may be seen as:

It could be considered as a stack of fully connected subnetworks, with each computing a utility score for each alternative.

The key modification involved addition to the *Fully-connected Deep Network* (FDN) layers between inputs and deterministic utilities, thus ensuring that the $V_j = \mathcal{F}(\mathbf{x}_j)$, where \mathcal{F} reflects the FDN transformations. The FDN layer in the original paper are assumed to have ReLU activations.

Another change involves the relaxation of the convolution layer restrictions, as nothing in the work of S. Wang, Wang, and Zhao (2020) indicates on the usage of convolution layer in strict sense, but rather a dimension reduction of the hidden layers.

The resulting model thus respects the baseline utility theory, assuming that V_k is independent from the $\mathbf{j} \neq \mathbf{k}$. The alternative models are equally possible, but are deprived of logical interpretation. For example, one may introduce a Fully-connected Deep Network prior to convolution layers across all alternatives' inputs, which would imply that different alternatives' attributes influence the alternative specific deterministic utilities V . Or one could imagine the introduction of several FDN layers after the convolution and the probability computations, which would be nearly identical from the interpretation point of view.

Identically to the previous section we are able to extend this model by modifying the transformations applied on the deterministic utilities.

3.4.2.2. Performance comparison framework

To tackle issues in model performance analysis from a fresh perspective, we introduce a framework for performance comparison. This framework encompasses key steps, ranging from defining the research question to conducting performance comparisons within the given context. Drawing inspiration from the concepts outlined by Williams and Ortuzar (1982), we have revised and extended them.

We advocate for structuring this framework in alignment with the traditional scientific research procedure. In literature, irrespective of the case at hand, all research originates from a problem — whether it's a question that needs answering or a barrier that needs overcoming. Once the task is defined, various strategies can be employed to proceed. Some strategies are conventional and detailed in practical guides (Baltagi 2008; Hastie, Tibshirani, and Friedman 2009; Wooldridge 2012; Agresti 2013), while others are more unconventional and occasionally criticized for their unusual practices (Daly, Hess, and Ortúzar 2022). It's noteworthy that the topics we address here are primarily discussed in epistemological works rather than in more abundant applied studies. Nevertheless, gaining a general understanding of typical procedures and paths implemented in applied research is crucial for constructing the framework effectively.

The procedure may be in general divided into several major steps (Figure 3.7). First of all, every research starts with a *problematic* identification and *operational or economic question* definition¹⁴. Every investigation originates from a specific requirement - an “operational problem” that needs addressing. The initial step involves transforming the real-world issue into a more delimited context of a research-oriented question. Subsequently, the researcher must make assumptions about the data's nature and underlying processes. Typically, this stage involves defining a hypothetical interaction model based on theoretical assumptions or a preliminary analysis of the available data. Thus, the second step is an expansion of the problem's narrowing, translating it into numerical terms: identifying target metrics. These metrics should enable the researcher to respond to the research question. For instance, if causality exploration is the goal, this could be translated into analysing the significance of specific coefficients in an econometric model. Another example is a prediction task, where researchers aim to predict consumer behaviour (e.g., identifying market shares), translating into a comparison of various performance metrics for different predictive models. Once the target is defined, the research may take various paths, contingent on the available information. This step, summarily referred to as the “*data collection and analysis*” process, involves either using existing data to build the model or prebuilding the model and then collecting data. Ultimately, data analysis furnishes the actor with information on the target metrics (estimates), enabling them to address the initial question and provide a solution to the initial problem.

¹⁴In this context, we refrain from discussing the “research question” explicitly, as it might not always have a direct connection to the “economic question” addressed in the study. Furthermore, the “question” could solely be “operational”, lacking the generation of novel knowledge and being entirely context-specific to a particular application.

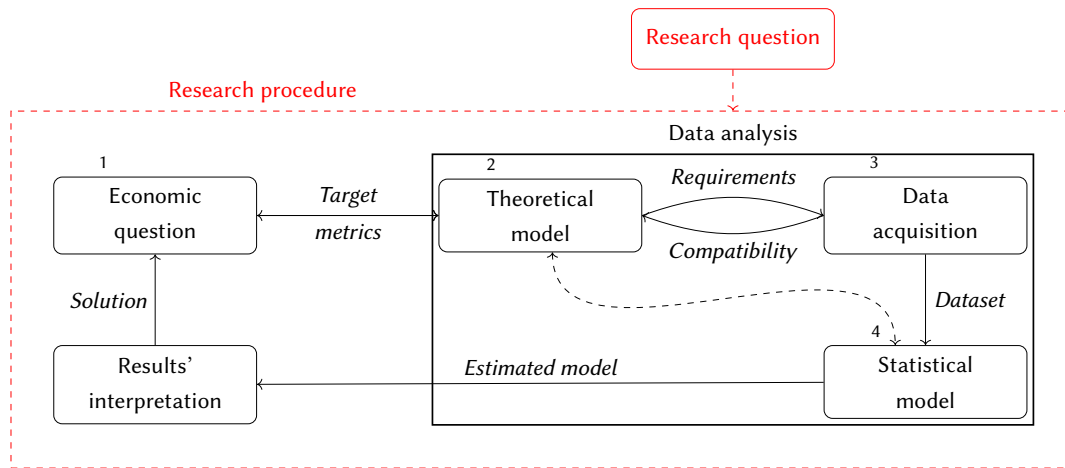


Figure 3.6.: Proposed performance comparison framework

Evaluating a model’s performance requires considering the specific context of the research. We redefine “*performance*” as the model’s capacity to provide accurate answers within the context of the explored problem. This change in the definition directs our focus away from the typical model performance comparison and towards another concept - *the procedure*. The procedure encompasses the entire process, from defining the research question to arriving at the corresponding answer, involving steps such as data collection, processing, and analysis. It also includes choices made during model configuration, selection, and refinement, whether arbitrary or not.

Consequently, the framework’s dependence on the research question is inevitable. In many applied cases, some models may lack the capability to answer certain questions, or there may be no established practices for their usage. Defining the research question thus becomes the initial phase in the suggested framework, providing the researcher with specific metrics for contemplating model comparisons. In this particular study this scientific context is given by the WTP estimation task, the analysis of performances will be performed with the WTP elicitation task as the main target. Speaking of targets frequent in literature we mostly encounter the general predictive quality (Lederrey et al. 2021). Although classification accuracy is typically emphasized in statistical modelling for performance assessment and comparison, it may not always be the best option. Some studies go further to assess the quality of the estimates (Sifringer, Lurkin, and Alahi 2020) or even the derived metrics (S. Wang, Mo, and Zhao 2020). Regardless of the research question, one will always have target metrics or criteria in mind for a complete comparison procedure. This necessitates comparing not only the models among themselves but also comparing the results with externally defined targets.

The subsequent stage involves the dataset choice or dataset generation procedure, encompassing assumptions and a priori choices on assumed individual behaviour, external effects, and potential biases. Opinions on this stage may vary, as typically the applied and theoretical studies rely on different data sources (Japkowicz and Shah 2011). While the first ones operate on the unexplored data, the second category involves usage of more well known datasets or even usage of simulated data.

The modelling procedure, constituting the next stage, involves choosing and implementing the model, configuring estimated utility functions, and more. Later, it becomes subject to numerical specificity, including the choice of the estimation algorithm, its implementation, the particular code base, and the approach to problem-solving. In this study we explore the effects of performances of several relatively distant modelling strategies. While all of the models are supposed to be RUM-compliant, the

estimation procedures vary drastically and the model parametrisation is not trivial in most cases. The hyperparameters are pretrained on a subset of data and then a single set of them is employed for the subsequent model estimation in bulk. Finally, the post-treatment of obtained estimates comes into play. In this case, it involves WTP calculation if the model was not estimated directly in the WTP space. Indicators obtained at this step should be evaluated in the context of the research question and, if possible, compared to target values used as inputs for the simulation task. Now that the framework is fully described, we can proceed with the application.

While the overall structure is dictated by the classical academic guidelines (Hastie, Tibshirani, and Friedman 2009; Agresti 2013), we also rely on a series of unstructured interviews¹⁵ with practising researchers and engineers for framework construction. The *data analysis* stage of the *scientific procedure* may be further divided into several major steps. Without loss of generality this step may be summarised as *data analysis* process. In the discussions with practising researchers three key elements were recurrent: (1) theoretical foundation, incorporating theoretical model structure and assumptions over the real world state; (2) data, alongside with its collection and pre-processing methods; and (3) statistical models and other analysis methods. All those elements are put in evidence in the structure proposed (Figure 3.7).

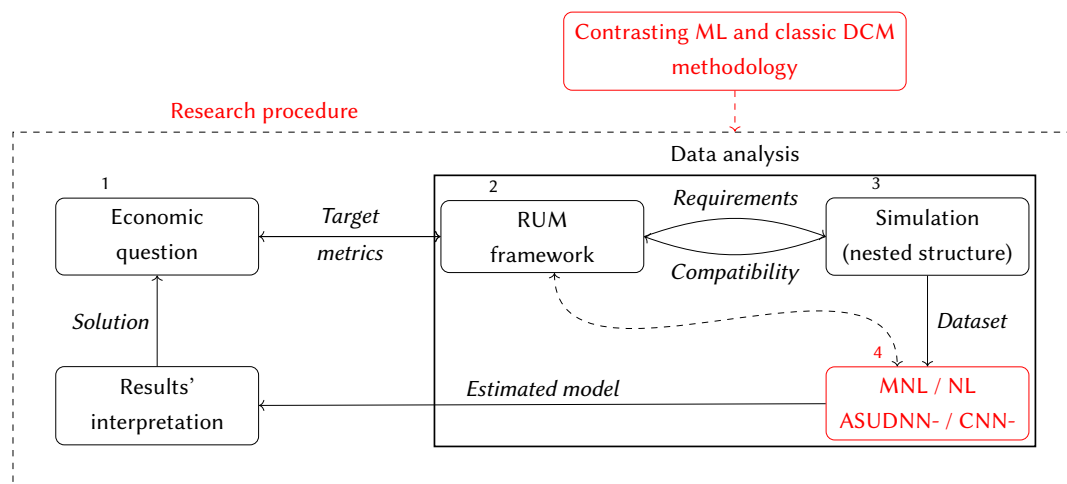


Figure 3.7.: Proposed performance comparison framework

3.4.3. Application

To exemplify, we employ the widely known and publicly accessible dataset *swissmetro*. This dataset was initially introduced in the work of M. Bierlaire, Axhausen, and Abay (2001), where it served to evaluate the acceptance of the state-proposed modal innovation, as discussed by A. Nash et al. (2007). A detailed description of the dataset, along with the dataset itself, can be found on the *biogeme* project website. This data has been utilized in numerous illustrations showcasing newly developed model capabilities, as well as in several tasks comparing model performance. The most closely related works to our use-case are the further works of M. Bierlaire, Bolduc, and McFadden (2008) and Newman, Ferguson, and Garrow (2013), which inspire our data simulation procedure.

We rely on preceding works to construct artificial datasets of different sample sizes and configurations.

¹⁵A series of more than 20 in person unstructured interviews was conducted in a period from May 2022 to March 2023 among researchers from University Grenoble Alpes, University Paris Saclay, University of Montreal, University of Laval and Polytechnic of Montreal.

The first simulation follows the previous works in imposing a conventional *Nested Logit* (NL) structure, which reflects quite common in reality decision rule structure. Several models are then estimated over the resulting datasets. The tests for overall model accuracy as well as the WTP estimates validity are then performed. The obtained estimates are then compared with expected target results. The overall significance of estimates is equally taken into account.

3.4.3.1. Data simulation strategy

The initial dataset, presented by M. Bierlaire, Axhausen, and Abay (2001), amalgamates data from both revealed preferences (RP) and stated preferences (SP) collected in Switzerland in March 1998. The first phase involved gathering initial observational information on subjects' trips. Subsequently, during the SP data collection, participants were presented with a novel hypothetical alternative, the *swissmetro*. To ensure the relevance of this new hypothetical transportation mode, subjects were approached while travelling on the target routes. A total of 470 observations (435 deemed suitable) were collected on the train between St. Gallen and Geneva. An additional 770 usable SP surveys were gathered from car users, facilitated by mail with the support of the central Swiss car license agency. In the SP segment, the authors employed a fractional factorial design, presenting alternatives such as rail (TRAIN), *swissmetro* (SM), and car (CAR, exclusively for car owners). These alternatives were characterized by travel time, fare/cost, and headway (for rail-based alternatives only).

For this study we adopt the approach described by M. Bierlaire, Bolduc, and McFadden (2008) and later used by Newman, Ferguson, and Garrow (2013). The original dataset will be used for simulation purposes, which allows us to observe the model performances in a more controlled environment. Prior to simulation the dataset is filtered, excluding the observations for which there is no choice made and limiting our attention to the commute and business purpose trips. We equally filter out the cases where one of the alternative is not available to the subject, in order to simplify the resulting data structure and lower the complexity of the created NN models. While standard statistical libraries, such as `apollo` or `biogeme`, have the toolset for management of such cases, this may pose some issues in NN models, resulting to inappropriate gradient values and consequently non-convergence of the optimisation algorithm. This may result in divergence from the existing literature, when assessing the results.

We proceed with a simulated dataset, which is based on the original one. The simulation approach adopted is identical to the one performed by M. Bierlaire, Bolduc, and McFadden (2008). Each observation is replicated 100 times to provide us with synthetic observations. The alternative attributes values were overwritten by draws from normal distribution $N(\mu, \sigma^2)$, where μ is the value of the corresponding attribute in the original dataset, and $\sigma = 0.05\mu$ (M. Bierlaire, Bolduc, and McFadden 2008).

Speaking about the decision rules, we decide to adopt the identical nested logit structure as in the other studies (M. Bierlaire, Bolduc, and McFadden 2008; M. Bierlaire, Axhausen, and Abay 2001). The choice model specification is given in the Table 3.16¹⁶.

¹⁶The data structure is subject to verification, other studies, even the ones introducing similar CNN-NL models have not presented their estimates.

Table 3.16.: Utility specification

| Utility | Value | TRAIN | SM | CAR |
|-----------------------|---------|-------|------|------|
| Parameter | | | | |
| ASC_{CAR} | -0.1880 | 0 | 0 | 1 |
| ASC_{SM} | 0.1470 | 0 | 1 | 0 |
| β_{TRAIN_TIME} | -0.0107 | TT | 0 | 0 |
| β_{SM_TIME} | -0.0081 | 0 | TT | 0 |
| β_{CAR_TIME} | -0.0071 | 0 | 0 | TT |
| β_{COST} | -0.0083 | COST | COST | COST |
| Nests (NL) | | | | |
| $\lambda_{EXISTING}$ | 0.4405 | 1 | 0 | 1 |
| λ_{FUTURE} | 1.0000 | 0 | 1 | 0 |

Nesting structure was introduced through error components following the specification provided by M. Bierlaire, Bolduc, and McFadden (2008)¹⁷. This structure assumed that alternatives can be separated according to their real availability. Meaning that while error components behave identically for existing transportation modes (CAR and TRAIN), the effects may differ for non-existing (*future*) alternative.

The final step includes drawing random observations from the resulting database to compose individual datasets of desired size and class-distribution. We vary the sample size from 200 observations per mode choice, a number quite often encountered in econometric studies, to 20000 observations, which approaches the frontier of the datasets available for some very simple ML tasks. The focus is made solely on the perfectly balanced datasets, not taking into account any cases requiring adjustment to the class balance at runtime. For each dataset configuration we randomly draw 100 datasets and estimate selected model over them. Such approach to simulation allows us not only to obtain a consistent baseline for performance assessment, but also the possibility to compare our results with similar papers, where identical simulation strategy was implemented.

3.4.3.2. Model estimation

For the purposes of this study we implement several closely related econometric models, which might be potentially used by novices in choice modelling. Among them: (1) multinomial logistic regression based models, both with and without nesting structure; (2) CNN-MNL and -NL models introducing ML driven estimation to the classic econometric models; (3) ASUDNN-MNL and -NL models representing the combination of ML and classic DCM analysis. The NL model follows the specification used during the simulation step and is expected to perform the best on the available data. The MNL model differs from it only by the absence of the nests, meaning the nesting parameter α is omitted.

Table 3.17.: Tested model configurations

| | Homogeneous | Nesting structure |
|------------|-------------|-------------------|
| DCM | | |

¹⁷For this purpose we used the `evd::rmvevd()` function in R

| | Homogeneous | Nesting structure |
|-------------|-------------|-------------------|
| Regressions | MNL | NL |
| NN | | |
| Estimation | CNN-MNL | CNN-NL |
| Combination | ASUDNN-MNL | ASUDNN-NL |

For the NN model estimation *Adam* algorithm is used with the adjusted hyperparameters. The hyperparameter selection task was performed prior to the model estimation for the purposes of model comparison, although in real-world scenario this entail a significant change in the time requirements for the model estimation.

3.4.3.3. Results

As we have previously shown, in the literature there is no known consensus on the performance metrics and the “*model performance*” definition. As our study focuses on the WTP estimates, we assume that the objective of a model can be viewed as correct estimation of the target metrics. The WTP in its turn relies on the correct estimation of the effects within the model, assuming that the functional form is known and true. The resulting pool of performance indicators includes: (1) overall accuracy; (2) direct effect estimates for concerned models; (3) execution times and resource efficiency.

Thus we are interested to observe the shares of estimation routines which manage to correctly identify the effects. For each sample size we estimate a series of models to get the idea of resulting performance indicators distribution, as in the end the researchers would be interested in a model that consistently performs up to the declared quality.

A. Prediction quality

In terms of predictive quality, the expected results of comparing a ML and classic DCM models depend on the complexity of the underlying data patterns. At this point one of the first errors becomes evident, as the data generating process is extremely simplistic for MNL not to detect the underlying structure. The ML empowered techniques in this case are in disadvantage, as the burden of learning an underlying functional form within the data is identical and facing a simple relationship does not speed up the procedure. This could be observed in the accuracy of the selected algorithms on Figure 3.18. CNN-MNL and -NL perform on par with the basic MNL and NL models given a sufficient amount of data to be able to learn the underlying functional form.

However, the ASUDDN-MNL and -NL model, empowered with additional hidden layers for learning complex functional forms, remain only slightly better than random guessing. At this point another error becomes apparent, because the theoretically more powerful algorithm struggles to identify the correct relationships within the data. This could be explained by the potentially incorrect choice of hyperparameters for the model. The problem of hyperparameter choice in NNs is a critical and challenging aspect of designing effective models. One of the primary challenges is the sheer number of hyperparameters involved, including: learning rate, batch size, number of layers, number of neurons per layer, activation functions, and regularization parameters. The set of hyperparameters for this application was fixed based on results of a preliminary exploration of the data and did not undergo any

Table 3.18.: Accuracy

| N | ASUDNN-MNL | ASUDNN-NL | CNN-MNL | CNN-NL | MNL | NL |
|-------|------------|-----------|-----------|-----------|-----------|-----------|
| 600 | 0.3290333 | 0.3333333 | 0.4424667 | 0.4577500 | 0.5066333 | 0.5146167 |
| 1500 | 0.3418267 | 0.3328733 | 0.4702800 | 0.4887267 | 0.5026000 | 0.5106200 |
| 6000 | 0.3549583 | 0.3378867 | 0.5009183 | 0.5017133 | 0.5026617 | 0.5103267 |
| 15000 | 0.3767073 | 0.3517700 | 0.5029553 | 0.5022580 | 0.5033153 | 0.5099480 |
| 60000 | 0.3899020 | 0.3683357 | 0.5028212 | 0.5023437 | 0.5027228 | 0.5094365 |

| N | ASUDNN-MNL | ASUDNN-NL | CNN-MNL | CNN-NL | MNL | NL |
|-------|------------|-----------|-----------|-----------|-----------|-----------|
| 600 | 0.0361761 | 0.0000000 | 0.0167665 | 0.0221430 | 0.0192596 | 0.0158111 |
| 1500 | 0.0427372 | 0.0108435 | 0.0135571 | 0.0157879 | 0.0124695 | 0.0126083 |
| 6000 | 0.0394478 | 0.0152220 | 0.0066681 | 0.0058089 | 0.0060427 | 0.0058428 |
| 15000 | 0.0453254 | 0.0304658 | 0.0042481 | 0.0039502 | 0.0038734 | 0.0039025 |
| 60000 | 0.0575745 | 0.0391024 | 0.0019716 | 0.0018236 | 0.0020964 | 0.0018348 |

changes throughout the experimentation. It should be acknowledged that the optimal hyperparameter values often depend on the specific characteristics of the dataset, implying that the hyperparameters should have been *learned* in a case-by-case scenario. Unfortunately this illustrates the issues of complex algorithms usage for simplistic problems.

Finally, the third error related to the ML learning algorithms implementation concerns the data pre-treatment procedure. It is advised in the literature (Hastie, Tibshirani, and Friedman 2009) to apply scaling to the inputs before proceeding with the *learning* stage. Scaling in the context of NNs typically refers to the process of normalizing or standardizing input data to improve the training and performance of the model. In the classic DCM methods this requirement does not play such important role, although scaling is still sometimes for convenience. In the NNs and related techniques the scaling plays a crucial role in the optimisation procedure as it ensures that the weights are updated in more or less uniform way. While in the simple models, such as CNN-MNL and -NL the scaling does not play such important role, in ASUDNN- architecture it would ensure a more optimal propagation of weight updates across the hidden layers.

B. Direct effects

In classic DCM, estimating direct effects involves determining the impact of changes in explicative variables on the probability of choosing a particular alternative, a transportation mode in this case. In the most simple models, the effects correspond to the estimates obtained. This allows to extract the associated weights from the simplest models to get the first ideas of the underlying relationship. The results are represented in the Table 3.19, where more errors become apparent.

First of all, the effects observed on estimation for all of the explored models appear biased when compared to the original inputs in Figure 3.16. While the changes in magnitude are expected due to the scaling of the inputs, the constant bias observed across all the models is more difficult to explain.

Next on the list is the extremely high value of λ observed for the ensemble of the CNN-NL models, which could indicate on an internal problem with the parameter estimation. Inside the neural network this parameter is used in a shared layer with a conditional transformation function. This could potentially lead to the errors in application of standard ML algorithms for the effect estimation.

Table 3.19.: Estimates

| | CNN-MNL | MNL | | CNN-NL | NL |
|--------------|------------|------------|--------------|------------|------------|
| ASC_TRAIN | 0.0000000 | 0.0000000 | ASC_TRAIN | 0.0000000 | 0.0000000 |
| ASC_CAR | -0.6933187 | -1.2539412 | ASC_CAR | -0.7762713 | -0.5866941 |
| ASC_SM | -0.8839766 | -1.4789945 | ASC_SM | -0.9058268 | -1.2860138 |
| B_CAR_TIME | -0.6879133 | -0.9706347 | B_CAR_TIME | -0.7107690 | -0.6279280 |
| B_SM_TIME | -0.6729723 | -1.0665629 | B_SM_TIME | -0.6632983 | -0.5471086 |
| B_TRAIN_TIME | -1.0419996 | -1.6043070 | B_TRAIN_TIME | -1.0311351 | -0.9221138 |
| B_COST | -1.0520054 | -1.2398357 | B_COST | -1.0394767 | -0.6899825 |
| | | | LAMBDA | 0.8204898 | 0.3782743 |

Table 3.20.: Execution time

| N | ASUDNN-MNL | ASUDNN-NL | CNN-MNL | CNN-NL | MNL | NL |
|-------|------------|-----------|----------|----------|---------|----------|
| 600 | 4.13996 | 4.74054 | 3.18671 | 4.00651 | 0.34394 | 0.95066 |
| 1500 | 5.00537 | 5.60575 | 3.89620 | 4.88473 | 0.40095 | 1.13067 |
| 6000 | 10.58837 | 11.66464 | 8.34579 | 10.80898 | 0.60355 | 1.96570 |
| 15000 | 22.50346 | 25.18025 | 18.05468 | 23.32200 | 0.97665 | 3.59499 |
| 60000 | 79.19823 | 89.97170 | 64.85465 | 85.83365 | 3.10462 | 10.74091 |

| N | ASUDNN-MNL | ASUDNN-NL | CNN-MNL | CNN-NL | MNL | NL |
|-------|------------|-----------|-----------|-----------|-----------|-----------|
| 600 | 0.2987457 | 0.2615996 | 0.1797158 | 0.2510467 | 0.0345148 | 0.0772726 |
| 1500 | 0.3260201 | 0.4175809 | 0.1889260 | 0.3159278 | 0.0228599 | 0.0818590 |
| 6000 | 0.3476757 | 0.5007037 | 0.3297749 | 0.4615765 | 0.0280301 | 0.0545128 |
| 15000 | 0.4166670 | 0.8005042 | 0.5210972 | 1.0199491 | 0.0494722 | 0.2939483 |
| 60000 | 1.5261298 | 2.0289448 | 1.0106308 | 2.0140425 | 0.1298905 | 1.1088749 |

C. Resource efficiency

The final target in this research involved the observation of the estimation times. Among the expected results were, according to the literature, a potentially lower estimation times for the ML algorithms. The Table 3.20 once again proves our expectations as erroneous, with high estimation times for all of the NN-related techniques.

Estimation times for NN models and classic DCM, such as MNL or NL, can significantly differ based on the complexity of the models and the size of the datasets. The complex NNs might have rather high computational intensity during training due to the large number of parameters and the iterative nature of backpropagation. However in large samples and with relatively low NN's complexity the ML algorithms should be able to outperform the classic DCM counterparts. As it can be observed from the obtained results, the samples of size below 60000 entries are still extremely small to represent a computational burden for well optimised DCM toolset. At the same time, the NN-related algorithms from TensorFlow suite are less optimised, and in our benchmarks do not take advantage in usage of *Graphic Processing Units* (GPU). While we can observe that the increase in estimation time for CNN- and ASUDNN- models is less steep with sample size augmentation, the maximum sample size limit in this study does not allow the models to reach a switching point.

3.4.4. Conclusion

In terms of immediate results we follow our previous work comparing the MNL, NL and MMNL model precision in the context of sample size and balance variations (Gusarov, Joly, and Lemaire 2023). This paper extends the experimental procedure to the broader spectre of the RUM compliant models, offering a more in depth focus on the differences in RUM-compliant models available nowadays.

This intermediary work serves as a remainder of eventual complexities associated with interdisciplinary statistical modelling. While the proposed framework allows to systematise the elements and facilitates the design of scientific procedure and associated testing, the task remains extremely complex to be accomplished in a short amount of time.

In the complete work we expect to confirm our previous observations with larger confidence intervals for the restricted ASU-DNN model specification, as it was partially demonstrated in the work of S. Wang, Wang, and Zhao (2020). The unbalanced configuration of the simulated dataset was expect to contribute towards biased estimates regardless of the model type due to the common RUM-compliant structure. Those findings should complement the most recent findings (S. Wang et al. 2021; Gonzalez-Valdes, Heydecker, and Ortúzar 2022), providing a discussion interlinking the analysis of the novel model families to the more classical methods.

3.4.5. Discussion

This study attempts to offer a comprehensive comparison of traditional DCM and emerging ML techniques, insisting on the performance comparison framework's role in the process. In its final form this paper is intended to present the full potential of the performance comparison framework in the context of economics applied and theoretical studies. For this purposes various models with DCM and ML related background, among which the ASUDNN- and CNN- extensions of the classic MNL and NL models, are applied to transportation-related research questions. However, in the current stage the work presents only a fraction of the intended results, as it puts in evidence the eventual errors committed in the performance analysis task performed under limited time constraints.

The performance comparison framework, enriched by a simulation-driven application, sheds light on the intricacy of the model comparison task. The paper discusses diverse issues in the adopted scientific procedure for detection of differences in interpretability, predictive quality, and estimation times for the different modelling approaches.

The obtained results reveal pitfalls, such as sensitivity to hyperparameter choice or the misconception in the experimental setting, favouring the classic DCM approaches. The study exposes errors in terms of predictive quality, with ML techniques struggling to outperform basic DCMs when faced with a simplistic data generation process. The analysis also unravels issues related the critical role of scaling in NN models, emphasizing the need for meticulous tuning and data preprocessing in future experiments.

What is more, the examination of direct effects estimates reveals biases and anomalies across models, confirming potential challenges in applying standard ML algorithms to effect estimation tasks in small samples. The estimation times, a key consideration in model selection, also diverge from initial expectations. The NN-related techniques in this case exhibiting unexpectedly high computational demands, highlighting the importance of optimization and hardware considerations, as well as software choice. Such observation partially contradict to the previously obtained ones. For example, in the Section 3.2,

Table 3.5, the comparison of CNN- based model against MNL one produce different results. This could be explained by a larger sample size considered in previous study, as well as the eventual differences in used software and its versions.

In essence, this work contributes insights into the complex landscape of model performance, offering a cautionary tale on the rushed application of advanced techniques to seemingly straightforward problems. The research underscores the need for a nuanced approach, considering the specific characteristics of the data, hyperparameter tuning, and computational efficiency, to make informed decisions in empirical studies.

Conclusion

The advancements in computing efficiency and increased data availability, previously resource-intensive data analysis methods have gained in popularity. Those changes particularly affected the disciplines reliant on statistical learning, among which the economics related studies. Once computationally demanding models can now be estimated in a fraction of time, leading current research to focus more on big data and automation. While there exist a number of studies focusing on bridging the interdisciplinary gap between the disciplines and striving to popularise innovative approaches to data analysis, the applied studies often remain constrained by the more accessible modelling techniques. The availability of new approaches to data analysis does not reduce the burden associated to the model selection, and leads to a choice overload for the non-proficient users.

In this study the problem of model performance assessment in the context of the consumer choice modelling is addressed. The Discrete Choice Modelling remains a rather complex task and the increased availability of advanced ML toolset only increases the flexibility in modelling techniques selection. The plethora of available data analysis strategies can be overwhelming for non-experts trying to choose the optimal solution. As option has its own strengths and weaknesses, making it easy for inexperienced users to overlook key elements, there is a growing need for a better understanding of the strengths and weaknesses of various models.

The first part of the work, particularly Chapter 1.1, offers an overview of current state of practices in choice modelling applications in economics, taking into account the interdisciplinary context as well. This puts in evidence the complexities and discrepancies that exist among the different disciplines and application contexts. From the baseline economics applications to the transportation research and preference studies, the choice modelling techniques are widely spread resulting in major differences in practices, dictated directly by the underlying use-cases. Those differences expand affecting not only the established conventions and practices, but going deeper to the vocabulary and terminology used by researchers.

The fast-paced research environment does not facilitate the search for common ground between the disciplines. Each year more and more novel data analysis approaches, their combinations and transformation emerge. This makes the initially implied task of the modelling approaches taxonomisation nearly impossible, due to the natural limitations of cognitive capabilities of a single researcher. The constant monitoring of the literature across several fields remains extremely difficult task, thus obliging to search for alternative solutions.

In this work one of natural answers to the problematic is proposed under a format of a performance testing and comparison framework, construction of which is outlined in the Chapter 2. While it is nearly impossible to reunite all the growing amount of information on different modelling techniques, it is still possible to provide a toolset for the model comparison and selection to the community of applied users. Even though it does not provide an answer to the knowledge acquisition problem in

highly-paced environment, it offers a toolset for modelling technique selection and fine-tuning.

The performance comparison lies at the heart of many optimisation tasks and is widely used in various statistical analysis contexts. The choice modelling community has previously relied mostly on the baseline statistical, and more precisely - econometric, approaches to the model performance comparison. Unfortunately, the focus on the model, as in *statistical model* or *econometric model*, imposed a series of constraints to the implications of such comparisons. The basic comparison of accuracy or of the confidence intervals for the estimates might not be always optimal, as the performance is not always defined in the numerical terms.

This thesis offers an argumentation on why the performance perceptions should not be limited to the baseline statistical performance indicators. While the alternative approaches to the performance understanding made their appearance in the scientific literature for decades, there was no evidence of any work aggregating those practices and providing a comprehensive and user friendly procedure for their application. Only several rare publications provided the link between the model estimates and the implications for public policies or other results of the estimates usage. The focus in particular is made on this exact dimensions: the link between the research question and the performance of applied techniques. Due to the complexities and particularities of the most advanced models, an argument is made on why the performance assessment should be performed not on the model, but on the data analysis procedure in whole.

The main contribution to the model performances comparison in the format of procedure focused performance analysis framework is presented in Section 2.5. It provides the guidelines for both expert and non-proficient users on the model performance assessment and comparison. Several illustration of its usage are provided in different contexts. First of all, several existing state-of-the-art studies are positioned according to the proposed framework to better illustrate how the existing practices in the model performance analysis are taken into the account by this toolset. A series of applications is then performed to illustrate how the future studies might be guided by the proposed framework. A discussion is provided for each of the explore cases and the evolution of the framework might be traced across the illustrations.

In the final Chapter 3, the performance comparison framework introduced in the previous section is explored further through a series of case studies. Each case study delves into different elements of the framework, focusing on modeling stage relationships, data acquisition issues, and statistical modeling within the broader data analysis procedure. The first study combines econometrics and machine learning models for consumer choice preference modeling, introducing a simulation and theory-testing framework. The second study concentrates on the WTP elicitation task, systematically evaluating model performance in this context by considering potential misspecifications, changes in sample size, and dataset balance. The third case study explores the comparison between econometric and machine learning models in the context of commute mode choice modeling, using the *swissmetro* dataset and synthetic samples to contrast conventional discrete choice models with emerging machine learning alternatives in the WTP estimation task. Despite potential inconsistencies in the framework's vision and presentation due to the varying maturity stages of the individual papers, the chapter aims to offer insights into model performance analysis and comparison for effective model selection, illustrating the evolution of the framework over different works. Each section offers a short discussion on insights obtained and differences from the final framework version.

In conclusion, this thesis addresses the challenges arising from the evolving landscape of data analysis techniques, especially in the context of consumer choice modeling. The proposed performance testing and comparison framework, detailed in Chapter 2, emerges as a valuable toolset for systematic model performance exploration. By shifting the focus from mere statistical performance indicators to a comprehensive understanding of the implications for explored research questions, the framework provides a nuanced approach to model performance assessment. The case studies in Chapter 3 further illustrate the framework's application in diverse contexts, shedding light on modeling stage relationships, data acquisition issues, and statistical modeling within the broader data analysis procedure. Despite the evolving nature of the framework across the studies, this work contributes to the ongoing discourse on model performance evaluation and consequently on model selection.

Glossary

Acronyms

ABS Agent Based Simulation.

AE Auto Encoding.

AIC Akaike Information Criterion.

ASUDNN Alternative Specific Utility Deep Neural Network.

AVC Asymptotic Variance-Covariance.

BA Bayesian Analysis.

BIC Bayesian Information Criterion.

CE Choice Experiment.

CM Choice Modelling.

CNL Cross-Nested Logit.

CNN Convolutional Neural Network.

CS Computer Science.

DCA Discrete Choice Analysis.

DCE Discrete Choice Experiment.

DCM Discrete Choice Modelling.

DFT Decision Field Theory.

DL Deep Learning.

DNN Deep Neural Network.

DoE Design of Experiments.

DSCM Dynamic Structural Choice Models.

DT Decision Theory.

Dyn.CM Dynamic Choice Models.

ED Experimental Design.

ESML Exogenous Sample Maximum Likelihood.

EU Expected Utility.

EV Extreme Value.

FF Full Factorial.

GAM Generalised Additive Models.

GCM Generalised Choice Models.

GEV Generalised Extreme Value.

GLM Generalised Linear Models.

GPL GNU General Public License.

GPU Graphic Processing Unit.

HCM Hybrid Choice Models.

ICLV Integrated Choice and Latent Variable.

IIA Independence from Irrelevant Alternatives.

IID identically and independently distributed.

LCCM Latent Class Choice Model.

LCM Latent Class Model.

LCRPL Latent Class Random Parameter Logit.

Logit Logistic Regression.

LR Linear Regression.

LVCM Latent Variable Choice Model.

MaaS Mobility-as-a-Service.

MC Monte-Carlo based Simulation.

MDFT Multialternative Decision Field Theory.

ML Machine Learning.

ML Maximum Likelihood.

MLP Multilayer Perceptron.

MMNL Mixed Logit.

MNL Multinomial Logistic Regression.

MSLE Maximum Simulated Likelihood.

MSM Method of Simulated Moments.

NL Nested Logit.

NN Neural Network.

OCM Online Choice Models.

OLS Ordinary Least Squares.

OOP Object Oriented Programming.

OSI Open Source Initiative.

PCA Principal Component Analysis.

PL Preference Learning.

PPR Project Pursuit Regression.

Probit Probability Unit Regression.

PSF Python Software Foundation.

PT Prospect Theory.

QC Quantum Choice.

QDT Quantum Decision Theory.

QP Quantum Probability.

QUM Quantum Utility Model.

R-FF Randomised Full Factorial.

RAM Random Advantage Maximisation.

RF Random Forest.

RP Revealed Preferences.

RRM Random Regret Minimisation.

RUM Random Utility Maximisation.

Sc Stated Choice.

SEM Structural Equation Models.

SHS Social and Human Sciences.

SL Statistical Learning.

SP Stated Preferences.

VTTS Value of Travel Time Savings.

WCML Weighted Conditional Maximum Likelihood.

WESML Weighted Exogenous Sample Maximum Likelihood.

WFH work from home.

WoS Web of Science.

WTP Willingness to Pay.

XML Mixed Logit.

XOR Exclusive OR.

Special terms

algorithm a step-by-step set of well-defined instructions or rules to perform a specific task or solve a particular problem.

backpropagation a supervised learning algorithm used to train artificial neural networks, involving iteratively updating the network's weights by computing the gradient of the loss function with respect to the weights and adjusting them in the opposite direction of the gradient.

behavioural model a representation that describes how individuals or entities are likely to act, make decisions, and interact based on observed patterns, psychological factors, and past behaviours.

bibliometrics a quantitative research method that involves the statistical analysis of publications, citations, and other bibliographic data to assess patterns, trends, and the impact of scientific or academic research.

biometrics a field of study within biology concerned with the theory and technique of measurement.

classification a statistical or machine learning task where the goal is to assign predefined categories or labels to input data based on its features.

computer science a systematic study of algorithms, data structures, computational processes, and the design and analysis of computer systems, with a focus on solving problems and advancing technology through the application of computational principles.

data analysis the process of acquiring, inspecting, cleaning, transforming, and modeling data to discover meaningful patterns, draw conclusions, and support decision-making in various domains.

data driven an approach or decision-making process that relies on empirical evidence, information, or insights derived from the analysis of data.

econometric model a statistical representation used to analyze and quantify the relationships among economic variables, incorporating both economic theory and empirical data to make predictions or test hypotheses about economic phenomena.

econometrics a field of study within economics concerned with the theory and technique of measurement.

economics a social science that studies the production, distribution, and consumption of goods and services.

estimate a calculated approximation or prediction of a value, often based on available data, observations, or statistical methods.

estimation algorithm a computational procedure used to calculate or infer the values of unknown parameters in a statistical model based on observed data.

Gumbel random variable a type of probability distribution used in statistics, particularly in extreme value theory, the probability density function of which is characterized by an exponential decay and is given by the formula:

$$f(x; \mu, \sigma) = \frac{1}{\sigma} \exp \left[-\frac{x - \mu}{\sigma} - \exp \left(-\frac{x - \mu}{\sigma} \right) \right]$$

informatics the interdisciplinary field that involves the study, design, and implementation of information systems and computational technologies to organize, analyze, and manage data, facilitating efficient information processing and decision-making across various domains.

marketing a strategic process of promoting, distributing, and selling products or services to target audiences, encompassing activities such as advertising, market research, branding, and sales to create awareness, generate demand, and build customer relationships.

mathematical model a formal representation that facilitates the description, analysis, and prediction of real-world phenomena using mathematical structures and relationships.

mode choice a decision-making process individuals go through when selecting a transportation mode.

model an informative representation of an object, person or system.

Normal distribution also known as Gaussian distribution, is a symmetric probability distribution that is characterized by a bell-shaped curve.

performance comparison the process of evaluating and contrasting the effectiveness, efficiency, or quality.

preference learning a type of machine learning focused on modeling and predicting an individual's preferences or ranking of items.

psychology the scientific study of the mind and behavior, exploring various aspects of human thought, emotion, perception, and social interactions to understand and explain mental processes.

psychometrics a field of study within psychology concerned with the theory and technique of measurement.

rational behaviour a concept that individuals make choices that maximize their utility or satisfaction, taking into account available information, preferences, and constraints, in a manner consistent with logical decision-making.

real world an objective environment that exists independently of individual perception, encompassing the physical universe, natural phenomena, and human society, where events, experiences, and interactions occur.

scientific procedure a systematic and structured series of steps, including hypothesis formulation, experimentation, data collection, analysis, and conclusion, used to investigate natural phenomena and test hypotheses.

sigmoid a mathematical function, often used in machine learning, which produces an S-shaped curve and is mathematically represented as:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

sociology a scientific study of society, human behavior, and the social structures, institutions, and processes that shape and influence individuals and groups within a community or society.

sociometrics a field of study within sociology concerned with the theory and technique of measurement.

state of the art the highest level of general development, as of a device, technique, or scientific field achieved at a particular time.

statistical model a mathematical model that embodies a set of statistical assumptions concerning the generation of sample data (and similar data from a larger population).

statistical modelling a mathematical modelling process that embodies a set of statistical assumptions concerning the generation of sample data (and similar data from a larger population).

target metrics a numerical or logical value(s) obtained as a result of data analysis, serving to answer the research question.

theoretical model a simplified and abstract representation of a system or phenomenon that is created to analyze, understand, or explain its fundamental principles, relationships, and behaviors.

theory driven an approach in research or problem-solving where the development and testing of hypotheses are guided by existing theories or conceptual frameworks.

universal approximator a mathematical model or algorithm that has the capability to approximate any continuous function with arbitrary precision, given a sufficient number of parameters or resources.

Tables

List of Figures

| | |
|--|-----|
| 1.1. Citation map on document level (1975 - 2021) | 13 |
| 1.2. Citation map on document level - focus on Economics (1975 - 2021) | 13 |
| 1.3. Real world | 17 |
| 1.4. Data analysis approaches | 17 |
| 1.5. Formal neuron (alternative representation) | 20 |
| 1.6. XOR problem | 21 |
| 1.7. Multilayer Perceptron | 22 |
| 1.8. Taxonomy of DCM as proposed by Hastie and Tibshirani (2009), simplified | 39 |
| 1.9. Taxonomy as proposed by Agresti (2013), simplified | 39 |
| 1.10. Taxonomy as proposed by Ayodele (2010) | 40 |
| | |
| 2.1. Research procedure in applied studies | 54 |
| 2.2. Research procedure, subject to external limitations | 55 |
| 2.3. Target metrics in research procedure | 56 |
| 2.4. Research procedure, complete | 56 |
| 2.5. Data analysis | 57 |
| 2.6. Data analysis | 63 |
| 2.7. Data acquisition strategies | 65 |
| 2.8. Data simulation | 66 |
| 2.9. Data collection | 70 |
| 2.10. Statistical modelling | 71 |
| 2.11. Statistical modelling in detail | 72 |
| 2.12. Introduction to framework | 78 |
| 2.13. All data analysis elements | 79 |
| 2.14. Data analysis stage simplified | 80 |
| 2.15. Research procedure in detail | 81 |
| 2.16. Proposed performance comparison framework | 82 |
| 2.17. Commuter choice model and WFH analysis research procedure | 84 |
| 2.18. WCML algorithm for GEV models in choice-based sample presence | 85 |
| 2.19. Sample size requirements exploration for SC experiments | 87 |
| | |
| 3.1. Proposed framework | 92 |
| 3.2. Positioning of DA2PL 2020 paper | 103 |
| 3.3. Proposed performance comparison framework | 107 |
| 3.4. Positioning of ITEA 2023 paper | 114 |

| | |
|---|-----|
| 3.5. Convolution Neural Network design | 120 |
| 3.6. Proposed performance comparison framework | 123 |
| 3.7. Proposed performance comparison framework | 124 |
| A.1. Number of publications by year (2011 - 2021) | 170 |
| A.2. Keyword network map (1975 - 2022) | 174 |
| A.3. Keyword density map (1975 - 2022) | 175 |
| A.4. Co-occurrences map, all keywords (1975 - 2021) | 178 |
| A.5. Co-occurrences map, all keywords - average citation year (1975 - 2021) | 178 |
| A.6. Co-occurrences map, all keywords - focus on Economics (1975 - 2021) | 179 |
| A.7. Citation map on document level (1975 - 2021) | 181 |
| A.8. Citation map on document level - focus on Economics (1975 - 2021) | 181 |
| B.1. Formal neuron | 192 |
| B.2. Formal neuron (alternative representation) | 193 |
| B.3. Example: gradient descent update (assuming $\eta = 0.5$) | 194 |
| B.4. XOR problem | 196 |
| B.5. XOR problem solution | 197 |
| B.6. MLP solution for XOR | 197 |
| B.7. Multilayer Perceptron | 198 |
| B.8. Convolution Neural Network design | 205 |
| B.9. Convolution Neural Network design | 206 |
| B.10. NN and DCM | 208 |
| F.1. Cadre proposé pour la comparaison des performances | 236 |

List of Tables

| | |
|--|-----|
| 1.1. Sources composition | 11 |
| 1.2. ML notation | 19 |
| 1.3. Discrete Choice Modelling (DCM) notation extension | 27 |
| 1.4. Expected Utility (EU) Theory notation | 33 |
| 1.5. Prospect Theory (PT) notation | 33 |
| 1.6. Decision Field Theory (DFT) notation | 34 |
| 1.7. Quantum Decision Theory (QDT) notation | 36 |
| 1.8. Performance gains (computation time) by Arteaga et al. (2022) | 47 |
| 2.1. Graphical representation conventions | 54 |
| 2.2. Data types in choice analysis studies | 68 |
| 3.1. The assumed relative utility function parameters | 97 |
| 3.2. Individuals' descriptive statistics by dataset | 97 |
| 3.3. Alternatives' descriptive statistics by dataset | 98 |
| 3.4. Estimation results | 99 |
| 3.5. General performance measures | 101 |
| 3.6. Performance in terms of WTP and premiums | 101 |

| | |
|---|-----|
| 3.7. Descriptive statistics | 109 |
| 3.8. Utility specification | 109 |
| 3.9. True WTP (VOT) values | 110 |
| 3.10. Utility specification for MNL model | 111 |
| 3.11. Utility specification for MMNL model | 111 |
| 3.12. Shares of WTP estimates not different from target, by sample size. | 112 |
| 3.13. Shares of WTP estimates different from zero, by sample size. | 112 |
| 3.14. Shares of all correctly estimated WTP by sample size. | 112 |
| 3.15. Shares of all correctly estimated WTP by dataset balance. | 113 |
| 3.16. Utility specification | 126 |
| 3.17. Tested model configurations | 126 |
| 3.18. Accuracy | 128 |
| 3.19. Estimates | 129 |
| 3.20. Execution time | 129 |
| | |
| A.1. Total number of keyword appearances (2011 - 2021) | 171 |
| A.2. The most occurring keywords by year (2011 - 2021) | 171 |
| A.3. Sources composition | 173 |
| A.4. The most occurring keywords (1975 - 2022) | 174 |
| A.5. Keywords related to Economics (1975 - 2022) | 176 |
| A.6. Keywords related to Sociology and Psychology (1975 - 2022) | 176 |
| A.7. Keywords related to Transportation (1975 - 2022) | 177 |
| A.8. The most occurring keywords in Economics related cluster (1975 - 2021) | 180 |
| A.9. Top 10 most cited works in Economics cluster (1975 - 2021) | 182 |
| A.10. Policy adoption or assessment related keywords (1975 - 2022) | 184 |
| A.11. The most cited documents related to Policy (1975 - 2022) | 184 |
| A.12. Preference or Attitude related keywords (1975 - 2022) | 186 |
| A.13. Most cited works concerning Preference or Attitudes (1975 - 2022) | 187 |
| | |
| B.1. Notation | 192 |
| B.2. Extracting information | 208 |

Index

- RUM-compliant, 25
- alternative, 25
- answer, 56, 81
- attributes, 27
- Backpropagation, 22
- Bayesian Estimation, 26
- behavioural model, 17
- characteristics, 27
- Choice Modelling, 24
- computer science, 46
- Conditional Logit, 28
- Cross-Nested Logit, 26
- data analysis, 54, 79
- data driven, 8
- data transformation, 74
- Deep Neural Networks, 22
- Discrete Choice Analysis, 25
- Discrete Choice Modelling, 24
- Dynamic Choice Models, 26
- Dynamic Structural Choice Models, 26
- economic question, 81
- estimates, 60
- estimation results, 81
- Exclusive OR, 21
- experimental design, 68
- Extreme Value, 25
- Generalised Additive Models, 22
- Generalised Linear Models, 22
- Generalised RUM models, 26
- goodness of fit, 52
- Hybrid Choice Models, 26
- identically and independently distributed, 28
- IIA, 28
- Independence from Irrelevant Alternatives, 26
- informatics, 46
- Integrated Choice and Latent Variable models, 26
- Julia, 78
- Linear Regression, 21
- Logistic Regression, 25
- Logit, 21, 25
- Machine Learning, 1, 18, 19
- Mixed MNL, 26
- Mixed Multinomial Logit, 29
- model taxonomy, 37
- Multilayer Perceptron, 22
- Multinomial Logistic Regression, 25
- Multinomial Logit, 27
- Nested Logit, 26, 29
- Neural Network, 18
- Online Choice Models, 26
- perceptron, 20
- performance, 48, 52, 58
- Project Pursuit Regression, 22, 23
- prospect, 32
- Python, 77
- R, 76
- random utility, 25
- Random Utility Maximisation, 25, 26
- research procedure, 52
- research question, 52, 82
- revealed preferences, 67
- SAS, 77
- Stata, 77

stated preferences, 67
statistical bias, 64
Statistical Learning, 1
statistical learning, 73
statistical model, 45
statistical modelling, 72
Structural equation Models, 29
target metrics, 15, 52, 56, 81
theoretical bias, 64
theory driven, 8
universal approximator, 23
utility, 25

Bibliography

- Abdelwahab, Hassan T., and Mohamed A. Abdel-Aty. 2002. "Artificial Neural Networks and Logit Models for Traffic Safety Analysis of Toll Plazas." *Transportation Research Record* 1784 (1): 115–25. <https://doi.org/10.3141/1784-15>.
- Aboutaleb, Youssef M, Mazen Danaf, Yifei Xie, and Moshe E Ben-Akiva. 2021. "Discrete Choice Analysis with Machine Learning Capabilities." *Machine Learning*, 19.
- Abou-Zeid, Maya, and Moshe Ben-Akiva. 2014. "Hybrid Choice Models." In *Handbook of Choice Modelling*, 383–412. Edward Elgar Publishing. <https://doi.org/10.4337/9781781003152.00025>.
- Agrawal, Ajay, Joshua Gans, and Avi Goldfarb. 2019. *The Economics of Artificial Intelligence: An Agenda*. Book. University of Chicago Press. <https://doi.org/10.7208/chicago/9780226613475.001.0001>.
- Agresti, Alan. 2007. *An Introduction to Categorical Data Analysis, Second Edition*.
———. 2013. *Categorical Data Analysis, Third Edition*.
- Aguirregabiria, Victor, and Pedro Mira. 2010. "Dynamic Discrete Choice Structural Models: A Survey." *Journal of Econometrics* 156 (1): 38–67. <https://doi.org/10.1016/j.jeconom.2009.09.007>.
- Albert, James H., and Siddhartha Chib. 1991. "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association*.
———. 1993. "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association* 88 (422): 669–79. <https://doi.org/10.1080/01621459.1993.10476321>.
- Allais, M. 1953. "Le Comportement de l'Homme Rationnel Devant Le Risque: Critique Des Postulats Et Axiomes de l'Ecole Americaine." *Econometrica* 21 (4): 503–46. <https://doi.org/10.2307/1907921>.
- Allenby, Greg M., and Peter E. Rossi. 1998. "Marketing Models of Consumer Heterogeneity." *Journal of Econometrics* 89 (1-2): 57–78. [https://doi.org/10.1016/S0304-4076\(98\)00055-4](https://doi.org/10.1016/S0304-4076(98)00055-4).
- Allouche, Omri, Asaf Tsoar, and Ronen Kadmon. 2006. "Assessing the Accuracy of Species Distribution Models: Prevalence, Kappa and the True Skill Statistic (TSS)." *Journal of Applied Ecology* 43 (6): 1223–32. <https://doi.org/10.1111/j.1365-2664.2006.01214.x>.
- Alwosheel, Ahmad, Sander van Cranenburgh, and Caspar 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: 167–82.
- Amini, Massih-Reza, and Nicolas Usunier. 2015. *Learning with Partially Labeled and Interdependent Data*. Springer.
- Andersson, Arne, Paul Davidsson, and Johan Lindén. 1999. "Measure-Based Classifier Performance Evaluation." *Pattern Recognition Letters* 20 (11): 1165–73. [https://doi.org/10.1016/S0167-8655\(99\)00084-7](https://doi.org/10.1016/S0167-8655(99)00084-7).
- Andrews, R. L., and A. K. Manrai. 1998. "Simulation Experiments in Choice Simplification: The Effects of Task and Context on Forecasting Performance." *Journal of Marketing Research* 35 (2): 198–209. <https://doi.org/10.2307/3151848>.
- Ardeshiri, Ali, Farshid Safarighouzhdi, and Taha Hossein Rashidi. 2021. "Measuring Willingness to Pay for Shared Parking." *Transportation Research Part A: Policy and Practice* 152: 186–202. <https://doi.org/10.1016/j.tra.2021.04.011>.

- [//doi.org/10.1016/j.tra.2021.08.014](https://doi.org/10.1016/j.tra.2021.08.014).
- Arellano, Manuel, and Bo Honoré. 2001. "Panel Data Models: Some Recent Developments." In *Handbook of Econometrics*, 5:3229–96. Elsevier. [https://doi.org/10.1016/S1573-4412\(01\)05006-1](https://doi.org/10.1016/S1573-4412(01)05006-1).
- Arrow, Kenneth. 1951. "Social Choice And Individual Values." <https://sites.duke.edu/niou/files/2014/06/Arrow-Social-Choice-And-Individual-Values.pdf>.
- Arteaga, Cristian, JeeWoong Park, Prithvi Bhat Beeramoole, and Alexander Paz. 2022. "Xlogit: An Open-Source Python Package for GPU-accelerated Estimation of Mixed Logit Models." *Journal of Choice Modelling* 42 (March): 100339. <https://doi.org/10.1016/j.jocm.2021.100339>.
- Askin, Oykum Esra, and Fulya Gokalp. 2013. "Comparing the Predictive and Classification Performances of Logistic Regression and Neural Networks: A Case Study on Timss 2011." *Procedia - Social and Behavioral Sciences* 106: 667–76. <https://doi.org/10.1016/j.sbspro.2013.12.076>.
- Athey, Susan. 2018. "The Impact of Machine Learning on Economics." In *The Economics of Artificial Intelligence: An Agenda*, 507–47. University of Chicago Press.
- Athey, Susan, and Guido Imbens. 2015. "A Measure of Robustness to Misspecification." *American Economic Review* 105 (5): 476–80. <https://doi.org/10.1257/aer.p20151020>.
- Athey, Susan, and Guido W. Imbens. 2019. "Machine Learning Methods That Economists Should Know About." *Annual Review of Economics* 11 (1): 685–725. <https://doi.org/10.1146/annurev-economics-080217-053433>.
- Athey, Susan, and Michael Luca. 2019. "Economists (and Economics) in Tech Companies." *Journal of Economic Perspectives* 33 (1): 209–30. <https://doi.org/10.1257/jep.33.1.209>.
- Athey, Susan, Julie Tibshirani, Stefan Wager, et al. 2019. "Generalized Random Forests." *The Annals of Statistics* 47 (2): 1148–78.
- Austen-Smith, D., and J. S. Banks. 1998. "Social Choice Theory, Game Theory, and Positive Political Theory." *Annual Review of Political Science* 1: 259–87. <https://doi.org/10.1146/annurev.polisci.1.1.259>.
- Ayalew, Lulseged, and Hiromitsu Yamagishi. 2005. "The Application of GIS-based Logistic Regression for Landslide Susceptibility Mapping in the Kakuda-Yahiko Mountains, Central Japan." *Geomorphology* 65 (1): 15–31. <https://doi.org/10.1016/j.geomorph.2004.06.010>.
- Ayodele, Taiwo Oladipupo. 2010. "Types of Machine Learning Algorithms." *New Advances in Machine Learning*, 19–48.
- Baesens, Bart, Sebastiaan Höppner, and Tim Verdonck. 2021. "Data Engineering for Fraud Detection." *Decision Support Systems*, Interpretable Data Science For Decision Making, 150 (November): 113492. <https://doi.org/10.1016/j.dss.2021.113492>.
- Bahamonde-Birke, Francisco J., and Juan de Dios Ortúzar. 2014. "On the Variability of Hybrid Discrete Choice Models." *Transportmetrica A: Transport Science* 10 (1): 74–88. <https://doi.org/10.1080/18128602.2012.700338>.
- Bai, Hui ren. 2022. "The Epistemology of Machine Learning." *Filosofija. Sociologija* 33 (1). <https://doi.org/10.6001/fil-soc.v33i1.4668>.
- Balbontin, Camila, David A. Hensher, and Matthew J. Beck. 2022. "Advanced Modelling of Commuter Choice Model and Work from Home During COVID-19 Restrictions in Australia." *Transportation Research Part E: Logistics and Transportation Review* 162 (June): 102718. <https://doi.org/10.1016/j.tr.e.2022.102718>.
- Baldi, Pierre, Søren Brunak, Yves Chauvin, Claus A. F. Andersen, and Henrik Nielsen. 2000. "Assessing the Accuracy of Prediction Algorithms for Classification: An Overview." *Bioinformatics* 16 (5): 412–

24. <https://doi.org/10.1093/bioinformatics/16.5.412>.
- Baltagi, Badi. 2008. *Econometrics, 4th Edition*. Berlin: Springer.
- Banfi, Silvia, Mehdi Farsi, Massimo Filippini, and Martin Jakob. 2008. "Willingness to Pay for Energy-Saving Measures in Residential Buildings." *Energy Economics* 30 (2): 503–16. <https://doi.org/10.1016/j.eneco.2006.06.001>.
- Bazzani, Claudia, Marco A. Palma, and Rodolfo M. Nayga. 2018. "On the Use of Flexible Mixing Distributions in WTP Space: An Induced Value Choice Experiment." *Australian Journal of Agricultural and Resource Economics* 62 (2): 185–98. <https://doi.org/10.1111/1467-8489.12246>.
- Belgiawan, Prawira Fajarindra, Ilka Dubernet, Basil Schmid, and Kay Axhausen. 2019. "Context-Dependent Models (CRRM, MuRRM, PRRM, RAM) Versus a Context-Free Model (MNL) in Transportation Studies: A Comprehensive Comparisons for Swiss and German SP and RP Data Sets." *Transportmetrica A: Transport Science* 15 (2): 1487–1521. <https://doi.org/10.1080/23249935.2019.1612968>.
- Ben-Akiva, Moshe, and Michel Bierlaire. 2003. "Discrete Choice Models with Applications to Departure Time and Route Choice." In *Handbook of Transportation Science*, edited by Randolph W. Hall, 56:7–37. Boston: Kluwer Academic Publishers. https://doi.org/10.1007/0-306-48058-1_2.
- Ben-Akiva, Moshe, Daniel McFadden, and Kenneth Train. 2019. "Foundations of Stated Preference Elicitation: Consumer Behavior and Choice-based Conjoint Analysis." *Foundations and Trends® in Econometrics* 10 (1-2): 1–144. <https://doi.org/10.1561/08000000036>.
- Ben-Akiva, Moshe, Joan Walker, Adriana T. Bernardino, Dinesh A. Gopinath, Taka Morikawa, and Amalia Polydoropoulou. 2002. "Integration of Choice and Latent Variable Models." In *In Perpetual Motion*, 431–70. Elsevier. <https://doi.org/10.1016/B978-008044044-6/50022-X>.
- Benson, Austin R., Ravi Kumar, and Andrew Tomkins. 2016. "On the Relevance of Irrelevant Alternatives." In *Proceedings of the 25th International Conference on World Wide Web (Www'16)*, 963–73. New York: Assoc Computing Machinery.
- Bentz, Yves, and Dwight Merunka. 2000. "Neural Networks and the Multinomial Logit for Brand Choice Modelling: A Hybrid Approach." *Journal of Forecasting* 19 (3): 177–200. [https://doi.org/10.1002/\(SICI\)1099-131X\(200004\)19:3%3C177::AID-FOR738%3E3.0.CO;2-6](https://doi.org/10.1002/(SICI)1099-131X(200004)19:3%3C177::AID-FOR738%3E3.0.CO;2-6).
- Bergantino, Angela Stefania, Mauro Capurso, and Stephane Hess. 2020. "Modelling Regional Accessibility to Airports Using Discrete Choice Models: An Application to a System of Regional Airports." *Transportation Research Part A: Policy and Practice* 132 (February): 855–71. <https://doi.org/10.1016/j tra.2019.12.012>.
- Bergtold, Jason S., and Steven M. Ramsey. 2015. "Neural Network Estimators of Binary Choice Processes: Estimation, Marginal Effects and WTP." 2015 {{AAEA}} \& {{WAEA Joint Annual Meeting}}, {{July}} 26-28, {{San Francisco}}, {{California}} 205649. Agricultural and Applied Economics Association. <https://doi.org/10.22004/ag.econ.205649>.
- Bhat, Chandra R. 2001. "Quasi-Random Maximum Simulated Likelihood Estimation of the Mixed Multinomial Logit Model." *Transportation Research Part B: Methodological* 35 (7): 677–93. [https://doi.org/10.1016/S0191-2615\(00\)00014-X](https://doi.org/10.1016/S0191-2615(00)00014-X).
- . 2003. "Simulation Estimation of Mixed Discrete Choice Models Using Randomized and Scrambled Halton Sequences." *Transportation Research Part B: Methodological* 37 (9): 837–55. [https://doi.org/10.1016/S0191-2615\(02\)00090-5](https://doi.org/10.1016/S0191-2615(02)00090-5).
- Bhat, Chandra R., and Jessica Y. Guo. 2007. "A Comprehensive Analysis of Built Environment Characteristics on Household Residential Choice and Auto Ownership Levels." *Transportation Research*

- Part B: Methodological* 41 (5): 506–26. <https://doi.org/10.1016/j.trb.2005.12.005>.
- Bierlaire, M, KW Axhausen, and G Abay. 2001. “The Acceptance of Modal Innovation: The Case of Swissmetro,” 17.
- Bierlaire, M., D. Bolduc, and D. McFadden. 2008. “The Estimation of Generalized Extreme Value Models from Choice-Based Samples.” *Transportation Research Part B: Methodological* 42 (4): 381–94. <https://doi.org/10.1016/j.trb.2007.09.003>.
- Bierlaire, Michel. 2006. “A Theoretical Analysis of the Cross-Nested Logit Model.” *Annals of Operations Research* 144 (1): 287–300. <https://doi.org/10.1007/s10479-006-0015-x>.
- Bierlaire, Michel, and Rico Krueger. 2020. “Sampling and Discrete Choice,” 31.
- Birchall, Cameron, and Frank Verboven. 2022. “Estimating Substitution Patterns and Demand Curvature in Discrete-Choice Models of Product Differentiation.” {{SSRN Scholarly Paper}}. Rochester, NY.
- Birol, Ekin, Katia Karousakis, and Phoebe Koundouri. 2006. “Using a Choice Experiment to Account for Preference Heterogeneity in Wetland Attributes: The Case of Cheimaditida Wetland in Greece.” *Ecological Economics* 60 (1): 145–56. <https://doi.org/10.1016/j.ecolecon.2006.06.002>.
- Blacklow, Paul, Amy Beth Corman, and Hugh Sibly. 2021. “The Demand and Supply of Esteem: An Experimental Analysis.” *Journal of Behavioral and Experimental Economics* 95: 101759. <https://doi.org/10.1016/j.socec.2021.101759>.
- Blades, Natalie J., G. Bruce Schaalje, and William F. Christensen. 2015. “The Second Course in Statistics: Design and Analysis of Experiments?” *The American Statistician* 69 (4): 326–33. <https://doi.org/10.1080/00031305.2015.1086437>.
- Blau, Julian H. 1971. “Arrow’s Theorem with Weak Independence.” *Economica* 38 (152): 413–20. <https://doi.org/10.2307/2551881>.
- Bliss, C. I. 1934. “The Method of Probits.” *Science* 79 (2037): 38–39. <https://doi.org/10.1126/science.79.2037.38>.
- Bode, Christoph, John R. Macdonald, and Maximilian Merath. 2022. “Supply Disruptions and Protection Motivation: Why Some Managers Act Proactively (and Others Don’t).” *Journal of Business Logistics* 43 (1): 92–115. <https://doi.org/10.1111/jbl.12293>.
- Bodea, Tudor D, and Laurie A Garrow. 2006. “The Importance of Synthetic Datasets in Empirical Testing: Comparison of NL with MNL with Error Components Models.” In *Proceedings of the European Transport Conference, Strasbourg*.
- Bose, R. C. 1947. “Mathematical Theory of the Symmetrical Factorial Design.” *Sankhyā: The Indian Journal of Statistics (1933-1960)* 8 (2): 107–66. <https://www.jstor.org/stable/25047939>.
- Boto-García, David, Petr Mariel, José Baños Pino, and Antonio Alvarez. 2022. “Tourists’ Willingness to Pay for Holiday Trip Characteristics: A Discrete Choice Experiment.” *Tourism Economics* 28 (2): 349–70. <https://doi.org/10.1177/1354816620959901>.
- Bouscasse, H el ene. 2018. “Integrated Choice and Latent Variable Models: A Literature Review on Mode Choice,” 27.
- Bouscasse, H el ene, Iraga el Joly, and Jean Peyhardi. 2019. “A New Family of Qualitative Choice Models: An Application of Reference Models to Travel Mode Choice.” *Transportation Research Part B: Methodological* 121 (C): 74–91. <https://doi.org/10.1016/j.trb.2018.12.010>.
- Boxall, Peter C., and Wiktor L. Adamowicz. 2002. “Understanding Heterogeneous Preferences in Random Utility Models: A Latent Class Approach.” *Environmental and Resource Economics* 23 (4): 421–46. <https://doi.org/10.1023/A:1021351721619>.

- Boyd, J. Hayden, and Robert E. Mellman. 1980. "The Effect of Fuel Economy Standards on the U.S. Automotive Market: An Hedonic Demand Analysis." *Transportation Research Part A: General* 14 (5): 367–78. [https://doi.org/10.1016/0191-2607\(80\)90055-2](https://doi.org/10.1016/0191-2607(80)90055-2).
- Brathwaite, Timothy, Akshay Vij, and Joan L. Walker. 2017. "Machine Learning Meets Microeconomics: The Case of Decision Trees and Discrete Choice." arXiv. <https://doi.org/10.48550/arXiv.1711.04826>.
- Breiman, Leo et al. 2001. "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)." *Statistical Science* 16 (3): 199–231.
- Brownstone, David, David S. Bunch, and Kenneth Train. 2000. "Joint Mixed Logit Models of Stated and Revealed Preferences for Alternative-Fuel Vehicles." *Transportation Research Part B: Methodological* 34 (5): 315–38. [https://doi.org/10.1016/S0191-2615\(99\)00031-4](https://doi.org/10.1016/S0191-2615(99)00031-4).
- Brownstone, David, and Kenneth Train. 1998. "Forecasting New Product Penetration with Flexible Substitution Patterns." *Journal of Econometrics* 89 (1-2): 109–29. [https://doi.org/10.1016/S0304-4076\(98\)00057-8](https://doi.org/10.1016/S0304-4076(98)00057-8).
- Bur, Oliver Thomas, Tobias Krieger, Steffen Moritz, Jan Philipp Klein, and Thomas Berger. 2022. "Optimizing the Context of Support of Web-Based Self-Help in Individuals with Mild to Moderate Depressive Symptoms: A Randomized Full Factorial Trial." *Behaviour Research and Therapy* 152 (May): 104070. <https://doi.org/10.1016/j.brat.2022.104070>.
- Burgess, Leonie, and Deborah J. Street. 2006. "The Optimal Size of Choice Sets in Choice Experiments." *Statistics* 40 (6): 507–15. <https://doi.org/10.1080/02331880601013841>.
- Busemeyer, J. R., and J. T. Townsend. 1993. "Decision Field Theory: A Dynamic-Cognitive Approach to Decision Making in an Uncertain Environment." *Psychological Review* 100 (3): 432–59. <https://doi.org/10.1037/0033-295x.100.3.432>.
- Busemeyer, Jerome R., and Adele Diederich. 2002. "Survey of Decision Field Theory." *Mathematical Social Sciences*, Random Utility Theory and Probabilistic measurement theory, 43 (3): 345–70. [https://doi.org/10.1016/S0165-4896\(02\)00016-1](https://doi.org/10.1016/S0165-4896(02)00016-1).
- Cardell, N.Scott, and Frederick C. Dunbar. 1980. "Measuring the Societal Impacts of Automobile Downsizing." *Transportation Research Part A: General* 14 (5-6): 423–34. [https://doi.org/10.1016/0191-2607\(80\)90060-6](https://doi.org/10.1016/0191-2607(80)90060-6).
- Caron, Alexandre, Vincent Vandewalle, Romaric Marcilly, Jessica Rochat, and Benoit Dervaux. 2021. "The Optimal Sample Size for Usability Testing, From the Manufacturer's Perspective: A Value-of-Information Approach." *Value in Health*. <https://doi.org/10.1016/j.jval.2021.07.010>.
- Carson, Richard T., and Mikołaj Czajkowski. 2019. "A New Baseline Model for Estimating Willingness to Pay from Discrete Choice Models." *Journal of Environmental Economics and Management* 95 (May): 57–61. <https://doi.org/10.1016/j.jeem.2019.03.003>.
- Cato, Susumu. 2013. "Independence of Irrelevant Alternatives Revisited."
- Caussade, Sebastián, Juan de Dios Ortúzar, Luis I. Rizzi, and David A. Hensher. 2005. "Assessing the Influence of Design Dimensions on Stated Choice Experiment Estimates." *Transportation Research Part B: Methodological* 39 (7): 621–40. <https://doi.org/10.1016/j.trb.2004.07.006>.
- Chan, Wai Kin Victor, Young-Jun Son, and Charles M Macal. 2010. "Agent-Based Simulation Tutorial-Simulation of Emergent Behavior and Differences Between Agent-Based Simulation and Discrete-Event Simulation." In *Proceedings of the 2010 Winter Simulation Conference*, 135–50. IEEE. <https://doi.org/10.1109/WSC.2010.5679168>.
- Chauhan, Rahul S. 2022. "Unstructured Interviews: Are They Really All That Bad?" *Human Resource Development International* 25 (4): 474–87. <https://doi.org/10.1080/13678868.2019.1603019>.

- Chen, Yushi, Zhouhan Lin, Xing Zhao, Gang Wang, and Yanfeng Gu. 2014. "Deep Learning-Based Classification of Hyperspectral Data." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 7 (6): 2094–2107. <https://doi.org/10.1109/JSTARS.2014.2329330>.
- Cheng, Long, Xuewu Chen, Jonas De Vos, Xinjun Lai, and Frank Witlox. 2019. "Applying a Random Forest Method Approach to Model Travel Mode Choice Behavior." *Travel Behaviour and Society* 14 (January): 1–10. <https://doi.org/10.1016/j.tbs.2018.09.002>.
- Cheng, Simon, and J. Scott Long. 2007. "Testing for IIA in the Multinomial Logit Model." *Sociological Methods & Research* 35 (4): 583–600. <https://doi.org/10.1177/0049124106292361>.
- Choo, Sangho, and Patricia L Mokhtarian. 2004. "What Type of Vehicle Do People Drive? The Role of Attitude and Lifestyle in Influencing Vehicle Type Choice." *Transportation Research Part A: Policy and Practice* 38 (3): 201–22. <https://doi.org/10.1016/j.tra.2003.10.005>.
- Chorus, Caspar G. 2010. "A New Model of Random Regret Minimization." *European Journal of Transport and Infrastructure Research* 10 (2): 181–96. <https://doi.org/10.18757/ejtir.2010.10.2.2881>.
- Chung, Hui-Kuan, Tomas Sjöström, Hsin-Ju Lee, Yi-Ta Lu, Fu-Yun Tsoo, Tzai-Shuen Chen, Chi-Fu Chang, Chi-Hung Juan, Wen-Jui Kuo, and Chen-Ying Huang. 2017. "Why Do Irrelevant Alternatives Matter? An fMRI-TMS Study of Context-Dependent Preferences." *Journal of Neuroscience* 37 (48): 11647–61. <https://doi.org/10.1523/JNEUROSCI.2307-16.2017>.
- Condorcet, Marquis De. 1785. *Essai Sur l'application de l'analyse à La Probabilité Des décisions Rendues à La Pluralité Des Voix*.
- Costa, Eduardo, Ana Lorena, ACPLF Carvalho, and Alex Freitas. 2007. "A Review of Performance Evaluation Measures for Hierarchical Classifiers." In *Evaluation Methods for Machine Learning II: Papers from the AAAI-2007 Workshop*, 1–6.
- Coussement, Kristof, Dries F. Benoit, and Dirk Van den Poel. 2010. "Improved Marketing Decision Making in a Customer Churn Prediction Context Using Generalized Additive Models." *Expert Systems with Applications* 37 (3): 2132–43. <https://doi.org/10.1016/j.eswa.2009.07.029>.
- Cybenko, G. 1989. "Approximation by Superpositions of a Sigmoidal Function," 12.
- Daly, Andrew, Stephane Hess, and Gerard de Jong. 2012. "Calculating Errors for Measures Derived from Choice Modelling Estimates." *Transportation Research Part B: Methodological*, Emerging and Innovative Directions in Choice Modeling, 46 (2): 333–41. <https://doi.org/10.1016/j.trb.2011.10.008>.
- Daly, Andrew, Stephane Hess, and Juan De Dios Ortúzar. 2022. "Estimating Willingness-to-Pay from Discrete Choice Models: Setting the Record Straight." *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4281881>.
- Daly, Andrew, Stephane Hess, and Kenneth Train. 2012. "Assuring Finite Moments for Willingness to Pay in Random Coefficient Models." *Transportation* 39 (1): 19–31. <https://doi.org/10.1007/s11116-011-9331-3>.
- Danaf, Mazen, Bilge Atasoy, and Moshe Ben-Akiva. 2020. "Logit Mixture with Inter and Intra-Consumer Heterogeneity and Flexible Mixing Distributions." *Journal of Choice Modelling* 35: 100188. <https://doi.org/10.1016/j.jocm.2019.100188>.
- Danaf, Mazen, Felix Becker, Xiang Song, Bilge Atasoy, and Moshe Ben-Akiva. 2019. "Online Discrete Choice Models: Applications in Personalized Recommendations." *Decision Support Systems* 119 (April): 35–45. <https://doi.org/10.1016/j.dss.2019.02.003>.
- Daunou, P. C. F. 1803. "Mémoire Sur Les Élections Au Scrutin. Baudouin, Imprimeur de l'Institut National, Paris. Translated and Reprinted in I. McLean and AB Urken, Eds.(1995), *Classics of Social Choice*." University of Michigan Press, Ann Arbor.

- Daziano, Ricardo A., and Martin Achtnicht. 2014. "Accounting for Uncertainty in Willingness to Pay for Environmental Benefits." *Energy Economics* 44 (July): 166–77. <https://doi.org/10.1016/j.eneco.2014.03.023>.
- De Palma, André, Moshe Ben-Akiva, David Brownstone, Charles Holt, Thierry Magnac, Daniel McFadden, Peter Moffat, Nathalie Picard, Kenneth Train, and Peter Wakker. 2008. "Risk, Uncertainty and Discrete Choice Models." *Marketing Letters* 19 (3): 269–85. <https://doi.org/10.1007/s11002-008-9047-0>.
- dell'Olio, Luigi, Angel Ibeas, and Patricia Cecin. 2011. "The Quality of Service Desired by Public Transport Users." *Transport Policy* 18 (1): 217–27. <https://doi.org/10.1016/j.tranpol.2010.08.005>.
- DeShazo, J. R., and German Fermo. 2002. "Designing Choice Sets for Stated Preference Methods: The Effects of Complexity on Choice Consistency." *Journal of Environmental Economics and Management* 44 (1): 123–43. <https://doi.org/10.1006/jeem.2001.1199>.
- Domshlak, Carmel, Eyke Hüllermeier, Souhila Kaci, and Henri Prade. 2011. "Preferences in AI: An Overview." *Artificial Intelligence* 175 (7): 1037–52. <https://doi.org/10.1016/j.artint.2011.03.004>.
- Donoho, David. 2017. "50 Years of Data Science." *Journal of Computational and Graphical Statistics* 26 (4): 745–66.
- Dormann, Carsten F., Jane Elith, Sven Bacher, Carsten Buchmann, Gudrun Carl, Gabriel Carré, Jaime R. García Márquez, et al. 2013. "Collinearity: A Review of Methods to Deal with It and a Simulation Study Evaluating Their Performance." *Ecography* 36 (1): 27–46. <https://doi.org/10.1111/j.1600-0587.2012.07348.x>.
- Dubernet, Ilka, and Kay W. Axhausen. 2020. "The German Value of Time and Value of Reliability Study: The Survey Work." *Transportation* 47 (3): 1477–1513. <https://doi.org/10.1007/s11116-019-10052-4>.
- Durlauf, Steven N., and Lawrence E. Blume, eds. 2010. *Behavioural and Experimental Economics*. London: Palgrave Macmillan UK. <https://doi.org/10.1057/9780230280786>.
- El-Badawy, Sherif, Marwa Elharoun, and Usama Shahdah. 2021. "Captivity Impact on Modelling Mode Choice Behavior." *Advances in Transportation Studies* LIII (April): 85–102.
- Erişkin, Levent. 2021. "Preference Modelling in Sorting Problems: Multiple Criteria Decision Aid and Statistical Learning Perspectives." *Journal of Multi-Criteria Decision Analysis* n/a (n/a). <https://doi.org/10.1002/mcda.1737>.
- Fifer, Simon, John Rose, and Stephen Greaves. 2014. "Hypothetical Bias in Stated Choice Experiments: Is It a Problem? And If so, How Do We Deal with It?" *Transportation Research Part A: Policy and Practice* 61 (March): 164–77. <https://doi.org/10.1016/j.tra.2013.12.010>.
- Firth, David. 1993. "Bias Reduction of Maximum Likelihood Estimates." *Biometrika* 80 (1): 27–38. <https://doi.org/10.1093/biomet/80.1.27>.
- Flach, Peter. 2019. "Performance Evaluation in Machine Learning: The Good, the Bad, the Ugly, and the Way Forward." In *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:9808–14.
- Fosgerau, Mogens, and Michel Bierlaire. 2007. "A Practical Test for the Choice of Mixing Distribution in Discrete Choice Models." *Transportation Research Part B: Methodological* 41 (7): 784–94. <https://doi.org/10.1016/j.trb.2007.01.002>.
- Fosgerau, Mogens, Daniel McFadden, and Michel Bierlaire. 2013. "Choice Probability Generating Functions." *Journal of Choice Modelling* 8 (September): 1–18. <https://doi.org/10.1016/j.jocm.2013.05.002>.
- Friedman, Jerome H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *The Annals of Statistics* 29 (5): 1189–1232. <https://doi.org/10.1214/aos/1013203451>.

- Friedman, Jerome H., and Werner Stuetzle. 1981. "Projection Pursuit Regression." *Journal of the American Statistical Association* 76 (376): 817–23. <https://doi.org/10.1080/01621459.1981.10477729>.
- Fry, Tim R. L., and Derek Chong. 2005. *A Tale of Two Logits, Compositional Data Analysis and Zero Observations*. Universitat de Girona. Departament d'Informàtica i Matemàtica Aplicada.
- Fürnkranz, J., and E. Hüllermeier. 2010. *Preference Learning*. Springer Verlag, Berlin.
- Gangi, Massimo Di, and Antonino Vitetta. 2021. "Quantum Utility and Random Utility Model for Path Choice Modelling: Specification and Aggregate Calibration from Traffic Counts." *Journal of Choice Modelling* 40: 100290. <https://doi.org/10.1016/j.jocm.2021.100290>.
- García-García, José Carlos, Ricardo García-Ródenas, Julio Alberto López-Gómez, and José Ángel Martín-Baos. 2022. "A Comparative Study of Machine Learning, Deep Neural Networks and Random Utility Maximization Models for Travel Mode Choice Modelling." *Transportation Research Procedia* 62: 374–82. <https://doi.org/10.1016/j.trpro.2022.02.047>.
- Gardiner, Tony. 2016. *Teaching Mathematics at Secondary Level*. OBP Series in Mathematics. Open Book Publishers. <https://doi.org/10.11647/OBP.0071>.
- Garrow, Laurie A, Tudor D Bodea, and Misuk Lee. 2010. "Generation of Synthetic Datasets for Discrete Choice Analysis." *Transportation* 37 (2): 183–202.
- Gatta, Valerio, Edoardo Marcucci, and Luisa Scaccia. 2015. "On Finite Sample Performance of Confidence Intervals Methods for Willingness to Pay Measures." *Transportation Research Part A: Policy and Practice* 82 (December): 169–92. <https://doi.org/10.1016/j.tra.2015.09.003>.
- Gellrich, Mario, Priska Baur, Barbara Koch, and Niklaus E. Zimmermann. 2007. "Agricultural Land Abandonment and Natural Forest Re-Growth in the Swiss Mountains: A Spatially Explicit Economic Analysis." *Agriculture, Ecosystems & Environment* 118 (1): 93–108. <https://doi.org/10.1016/j.agee.2006.05.001>.
- Gensch, D. H., and S. Ghose. 1997. "Differences in Independence of Irrelevant Alternatives at Individual Vs Aggregate Levels, and at Single Pair Vs Full Choice Set." *Omega-International Journal of Management Science* 25 (2): 201–14. [https://doi.org/10.1016/S0305-0483\(96\)00047-3](https://doi.org/10.1016/S0305-0483(96)00047-3).
- Geržinič, Nejc, Sander van Cranenburgh, Oded Cats, Emily Lancsar, and Caspar Chorus. 2021. "Estimating Decision Rule Differences Between 'Best' and 'Worst' Choices in a Sequential Best Worst Discrete Choice Experiment." *Journal of Choice Modelling* 41: 100307. <https://doi.org/10.1016/j.jocm.2021.100307>.
- Goff, Sandra H. 2021. "A Test of Willingness to Pay as Penance in the Demand for Ethical Consumption." *Journal of Behavioral and Experimental Economics* 94: 101744. <https://doi.org/10.1016/j.socec.2021.101744>.
- Goldberg, Pinelopi Koujianou. 1995. "Product Differentiation and Oligopoly in International Markets: The Case of the U.S. Automobile Industry." *Econometrica* 63 (4): 891–951. <https://doi.org/10.2307/2171803>.
- Gong, Zhiqiang, Ping Zhong, and Weidong Hu. 2021. "Statistical Loss and Analysis for Deep Learning in Hyperspectral Image Classification." *IEEE Transactions on Neural Networks and Learning Systems* 32 (1): 322–33. <https://doi.org/10.1109/TNNLS.2020.2978577>.
- González, Rosa Marina, Concepción Román, and Ángel Simón Marrero. 2021. "Values of Travel Time for Recreational Trips Under Different Behavioural Rules." *Sustainability* 13 (12): 6831. <https://doi.org/10.3390/su13126831>.
- Gonzalez-Valdes, Felipe, Benjamin G. Heydecker, and Juan de Dios Ortúzar. 2022. "Quantifying Behavioural Difference in Latent Class Models to Assess Empirical Identifiability: Analytical Devel-

- opment and Application to Multiple Heuristics.” *Journal of Choice Modelling* 43 (June): 100356. <https://doi.org/10.1016/j.jocm.2022.100356>.
- Green, William. 2018. *Econometric Analysis*. Eighth edition. New York, NY: Pearson.
- Greene, William H., and David A. Hensher. 2003. “A Latent Class Model for Discrete Choice Analysis: Contrasts with Mixed Logit.” *Transportation Research Part B: Methodological* 37 (8): 681–98. [https://doi.org/10.1016/S0191-2615\(02\)00046-2](https://doi.org/10.1016/S0191-2615(02)00046-2).
- Guevara, C. Angelo, and Moshe E. Ben-Akiva. 2013. “Sampling of Alternatives in Multivariate Extreme Value (MEV) Models.” *Transportation Research Part B: Methodological* 48 (February): 31–52. <https://doi.org/10.1016/j.trb.2012.11.001>.
- Guillon, M. 2020. “Attitudes and Opinions on Quarantine and Support for a Contact-Tracing Application in France During the COVID-19 Outbreak.” *Public Health*, 11.
- Gundlach, Amely, Marius Ehrlinspiel, Svenja Kirsch, Alexander Koschker, and Julian Sagebiel. 2018. “Investigating People’s Preferences for Car-Free City Centers: A Discrete Choice Experiment.” *Transportation Research Part D: Transport and Environment* 63 (August): 677–88. <https://doi.org/10.1016/j.trd.2018.07.004>.
- Gusarov, Nikita. 2022. “A Discrete Choice Experiment (DCE) Simulator: ‘Dcesimulatr’.” Grenoble, France.
- Gusarov, Nikita, Iragaël Joly, and Pierre Lemaire. 2023. “Willingness to Pay Quality Estimates in Commute Mode Choice: Model Performance Comparison Under Sample Size and Balance Impacts.” In *ITEA 2023 Proceedings*. Santander, Spain: GAEL, G-SCOP, Univ. Grenoble Alpes, CNRS, INRAE, Grenoble INP.
- Gusarov, Nikita, Amirreza Talebijamalabad, and Iragaël Joly. 2020. “Exploration of Model Performances in the Presence of Heterogeneous Preferences and Random Effects Utilities Awareness.” In *DA2PL 2020 Proceedings*. Trento, Italy.
- Haboucha, Chana J., Robert Ishaq, and Yoram Shiftan. 2017. “User Preferences Regarding Autonomous Vehicles.” *Transportation Research Part C: Emerging Technologies* 78 (May): 37–49. <https://doi.org/10.1016/j.trc.2017.01.010>.
- Hackbarth, André, and Reinhard Madlener. 2013. “Consumer Preferences for Alternative Fuel Vehicles: A Discrete Choice Analysis.” *Transportation Research Part D: Transport and Environment* 25 (December): 5–17. <https://doi.org/10.1016/j.trd.2013.07.002>.
- Hagenauer, Julian, and Marco Helbich. 2017. “A Comparative Study of Machine Learning Classifiers for Modeling Travel Mode Choice.” *Expert Systems with Applications* 78: 273–82. <https://doi.org/10.1016/j.eswa.2017.01.057>.
- Haghani, Milad, Michiel C. J. Bliemer, and David A. Hensher. 2021. “The Landscape of Econometric Discrete Choice Modelling Research.” *Journal of Choice Modelling* 40 (September): 100303. <https://doi.org/10.1016/j.jocm.2021.100303>.
- Haghani, Milad, Michiel C. J. Bliemer, John M. Rose, Harmen Oppewal, and Emily Lancsar. 2021a. “Hypothetical Bias in Stated Choice Experiments: Part I. Macro-scale Analysis of Literature and Integrative Synthesis of Empirical Evidence from Applied Economics, Experimental Psychology and Neuroimaging.” *Journal of Choice Modelling*, 100309. <https://doi.org/10.1016/j.jocm.2021.100309>.
- . 2021b. “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 (December): 100322. <https://doi.org/10.1016/j.jocm.2021.100322>.
- Haghani, Milad, and Majid Sarvi. 2019. “Laboratory Experimentation and Simulation of Discrete Di-

- rection Choices: Investigating Hypothetical Bias, Decision-Rule Effect and External Validity Based on Aggregate Prediction Measures.” *Transportation Research Part A: Policy and Practice* 130 (December): 134–57. <https://doi.org/10.1016/j.tra.2019.09.040>.
- Haile, Kaleab Kebede, Nyasha Tirivayi, and Wondimagegn Tesfaye. 2020. “Farmers’ willingness to accept payments for ecosystem services on agricultural land: The case of climate-smart agroforestry in Ethiopia.” *DataverseNL*. <https://doi.org/10.34894/PF4NF4>.
- Hall, Randolph W., and Frederick S. Hillier, eds. 2003. *Handbook of Transportation Science*. Vol. 56. International Series in Operations Research & Management Science. Boston, MA: Springer US. <https://doi.org/10.1007/b101877>.
- Han, Yafei, Francisco Camara Pereira, Moshe Ben-Akiva, and Christopher Zegras. 2022. “A Neural-Embedded Discrete Choice Model: Learning Taste Representation with Strengthened Interpretability.” *Transportation Research Part B: Methodological* 163 (September): 166–86. <https://doi.org/10.1016/j.trb.2022.07.001>.
- Hancock, Thomas O., Jan Broekaert, Stephane Hess, and Charisma F. Choudhury. 2020. “Quantum Choice Models: A Flexible New Approach for Understanding Moral Decision-Making.” *Journal of Choice Modelling* 37 (December): 100235. <https://doi.org/10.1016/j.jocm.2020.100235>.
- Hancock, Thomas O., Stephane Hess, and Charisma F. Choudhury. 2018. “Decision Field Theory: Improvements to Current Methodology and Comparisons with Standard Choice Modelling Techniques.” *Transportation Research Part B: Methodological* 107 (January): 18–40. <https://doi.org/10.1016/j.trb.2017.11.004>.
- Hand, David J. 2012. “Assessing the Performance of Classification Methods.” *International Statistical Review* 80 (3): 400–414. <https://doi.org/10.1111/j.1751-5823.2012.00183.x>.
- Hannus, Veronika. 2020. “Data on Farmers’ Perception and Acceptance of Sustainability Standards.” *Data in Brief* 32 (August): 106250. <https://doi.org/10.1016/j.dib.2020.106250>.
- Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.
- Hausman, J., B. Hall, and Z. Griliches. 1984. “Economic Models for Count Data with an Application to the Patents-R&D Relationship.” *Econometrica* 52 (4): 909–38.
- Haynes, Kingsley E., David H. Good, and Tony Dignan. 1988. “Discrete Spatial Choice and the Axiom of Independence from Irrelevant Alternatives.” *Socio-Economic Planning Sciences* 22 (6): 241–51. [https://doi.org/10.1016/0038-0121\(88\)90006-7](https://doi.org/10.1016/0038-0121(88)90006-7).
- Head, Keith, John Ries, and Deborah Swenson. 1995. “Agglomeration Benefits and Location Choice: Evidence from Japanese Manufacturing Investments in the United States.” *Journal of International Economics* 38 (3): 223–47. [https://doi.org/10.1016/0022-1996\(94\)01351-R](https://doi.org/10.1016/0022-1996(94)01351-R).
- Hensher, D. A., and T. T. Ton. 2000. “A Comparison of the Predictive Potential of Artificial Neural Networks and Nested Logit Models for Commuter Mode Choice.” *Transportation Research Part E-Logistics and Transportation Review* 36 (3): 155–72. [https://doi.org/10.1016/S1366-5545\(99\)00030-7](https://doi.org/10.1016/S1366-5545(99)00030-7).
- Hensher, David A., and William H. Greene. 2002. “Specification and Estimation of the Nested Logit Model: Alternative Normalisations.” *Transportation Research Part B: Methodological* 36 (1): 1–17. [https://doi.org/10.1016/S0191-2615\(00\)00035-7](https://doi.org/10.1016/S0191-2615(00)00035-7).
- . 2003. “The Mixed Logit Model: The State of Practice.” *Transportation* 30 (2): 133–76. <https://doi.org/10.1023/A:1022558715350>.
- Hensher, David A., William H. Greene, and Caspar G. Chorus. 2013. “Random Regret Minimization or Random Utility Maximization: An Exploratory Analysis in the Context of Automobile Fuel Choice.”

- Journal of Advanced Transportation* 47 (7): 667–78. <https://doi.org/10.1002/atr.188>.
- Hensher, David A., John M. Rose, and William H. Greene. 2005. *Applied Choice Analysis: A Primer*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511610356>.
- Hess, Stephane, and Caspar Chorus. 2015. “Utility Maximisation and Regret Minimisation: A Mixture of a Generalisation.” In, 31–48. <https://doi.org/10.1108/978-1-78441-072-820151006>.
- Hess, Stephane, and John M. Rose. 2008. “Asymmetrical Preference Formation in Willingness to Pay Estimates in Discrete Choice Models.” *Transportation Research Part E*, 847–63.
- Heukelom, Floris. 2015. “Prospect Theory.” In *International Encyclopedia of the Social & Behavioral Sciences*, 261–69. Elsevier. <https://doi.org/10.1016/B978-0-08-097086-8.03193-7>.
- Hilhorst, Loes, Jip van der Stappen, Joran Lokkerbol, Mickaël Hiligsmann, Anna H. Risseeuw, and Bea G. Tiemens. 2022. “Patients’ and Psychologists’ Preferences for Feedback Reports on Expected Mental Health Treatment Outcomes: A Discrete-Choice Experiment.” *Administration and Policy in Mental Health and Mental Health Services Research* 49 (5): 707–21. <https://doi.org/10.1007/s10488-022-01194-2>.
- Hillel, Tim, Michel Bierlaire, Mohammed Z. E. B. Elshafie, and Ying Jin. 2021. “A Systematic Review of Machine Learning Classification Methodologies for Modelling Passenger Mode Choice.” *Journal of Choice Modelling* 38 (March): 100221. <https://doi.org/10.1016/j.jocm.2020.100221>.
- Hillel, Tim, and Emma Frejinger. 2021. “Dynamic Choice Models,” March, 24.
- Hole, Arne Risa. 2007. “A Comparison of Approaches to Estimating Confidence Intervals for Willingness to Pay Measures.” *Health Economics* 16 (8): 827–40. <https://doi.org/10.1002/hec.1197>.
- Hornik, Kurt. 1991. “Approximation Capabilities of Multilayer Feedforward Networks.” *Neural Networks* 4 (2): 251–57. [https://doi.org/10.1016/0893-6080\(91\)90009-T](https://doi.org/10.1016/0893-6080(91)90009-T).
- Horowitz, Joel. 1982. “Specification Tests for Probabilistic Choice Models.” *Transportation Research Part A: General* 16 (5): 383–94. [https://doi.org/10.1016/0191-2607\(82\)90066-8](https://doi.org/10.1016/0191-2607(82)90066-8).
- Hrnjic, Emir, and Nikodem Tomczak. 2019. “Machine Learning and Behavioral Economics for Personalized Choice Architecture.” *arXiv Preprint arXiv:1907.02100*. <https://arxiv.org/abs/1907.02100>.
- Hruschka, Harald, Werner Fettes, and Markus Probst. 2001. “Analyzing Purchase Data by a Neural Net Extension of the Multinomial Logit Model.” In *Artificial Neural Networks — ICANN 2001*, edited by Georg Dorffner, Horst Bischof, and Kurt Hornik, 790–95. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer. https://doi.org/10.1007/3-540-44668-0_110.
- Huber, Joel, and Klaus Zwerina. 1996. “The Importance of Utility Balance in Efficient Choice Designs.” *Journal of Marketing Research* 33 (3): 307–17. <https://doi.org/10.1177/002224379603300305>.
- Huls, Samare P. I., and Esther W. de Bekker-Grob. 2022. “Can Healthcare Choice Be Predicted Using Stated Preference Data? The Role of Model Complexity in a Discrete Choice Experiment about Colorectal Cancer Screening.” *Social Science & Medicine* 315 (December): 115530. <https://doi.org/10.1016/j.socscimed.2022.115530>.
- Hurtubia, Ricardo, My Hang Nguyen, Aurélie Glerum, and Michel Bierlaire. 2014. “Integrating Psychometric Indicators in Latent Class Choice Models.” *Transportation Research Part A: Policy and Practice* 64 (June): 135–46. <https://doi.org/10.1016/j.tra.2014.03.010>.
- Hynes, Stephen, Claire W. Armstrong, Bui Bich Xuan, Isaac Ankamah-Yeboah, Katherine Simpson, Robert Tinch, and Adriana Ressurreição. 2021. “Have Environmental Preferences and Willingness to Pay Remained Stable Before and During the Global Covid-19 Shock?” *Ecological Economics* 189: 107142. <https://doi.org/10.1016/j.ecolecon.2021.107142>.
- Ilahi, Anugrah, Prawira F. Belgiawan, Milos Balac, and Kay W. Axhausen. 2021. “Understanding Travel

- and Mode Choice with Emerging Modes; a Pooled SP and RP Model in Greater Jakarta, Indonesia.” *Transportation Research Part A: Policy and Practice* 150 (August): 398–422. <https://doi.org/10.1016/j.tra.2021.06.023>.
- Ish-Horowicz, Jonathan, Dana Udwin, Seth Flaxman, Sarah Filippi, and Lorin Crawford. 2019. “Interpreting Deep Neural Networks Through Variable Importance.”
- Jang, Sunghoon, Soora Rasouli, and Harry Timmermans. 2017. “Bias in Random Regret Models Due to Measurement Error: Formal and Empirical Comparison with Random Utility Model.” *Transportmetrica A: Transport Science* 13 (5): 405–34. <https://doi.org/10.1080/23249935.2017.1285366>.
- Janssen, Meike, and Ulrich Hamm. 2012. “Product Labelling in the Market for Organic Food: Consumer Preferences and Willingness-to-Pay for Different Organic Certification Logos.” *Food Quality and Preference* 25 (1): 9–22. <https://doi.org/10.1016/j.foodqual.2011.12.004>.
- Japkowicz, Nathalie, and Mohak Shah. 2011. *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511921803>.
- Jong, Valentijn M. T., Marinus J. C. Eijkemans, Ben Calster, Dirk Timmerman, Karel G. M. Moons, Ewout W. Steyerberg, and Maarten Smeden. 2019. “Sample Size Considerations and Predictive Performance of Multinomial Logistic Prediction Models.” *Statistics in Medicine* 38 (9): 1601–19. <https://doi.org/10.1002/sim.8063>.
- Kahneman, Daniel, and Amos Tversky. 1979. “Prospect Theory: An Analysis of Decision Under Risk.” *Econometrica* 47 (2): 263–91.
- . 2012. “Prospect Theory: An Analysis of Decision Under Risk.” In *Handbook of the Fundamentals of Financial Decision Making*, Volume 4:99–127. World Scientific Handbook in Financial Economics Series. WORLD SCIENTIFIC. https://doi.org/10.1142/9789814417358_0006.
- Kahng, Minsuk, Pierre Y. Andrews, Aditya Kalro, and Duen Horng Chau. 2018. “ActiVis: Visual Exploration of Industry-Scale Deep Neural Network Models.” *IEEE Transactions on Visualization and Computer Graphics* 24 (1): 88–97. <https://doi.org/10.1109/TVCG.2017.2744718>.
- Karlaftis, M. G., and E. I. Vlahogianni. 2011. “Statistical Methods Versus Neural Networks in Transportation Research: Differences, Similarities and Some Insights.” *Transportation Research Part C: Emerging Technologies* 19 (3): 387–99. <https://doi.org/10.1016/j.trc.2010.10.004>.
- Kessels, Roselinde, Bradley Jones, Peter Goos, and Martina Vandebroek. 2011. “The Usefulness of Bayesian Optimal Designs for Discrete Choice Experiments.” *Applied Stochastic Models in Business and Industry* 27 (3): 173–88. <https://doi.org/10.1002/asmb.906>.
- Khan, Nazmul Arefin, Muhammad Ahsanul Habib, and Shaila Jamal. 2020. “Effects of Smartphone Application Usage on Mobility Choices.” *Transportation Research Part A: Policy and Practice* 132 (February): 932–47. <https://doi.org/10.1016/j.tra.2019.12.024>.
- Kingma, Diederik P., and Jimmy Ba. 2017. “Adam: A Method for Stochastic Optimization.” arXiv. <https://doi.org/10.48550/arXiv.1412.6980>.
- Kotsiantis, Sotiris, I. Zaharakis, and P. Pintelas. 2006. “Machine Learning: A Review of Classification and Combining Techniques.” *Artificial Intelligence Review* 26 (November): 159–90. <https://doi.org/10.1007/s10462-007-9052-3>.
- Kreutz, Clemens, and Jens Timmer. 2009. “Systems Biology: Experimental Design.” *The FEBS Journal* 276 (4): 923–42. <https://doi.org/10.1111/j.1742-4658.2008.06843.x>.
- Krueger, Rico, Taha H. Rashidi, and John M. Rose. 2016. “Preferences for Shared Autonomous Vehicles.” *Transportation Research Part C: Emerging Technologies* 69 (August): 343–55. <https://doi.org/10.1016/j.trc.2016.06.015>.

- Kuhfeld, Warren F., Randall D. Tobias, and Mark Garratt. 1994. "Efficient Experimental Design with Marketing Research Applications." *Journal of Marketing Research* 31 (4): 545–57. <https://doi.org/10.2307/3151882>.
- Lancsar, Emily, Denzil G. Fiebig, and Arne Risa Hole. 2017. "Discrete Choice Experiments: A Guide to Model Specification, Estimation and Software." *PharmacoEconomics* 35 (7): 697–716. <https://doi.org/10.1007/s40273-017-0506-4>.
- Lederrey, Gael, Virginie Lurkin, and Michel Bierlaire. 2018. "SNM: Stochastic Newton Method for Optimization of Discrete Choice Models." In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 3199–3204. <https://doi.org/10.1109/ITSC.2018.8569539>.
- Lederrey, Gael, Virginie Lurkin, Tim Hillel, and Michel Bierlaire. 2021. "Estimation of Discrete Choice Models with Hybrid Stochastic Adaptive Batch Size Algorithms." *Journal of Choice Modelling* 38 (March): 100226. <https://doi.org/10.1016/j.jocm.2020.100226>.
- Lee, Dongwoo, Sybil Derrible, and Francisco Camara Pereira. 2018. "Comparison of Four Types of Artificial Neural Network and a Multinomial Logit Model for Travel Mode Choice Modeling." *Transportation Research Record* 2672 (49): 101–12. <https://doi.org/10.1177/0361198118796971>.
- Leong, Waiyan, and David A. Hensher. 2015. "Contrasts of Relative Advantage Maximisation with Random Utility Maximisation and Regret Minimisation." *Journal of Transport Economics and Policy (JTEP)* 49 (1): 167–86.
- Lewbel, Arthur. 2019. "The Identification Zoo: Meanings of Identification in Econometrics." *Journal of Economic Literature* 57 (4): 835–903.
- Lewin Kurt. 1935. *A Dynamic Theory Of Personality*.
- Li, Zheng, Siwen Wang, Wei Shan Chin, Luke E. Achenie, and Hongliang Xin. 2017. "High-Throughput Screening of Bimetallic Catalysts Enabled by Machine Learning." *Journal of Materials Chemistry A* 5 (46): 24131–38. <https://doi.org/10.1039/C7TA01812F>.
- Lipovetsky, Stan. 2018. "Quantum Paradigm of Probability Amplitude and Complex Utility in Entangled Discrete Choice Modeling." *Journal of Choice Modelling* 27 (June): 62–73. <https://doi.org/10.1016/j.jocm.2017.10.003>.
- Lipton, Zachary C. 2017. "The Mythos of Model Interpretability." *arXiv:1606.03490 [Cs, Stat]*, March. <https://arxiv.org/abs/1606.03490>.
- Liu, Yan, and Tian Xie. 2019. "Machine Learning Versus Econometrics: Prediction of Box Office." *Applied Economics Letters* 26 (2): 124–30. <https://doi.org/10.1080/13504851.2018.1441499>.
- Lorenzo Varela, Juan Manuel. 2018. "Parameter Bias in Misspecified Hybrid Choice Models: An Empirical Study." *Transportation Research Procedia*, XIII Conference on Transport Engineering, CIT2018, 33 (January): 99–106. <https://doi.org/10.1016/j.trpro.2018.10.081>.
- Louviere, Jordan, David Hensher, and Joffre Swait. 2000. *Stated Choice Methods: Analysis and Application*. Vol. 17. <https://doi.org/10.1017/CBO9780511753831.008>.
- Louviere, Jordan, and Harry Timmermans. 1990. "A REVIEW OF RECENT ADVANCES IN DECOMPOSITIONAL PREFERENCE AND CHOICE MODELS." *Tijdschrift Voor Economische En Sociale Geografie* 81 (3): 214–24. <https://doi.org/10.1111/j.1467-9663.1990.tb00772.x>.
- Luce, R Duncan. 1957. "A Theory of Individual Choice Behavior." COLUMBIA UNIV NEW YORK BUREAU OF APPLIED SOCIAL RESEARCH.
- . 1977. "The Choice Axiom After Twenty Years." *Journal of Mathematical Psychology* 15 (3): 215–33.
- Ludwig, Kristina, Juan M. Ramos-Goñi, Mark Oppe, Simone Kreimeier, and Wolfgang Greiner. 2021.

- “To What Extent Do Patient Preferences Differ From General Population Preferences?” *Value in Health* 24 (9): 1343–49. <https://doi.org/10.1016/j.jval.2021.02.012>.
- Lusk, Jayson L., Jutta Roosen, and John A. Fox. 2003. “Demand for Beef from Cattle Administered Growth Hormones or Fed Genetically Modified Corn: A Comparison of Consumers in France, Germany, the United Kingdom, and the United States.” *American Journal of Agricultural Economics* 85 (1): 16–29. <https://doi.org/10.1111/1467-8276.00100>.
- Lusk, Jayson L., and Ted C. Schroeder. 2004. “Are Choice Experiments Incentive Compatible? A Test with Quality Differentiated Beef Steaks.” *American Journal of Agricultural Economics* 86 (2): 467–82. <https://doi.org/10.1111/j.0092-5853.2004.00592.x>.
- Macal, Charles, and Michael North. 2014. “Introductory Tutorial: Agent-based Modeling and Simulation.” In *Proceedings of the Winter Simulation Conference 2014*, 6–20. IEEE.
- Magaldi, Danielle, and Matthew Berler. 2018. “Semi-Structured Interviews.” In *Encyclopedia of Personality and Individual Differences*, edited by Virgil Zeigler-Hill and Todd K. Shackelford, 1–6. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-28099-8_857-1.
- Malone, Trey, and Jayson L Lusk. 2018. “A Simple Diagnostic Measure of Inattention Bias in Discrete Choice Models.” *European Review of Agricultural Economics* 45 (3): 455–62. <https://doi.org/10.1093/erae/jby005>.
- Manski, Charles F., and Steven R. Lerman. 1977. “The Estimation of Choice Probabilities from Choice Based Samples.” *Econometrica* 45 (8): 1977–88. <https://doi.org/10.2307/1914121>.
- “Mathematical Model.” 2021. Encyclopedia. *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Mathematical_model.
- . 2023. Encyclopedia. *Encyclopedia Britannica*. <https://www.britannica.com/science/mathematical-model>.
- McCullagh, Peter. 2002. “What Is a Statistical Model?” *The Annals of Statistics* 30 (5): 1225–1310. <https://doi.org/10.1214/aos/1035844977>.
- McCulloch, Warren S., and Walter Pitts. 1943. “A Logical Calculus of the Ideas Immanent in Nervous Activity.” *The Bulletin of Mathematical Biophysics* 5 (4): 115–33. <https://doi.org/10.1007/BF02478259>.
- McFadden, Daniel. 1974. “The Measurement of Urban Travel Demand.” *Journal of Public Economics* 3 (4): 303–28. [https://doi.org/10.1016/0047-2727\(74\)90003-6](https://doi.org/10.1016/0047-2727(74)90003-6).
- . 1981. “Econometric Models of Probabilistic Choice.” *Structural Analysis of Discrete Data with Econometric Applications* 198272.
- . 1987. “Regression-Based Specification Tests for the Multinomial Logit Model.” *Journal of Econometrics* 34 (1): 63–82. [https://doi.org/10.1016/0304-4076\(87\)90067-4](https://doi.org/10.1016/0304-4076(87)90067-4).
- . 2001. “Economic Choices.” *The American Economic Review* 91 (3): 351–78. <https://www.jstor.org/stable/2677869>.
- McFadden, Daniel, and Kenneth Train. 2000. “Mixed MNL Models for Discrete Response.” *Journal of Applied Econometrics* 15 (5): 447–70. [https://doi.org/10.1002/1099-1255\(200009/10\)15:5%3C447::AID-JAE570%3E3.0.CO;2-1](https://doi.org/10.1002/1099-1255(200009/10)15:5%3C447::AID-JAE570%3E3.0.CO;2-1).
- McLean, Iain. 1995. “Independence of Irrelevant Alternatives Before Arrow.” *Mathematical Social Sciences* 30 (2): 107–26. [https://doi.org/10.1016/0165-4896\(95\)00784-J](https://doi.org/10.1016/0165-4896(95)00784-J).
- Merkert, Rico, Michiel C. J. Bliemer, and Muhammad Fayyaz. 2022. “Consumer Preferences for Innovative and Traditional Last-Mile Parcel Delivery.” *International Journal of Physical Distribution & Logistics Management* 52 (3): 261–84. <https://doi.org/10.1108/IJPDLM-01-2021-0013>.
- Michaud, Celine, Daniel Llerena, and Iragael Joly. 2012. “Willingness to Pay for Environmental At-

- tributes of Non-Food Agricultural Products: A Real Choice Experiment.” *European Review of Agricultural Economics* 40 (2): 313–29. <https://doi.org/10.1093/erae/jbs025>.
- Mihailova, Darja, Iljana Schubert, Adan L. Martinez-Cruz, Adam X. Hearn, and Annika Sohre. 2022. “Preferences for Configurations of Positive Energy Districts – Insights from a Discrete Choice Experiment on Swiss Households.” *Energy Policy* 163 (April): 112824. <https://doi.org/10.1016/j.enpol.2022.112824>.
- Minsky, Marvin, and Seymour A. Papert. 1969. *Perceptrons: An Introduction to Computational Geometry*. Cambridge, MA, USA: MIT Press.
- “Model.” 2022. Encyclopedia. *Wikipedia*. <https://en.wikipedia.org/w/index.php?title=Model>.
- . 2023. Encyclopedia. *Encyclopedia Britannica*. <https://www.britannica.com/dictionary/model>.
- Mohammadi, Farzaneh, Hamidreza Pourzamani, Hossein Karimi, Maryam Mohammadi, Mohammad Mohammadi, Nahid Ardalan, Roya Khoshravesh, et al. 2021. “Artificial Neural Network and Logistic Regression Modelling to Characterize COVID-19 Infected Patients in Local Areas of Iran.” *Biomedical Journal* 44 (3): 304–316. <https://doi.org/10.1016/j.bj.2021.02.006>.
- Molina, Mario, and Filiz Garip. 2019. “Machine Learning for Sociology.” *Annual Review of Sociology* 45 (1): 27–45. <https://doi.org/10.1146/annurev-soc-073117-041106>.
- Mouter, Niek, Marion Collewet, G. Ardine de Wit, Adrienne Rotteveel, Mattijs S. Lambooi, and Roselinde Kessels. 2021. “Societal Effects Are a Major Factor for the Uptake of the Coronavirus Disease 2019 (COVID-19) Digital Contact Tracing App in The Netherlands.” *Value in Health* 24 (5): 658–67. <https://doi.org/10.1016/j.jval.2021.01.001>.
- Mueller, Anne E., and Daniel L. Segal. 2015. “Structured Versus Semistructured Versus Unstructured Interviews.” In *The Encyclopedia of Clinical Psychology*, edited by Robin L. Cautin and Scott O. Lilienfeld, 1–7. Hoboken, NJ, USA: John Wiley & Sons, Inc. <https://doi.org/10.1002/9781118625392.wbecp069>.
- Mühlbacher, Axel C., and Susanne Bethge. 2015. “Patients’ Preferences: A Discrete-Choice Experiment for Treatment of Non-Small-Cell Lung Cancer.” *The European Journal of Health Economics* 16 (6): 657–70. <https://doi.org/10.1007/s10198-014-0622-4>.
- Mullainathan, Sendhil, and Jann Spiess. 2017. “Machine Learning: An Applied Econometric Approach.” *Journal of Economic Perspectives* 31 (2): 87–106. <https://doi.org/10.1257/jep.31.2.87>.
- Munizaga, Marcela A., and Ricardo Alvarez-Daziano. 2005. “Testing Mixed Logit and Probit Models by Simulation.” *Transportation Research Record* 1921 (1): 53–62. <https://doi.org/10.1177/0361198105192100107>.
- Munizaga, Marcela A, and Ricardo Alvarez-Daziano. 2001. “Mixed Logit Vs. Nested Logit and Probit Models.”
- Nagasubramanian, Koushik, Sarah Jones, Asheesh K. Singh, Arti Singh, Baskar Ganapathysubramanian, and Soumik Sarkar. 2018. “Explaining Hyperspectral Imaging Based Plant Disease Identification: 3D CNN and Saliency Maps.” arXiv. <https://doi.org/10.48550/arXiv.1804.08831>.
- Nash, Andrew, Ulrich Weidmann, Stefan Buchmueller, and Markus Rieder. 2007. “Assessing Feasibility of Transport Megaprojects: Swissmetro European Market Study.” *Transportation Research Record* 1995 (1): 17–26. <https://doi.org/10.3141/1995-03>.
- Nash, John F. 1950. “The Bargaining Problem.” *Econometrica* 18 (2): 155–62. <https://doi.org/10.2307/1907266>.
- Ndebele, Tom, Dan Marsh, and Riccardo Scarpa. 2019. “Consumer Switching in Retail Electricity Markets: Is Price All That Matters?” *Energy Economics* 83 (September): 88–103. <https://doi.org/10>

- .1016/j.eneco.2019.06.012.
- Nerella, Sriharsha, and Chandra R Bhat. 2004. "Numerical Analysis of Effect of Sampling of Alternatives in Discrete Choice Models." *Transportation Research Record* 1894 (1): 11–19.
- Newman, Jeffrey P., Mark E. Ferguson, and Laurie A. Garrow. 2013. "Estimating GEV Models with Censored Data." *Transportation Research Part B: Methodological* 58 (December): 170–84. <https://doi.org/10.1016/j.trb.2013.09.002>.
- Novikoff, A. B. 1962. "On Convergence Proofs on Perceptrons." In *Proceedings of the Symposium on the Mathematical Theory of Automata*, 12:615–22. New York, NY, USA: Polytechnic Institute of Brooklyn.
- Ojeda-Cabral, Manuel, Stephane Hess, and Richard Batley. 2018. "Understanding Valuation of Travel Time Changes: Are Preferences Different Under Different Stated Choice Design Settings?" *Transportation* 45 (1): 1–21. <https://doi.org/10.1007/s11116-016-9716-4>.
- Omrani, Hichem, Fahed Abdallah, Omar Charif, and Nicholas T. Longford. 2015. "Multi-Label Class Assignment in Land-Use Modelling." *International Journal of Geographical Information Science* 29 (6): 1023–41. <https://doi.org/10.1080/13658816.2015.1008004>.
- Ortega, David L., H. Holly Wang, Laping Wu, and Nicole J. Olynk. 2011. "Modeling Heterogeneity in Consumer Preferences for Select Food Safety Attributes in China." *Food Policy* 36 (2): 318–24. <https://doi.org/10.1016/j.foodpol.2010.11.030>.
- Ortelli, Nicola, Tim Hillel, Francisco C. Pereira, Matthieu de Lapparent, and Michel Bierlaire. 2021. "Assisted Specification of Discrete Choice Models." *Journal of Choice Modelling* 39 (June): 100285. <https://doi.org/10.1016/j.jocm.2021.100285>.
- Paredes, Miguel, Erik Hemberg, Una-May O'Reilly, and Chris Zegras. 2017. "Machine Learning or Discrete Choice Models for Car Ownership Demand Estimation and Prediction?" In *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, 780–85. <https://doi.org/10.1109/MTITS.2017.8005618>.
- Parkes, Alison, Ade Kearns, and Rowland Atkinson. 2002. "What Makes People Dissatisfied with Their Neighbourhoods?" *Urban Studies* 39 (13): 2413–38. <https://doi.org/10.1080/0042098022000027031>.
- Parkin, John, Mark Wardman, and Matthew Page. 2008. "Estimation of the Determinants of Bicycle Mode Share for the Journey to Work Using Census Data." *Transportation* 35 (1): 93–109. <https://doi.org/10.1007/s11116-007-9137-5>.
- Patty, John W., and Elizabeth Maggie Penn. 2019. "A Defense of Arrow's Independence of Irrelevant Alternatives." *Public Choice* 179 (1-2): 145–64. <https://doi.org/10.1007/s11127-018-0604-7>.
- Paul, Justin, Weng Marc Lim, Aron O'Cass, Andy Wei Hao, and Stefano Bresciani. 2021. "Scientific Procedures and Rationales for Systematic Literature Reviews (SPAR-4-SLR)." *International Journal of Consumer Studies* 45 (4): O1–16. <https://doi.org/10.1111/ijcs.12695>.
- Pigozzi, Gabriella, Alexis Tsoukiàs, and Paolo Viappiani. 2016. "Preferences in Artificial Intelligence." *Annals of Mathematics and Artificial Intelligence* 77 (3-4): 361–401. <https://doi.org/10.1007/s10472-015-9475-5>.
- Quiggin, John. 1981. "Risk Perception and the Analysis of Risk Attitudes*." *Australian Journal of Agricultural Economics* 25 (2): 160–69. <https://doi.org/10.1111/j.1467-8489.1981.tb00393.x>.
- Radner, Roy, and Jacob Marschak. 1954. "Note on Some Proposed Decision Criteria." Wiley.
- Ramon y Cajal, Santiago. 2002. "Structure et connexions des neurones Conférence Nobel faite à Stockholm le 12 Décembre 1906" 35 (4): 22.
- Ray, Paramesh. 1973. "Independence of Irrelevant Alternatives." *Econometrica* 41 (5): 987–91. <https://doi.org/10.2307/2326111>.

- [//doi.org/10.2307/1913820](https://doi.org/10.2307/1913820).
- Raychaudhuri, Samik. 2008. "Introduction to Monte Carlo Simulation." In *2008 Winter Simulation Conference*, 91–100. IEEE.
- Reckers-Droog, Vivian, Job van Exel, and Werner Brouwer. 2021. "Willingness to Pay for Health-Related Quality of Life Gains in Relation to Disease Severity and the Age of Patients." *Value in Health* 24 (8): 1182–92. <https://doi.org/10.1016/j.jval.2021.01.012>.
- Reed Johnson, F., Emily Lancsar, Deborah Marshall, Vikram Kilambi, Axel Mühlbacher, Dean A. Regier, Brian W. Bresnahan, Barbara Kanninen, and John F. P. Bridges. 2013. "Constructing Experimental Designs for Discrete-Choice Experiments: Report of the ISPOR Conjoint Analysis Experimental Design Good Research Practices Task Force." *Value in Health* 16 (1): 3–13. <https://doi.org/10.1016/j.jval.2012.08.2223>.
- Revelt, David, and Kenneth Train. 1998. "Mixed Logit with Repeated Choices: Households' Choices of Appliance Efficiency Level." *The Review of Economics and Statistics* 80 (4): 647–57. <https://doi.org/10.1162/003465398557735>.
- Roe, Robert M., Jermone R. Busemeyer, and James T. Townsend. 2001. "Multialternative Decision Field Theory: A Dynamic Connectionist Model of Decision Making." *Psychological Review* 108 (2): 370–92. <https://doi.org/10.1037/0033-295X.108.2.370>.
- Rose, John M., and Michiel C. J. Bliemer. 2013. "Sample Size Requirements for Stated Choice Experiments." *Transportation* 40 (5): 1021–41. <https://doi.org/10.1007/s11116-013-9451-z>.
- Rose, John M., Michiel C. J. Bliemer, David A. Hensher, and Andrew T. Collins. 2008. "Designing Efficient Stated Choice Experiments in the Presence of Reference Alternatives." *Transportation Research Part B: Methodological* 42 (4): 395–406. <https://doi.org/10.1016/j.trb.2007.09.002>.
- Rosenblatt, F. 1958. "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain." *Psychological Review* 65 (6): 386–408. <https://doi.org/10.1037/h0042519>.
- Rubinstein, Reuven Y, and Dirk P Kroese. 2016. *Simulation and the Monte Carlo Method*. Vol. 10. John Wiley & Sons.
- Rumelhart, David E., and James L. McClelland. 1987. "Learning Internal Representations by Error Propagation." In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*, 318–62. MIT Press.
- Saari, Donald G. 1999. "Explaining All Three-Alternative Voting Outcomes." *Journal of Economic Theory* 87 (2): 313–55. <https://doi.org/10.1006/jeth.1999.2541>.
- Sanchez, Cory. 2014. "Unstructured Interviews." In *Encyclopedia of Quality of Life and Well-Being Research*, edited by Alex C. Michalos, 6824–25. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-0753-5_3121.
- Savage, L. J. 1951. "The Theory of Statistical Decision." *Journal of the American Statistical Association*, March.
- Savolainen, Peter T., Fred L. Mannering, Dominique Lord, and Mohammed A. Quddus. 2011. "The Statistical Analysis of Highway Crash-Injury Severities: A Review and Assessment of Methodological Alternatives." *Accident Analysis & Prevention* 43 (5): 1666–76. <https://doi.org/10.1016/j.aap.2011.03.025>.
- Scaccia, Luisa, Edoardo Marcucci, and Valerio Gatta. 2023. "Prediction and Confidence Intervals of Willingness-to-Pay for Mixed Logit Models." *Transportation Research Part B: Methodological* 167 (January): 54–78. <https://doi.org/10.1016/j.trb.2022.11.007>.
- Scarpa, Riccardo, and John M. Rose. 2008. "Design Efficiency for Non-Market Valuation with Choice

- Modelling: How to Measure It, What to Report and Why.” *Australian Journal of Agricultural and Resource Economics* 52 (3): 253–82. <https://doi.org/10.1111/j.1467-8489.2007.00436.x>.
- Scarpa, Riccardo, and Ken Willis. 2010. “Willingness-to-Pay for Renewable Energy: Primary and Discretionary Choice of British Households’ for Micro-Generation Technologies.” *Energy Economics* 32 (1): 129–36. <https://doi.org/10.1016/j.eneco.2009.06.004>.
- Schmeidler, David. 1989. “Subjective Probability and Expected Utility Without Additivity.” *Econometrica* 57 (3): 571–87. <https://doi.org/10.2307/1911053>.
- Scholz, Michael, Verena Dorner, Markus Franz, and Oliver Hinz. 2015. “Measuring Consumers’ Willingness to Pay with Utility-Based Recommendation Systems.” *Decision Support Systems* 72: 60–71. <https://doi.org/10.1016/j.dss.2015.02.006>.
- Schulz, Eric, Maarten Speekenbrink, and David R Shanks. 2014. “Predict Choice: A Comparison of 21 Mathematical Models.” In *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 36.
- Sen, Amartya. 1983. “Liberty and Social Choice.” *The Journal of Philosophy* 80 (1): 5–28. <https://doi.org/10.2307/2026284>.
- . 2014. “Arrow and the Impossibility Theorem.” In *ARROW AND THE IMPOSSIBILITY THEOREM*, 29–42. Columbia University Press. <https://doi.org/10.7312/mask15328-003>.
- Seroussi, Dominique-Esther. 1995. “Heuristic Hypotheses in Problem Solving: An Example of Conceptual Issues about Scientific Procedures.” *Science Education* 79 (6): 595–609. <https://doi.org/10.1002/sce.3730790603>.
- Sfeir, Georges, Maya Abou-Zeid, Filipe Rodrigues, Francisco Camara Pereira, and Isam Kaysi. 2021. “Latent Class Choice Model with a Flexible Class Membership Component: A Mixture Model Approach.” *Journal of Choice Modelling* 41: 100320. <https://doi.org/10.1016/j.jocm.2021.100320>.
- Sfeir, Georges, Filipe Rodrigues, and Maya Abou-Zeid. 2022. “Gaussian Process Latent Class Choice Models.” *Transportation Research Part C: Emerging Technologies* 136 (March): 103552. <https://doi.org/10.1016/j.trc.2022.103552>.
- Sifringer, Brian, Virginie Lurkin, and Alexandre Alahi. 2020. “Enhancing Discrete Choice Models with Representation Learning.” *Transportation Research Part B: Methodological* 140 (October): 236–61. <https://doi.org/10.1016/j.trb.2020.08.006>.
- Simonson, Itamar. 1989. “Choice Based on Reasons: The Case of Attraction and Compromise Effects.” *Journal of Consumer Research* 16 (2): 158–74. <https://doi.org/10.1086/209205>.
- Small, Kenneth A., and Cheng Hsiao. 1985. “Multinomial Logit Specification Tests.” *International Economic Review* 26 (3): 619–27. <https://doi.org/10.2307/2526707>.
- Small, Kenneth A., Clifford Winston, and Jia Yan. 2005. “Uncovering the Distribution of Motorists’ Preferences for Travel Time and Reliability.” *Econometrica* 73 (4): 1367–82. <https://doi.org/10.1111/j.1468-0262.2005.00619.x>.
- Soekhai, Vikas, Esther W. de Bekker-Grob, Alan R. Ellis, and Caroline M. Vass. 2019. “Discrete Choice Experiments in Health Economics: Past, Present and Future.” *PharmacoEconomics* 37 (2): 201–26. <https://doi.org/10.1007/s40273-018-0734-2>.
- Sokolova, Marina, and Guy Lapalme. 2009. “A Systematic Analysis of Performance Measures for Classification Tasks.” *Information Processing & Management* 45 (4): 427–37. <https://doi.org/10.1016/j.ipm.2009.03.002>.
- “Statistical Model.” 2021. Encyclopedia. *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Statistical_model&
- Stern, Steven. Decembre 1997. “Simulation-Based Estimation.” *Journal of Economic Literature* 35 (4):

- 2006–39.
- Talvitie, Antti. 1972. “Comparison of Probabilistic Modal-Choice Models: Estimation Methods and System Inputs.” *Highway Research Record*, no. 392.
- Thiene, Mara, and Riccardo Scarpa. 2009. “Deriving and Testing Efficient Estimates of WTP Distributions in Destination Choice Models.” *Environmental and Resource Economics* 44 (3): 379–95. <https://doi.org/10.1007/s10640-009-9291-7>.
- Townsend, James T., Jerome R. Busemeyer, and C Izawa. 1989. “Current Issues in Cognitive Processes: Tulane Flowerree Symposium on Cognition.”
- Train, Kenneth. 2002. *Discrete Choice Methods with Simulation*. University of California, Berkeley: Cambridge University Press.
- . 2016. “Mixed Logit with a Flexible Mixing Distribution.” *Journal of Choice Modelling* 19 (June): 40–53. <https://doi.org/10.1016/j.jocm.2016.07.004>.
- Train, Kenneth E. 1998. “Recreation Demand Models with Taste Differences over People.” *Land Economics* 74 (2): 230. <https://doi.org/10.2307/3147053>.
- . 2008. “EM Algorithms for Nonparametric Estimation of Mixing Distributions.” *Journal of Choice Modelling* 1 (1): 40–69. [https://doi.org/10.1016/S1755-5345\(13\)70022-8](https://doi.org/10.1016/S1755-5345(13)70022-8).
- Train, Kenneth, and Melvyn Weeks. 2005. “Discrete Choice Models in Preference Space and Willingness-to-Pay Space.” In *Applications of Simulation Methods in Environmental and Resource Economics*, edited by Riccardo Scarpa and Anna Alberini, 1–16. The Economics of Non-Market Goods and Resources. Dordrecht: Springer Netherlands. https://doi.org/10.1007/1-4020-3684-1_1.
- Tsouros, Ioannis, Athena Tsirimpa, Ioanna Pagoni, and Amalia Polydoropoulou. 2021. “MaaS Users: Who They Are and How Much They Are Willing-to-Pay.” *Transportation Research Part A: Policy and Practice* 148 (June): 470–80. <https://doi.org/10.1016/j.tra.2021.04.016>.
- Tyrinopoulos, Yannis, and Constantinos Antoniou. 2008. “Public Transit User Satisfaction: Variability and Policy Implications.” *Transport Policy* 15 (4): 260–72. <https://doi.org/10.1016/j.tranpol.2008.06.002>.
- van den Broek-Altenburg, Eline, and Adam Atherly. 2020. “Using Discrete Choice Experiments to Measure Preferences for Hard to Observe Choice Attributes to Inform Health Policy Decisions.” *Health Economics Review* 10 (1): 1–8. <https://doi.org/10.1186/s13561-020-00276-x>.
- Varian, Hal R. 1994. “What Use Is Economic Theory?” 9401001. *Method and Hist of Econ Thought*. University Library of Munich, Germany.
- . 2014. “Big Data: New Tricks for Econometrics.” *Journal of Economic Perspectives* 28 (2): 3–28. <https://doi.org/10.1257/jep.28.2.3>.
- Vicente-Molina, María Azucena, Ana Fernández-Sáinz, and Julen Izagirre-Olaizola. 2013. “Environmental Knowledge and Other Variables Affecting Pro-Environmental Behaviour: Comparison of University Students from Emerging and Advanced Countries.” *Journal of Cleaner Production*, Special Volume: Green Universities and Environmental Higher Education for Sustainable Development in China and Other Emerging Countries, 61 (December): 130–38. <https://doi.org/10.1016/j.jclepro.2013.05.015>.
- Vij, Akshay, and Rico Krueger. 2017. “Random Taste Heterogeneity in Discrete Choice Models: Flexible Nonparametric Finite Mixture Distributions.” *Transportation Research Part B: Methodological* 106 (December): 76–101. <https://doi.org/10.1016/j.trb.2017.10.013>.
- Vij, Akshay, and Joan L. Walker. 2016. “How, When and Why Integrated Choice and Latent Variable Models Are Latently Useful.” *Transportation Research Part B: Methodological* 90 (August): 192–217.

- <https://doi.org/10.1016/j.trb.2016.04.021>.
- Vijayakumar, Ranjith, and Mike W.-L. Cheung. 2019. "Assessing Replicability of Machine Learning Results: An Introduction to Methods on Predictive Accuracy in Social Sciences." *Social Science Computer Review* 34 (1): 0894439319888445. <https://doi.org/10.1177/0894439319888445>.
- Vitetta, Antonino. 2016. "A Quantum Utility Model for Route Choice in Transport Systems." *Travel Behaviour and Society* 3: 29–37.
- Von Neumann, John, and Oskar Morgenstern. 1947. *Theory of Games and Economic Behavior, 2nd Rev. Ed.* Theory of Games and Economic Behavior, 2nd Rev. Ed. Princeton, NJ, US: Princeton University Press.
- Walker, Joan L., Yanqiao Wang, Mikkel Thorhauge, and Moshe Ben-Akiva. 2018. "D-Efficient or Deficient? A Robustness Analysis of Stated Choice Experimental Designs." *Theory and Decision* 84 (2): 215–38. <https://doi.org/10.1007/s11238-017-9647-3>.
- Walker, Joan, and Moshe Ben-Akiva. 2002. "Generalized Random Utility Model." *Mathematical Social Sciences* 43 (3): 303–43.
- Walrave, Michel, Cato Waeterloos, and Koen Ponnet. 2020. "Adoption of a Contact Tracing App for Containing COVID-19: A Health Belief Model Approach." *JMIR Public Health and Surveillance* 6 (3): e20572. <https://doi.org/10.2196/20572>.
- Wang, Qi, Yue Ma, Kun Zhao, and Yingjie Tian. 2020. "A Comprehensive Survey of Loss Functions in Machine Learning." *Annals of Data Science*, 1–26.
- Wang, Shenhao, Baichuan Mo, and Jinhua Zhao. 2020. "Deep Neural Networks for Choice Analysis: Architecture Design with Alternative-Specific Utility Functions." *Transportation Research Part C: Emerging Technologies* 112 (March): 234–51. <https://doi.org/10.1016/j.trc.2020.01.012>.
- Wang, Shenhao, Qingyi Wang, Nate Bailey, and Jinhua Zhao. 2021. "Deep Neural Networks for Choice Analysis: A Statistical Learning Theory Perspective." *Transportation Research Part B: Methodological* 148 (June): 60–81. <https://doi.org/10.1016/j.trb.2021.03.011>.
- Wang, Shenhao, Qingyi Wang, and Jinhua Zhao. 2018. "Deep Neural Networks for Choice Analysis: Extracting Complete Economic Information for Interpretation."
- . 2020. "Deep Neural Networks for Choice Analysis: Extracting Complete Economic Information for Interpretation." *Transportation Research Part C: Emerging Technologies* 118 (September): 102701. <https://doi.org/10.1016/j.trc.2020.102701>.
- Welleck, Sean, Zixin Yao, Yu Gai, Jialin Mao, Zheng Zhang, and Kyunghyun Cho. 2017. "Loss Functions for Multiset Prediction." *arXiv Preprint arXiv:1711.05246*. <https://arxiv.org/abs/1711.05246>.
- Widrow, Bernard, and Marcian E. Hoff. 1960. "Adaptive Switching Circuits." *1960 IRE WESCON Convention Record, Part 4*, 96–104.
- Willems, J. C., and J. W. Polderman. 1998. *Introduction to Mathematical Systems Theory: A Behavioral Approach*. Texts in Applied Mathematics. New York: Springer-Verlag. <https://doi.org/10.1007/978-1-4757-2953-5>.
- Williams, H. C. W. L., and J. D. Ortuzar. 1982. "Behavioural Theories of Dispersion and the Mis-specification of Travel Demand Models." *Transportation Research Part B: Methodological* 16 (3): 167–219. [https://doi.org/10.1016/0191-2615\(82\)90024-8](https://doi.org/10.1016/0191-2615(82)90024-8).
- Wong, R. C. P., Linchuan Yang, and W. Y. Szeto. 2021. "Wearable Fitness Trackers and Smartphone Pedometer Apps: Their Effect on Transport Mode Choice in a Transit-Oriented City." *Travel Behaviour and Society* 22 (January): 244–51. <https://doi.org/10.1016/j.tbs.2020.10.006>.
- Wooldridge, Jeffrey M. 2012. "Introductory Econometrics: A Modern Approach," 910.

-
- Yu, Jiangbo Gabriel, and R. Jayakrishnan. 2018. "A Quantum Cognition Model for Bridging Stated and Revealed Preference." *Transportation Research Part B: Methodological* 118 (December): 263–80. <https://doi.org/10.1016/j.trb.2018.10.014>.
- Yukalov, Vyacheslav I, and Didier Sornette. 2017. "Quantum Probabilities as Behavioral Probabilities." *Entropy* 19 (3): 112.
- Zeng, Minhui, Ming Zhong, and John Douglas Hunt. 2018. "Analysis of the Impact of Sample Size, Attribute Variance and Within-Sample Choice Distribution on the Estimation Accuracy of Multinomial Logit Models Using Simulated Data." *Journal of Systems Science and Systems Engineering* 27 (6): 771–89. <https://doi.org/10.1007/s11518-018-5359-7>.
- Zhao, Xilei, Xiang Yan, Alan Yu, and Pascal Van Hentenryck. 2020. "Prediction and Behavioral Analysis of Travel Mode Choice: A Comparison of Machine Learning and Logit Models." *Travel Behaviour and Society* 20 (July): 22–35. <https://doi.org/10.1016/j.tbs.2020.02.003>.

A. Bibliometric study

This appendix introduces the bibliography analysis, based on a bibliometric study conducted in between February and May 2022. The study focuses on the exploration of *choice modelling* literature in economics. We explore the different use-cases there exist within the economics discipline for the choice modelling methodology. In this task we focus on the general research directions and subdomains within the economics discipline, as well as the closely related fields. For each of the identified subdomains a more in-depth analysis is performed. The most prominent works are explored more in detail for each of the application cases.

The bibliometric study delves into the intricate landscape of discrete choice analysis in social sciences and economics in particular. The work acknowledges the terminological ambiguity prevalent across diverse disciplines and fields, addressing it through a systematic bibliometric review on the Web of Science database. The strategic use of general and specific keywords, forms the basis for dataset construction and refinement. The adoption of VOSviewer for bibliometric analysis, despite computational constraints, ensures a robust and efficient examination of citation patterns, publishers, and disciplines. This methodical exploration offers a detailed exploration of trends, influential works, and applications, contributing significantly to the understanding of discrete choice analysis in economics.

A.1. Motivation and research objectives

Even though this background description already offers a fair enough overview of the problematic, we should dig deeper into the available bibliography. In order to obtain a more clear picture of the existing techniques and methodology we propose an exhaustive bibliometric study.

There exists an extreme ambiguity in the terminology between the different fields and disciplines. From one application domain, to another the understanding and perception of different element's meaning differs. Seemingly universal concept may have different connotation depending on the particular convention existing within the community. This fact rises to the extreme the complexity of the bibliography review. On the one hand we can-not use already known economics specific keywords, due to the risk of omission of some of the potential applications of the discrete choice models in economics, psychology, sociology and marketing. On the other hand we are equally constrained in the usage of too general terms, because it will increase the number of non-relevant studies in our dataset.

Such complications bring us to the idea of conducting the bibliography review using multiple stages. First of all, we will use the most general keywords to limit our scope to the potentially relevant applications of the discrete choice models. Once the preliminary dataset complete, we will have to analyse the keywords and choose, which ones should be filtered. For example, we don't want to include in our final dataset the strictly technical studies on biology applications, while at the same time we may desire to preserve the entries on the economics of health. The next stage will encompass the more through bibliometric analysis of the collected data.

For the whole analysis we use *Web of Science (WoS)* database exclusively, without adding *Scopus* or *Google Scholar*. This fact can potentially limit our results. However, we assume that the most important works for the scientific community may be encountered throughout all the three indexes and our results should not be impacted in significant manner. We equally suppose that intermediary filters may have greater impact than untracked (and supposedly less popular) works.

A.2. Data collection and preliminary analysis

There exist many different available strategies to perform the preliminary analysis. The key idea at this stage is to reduce the size of our bibliographic dataset¹ in a manner to make it manageable. Evidently a dataset including all the publication across all the disciplines is not that useful, so we should find some entry point for our literature review. As it was previously stated the main focus of this work is to identify and explore the different use-cases of *discrete choice models* in *social sciences*. This objective immediately offers us two entry points: (1) by toolset, because we can refine our search using keywords related to different models and modelling techniques of *discrete choice* analysis, and (2) by discipline or domain of application, restricting the search by field to retain only *social and human sciences* (ex: *economics, sociology* and *psychology*).

Unfortunately both strategies have their own advantages and drawbacks. For example, let's take a look at one of the possible strategies: refining the literature by field of application. In this case we risk to end up with extremely large dataset including all the available publications related to the explored application fields. Seemingly exhaustive, this dataset risks to lack some of the potentially interesting

¹The number of bibliographic entries to consider in our analysis.

use-cases and applications. What is more, the WoS database offers the possibility to refine the search by discipline with extremely fine granularity. This means that we risk to obtain a rather limited dataset even considering the boundaries of searched disciplines. Finally, the results of such search will contain all the eventual techniques and approaches. Those may not be closely related to the discrete choice modelling and we have no means to assure that clustering algorithms of bibliometrics software will identify the clusters in a convenient for us manner. The same reasoning may be applied to another approach: refining the literature by modelling approach. In this case we may potentially obtain a sufficiently complete dataset of all the potential applications of the queried modelling techniques. Unfortunately, this dataset will surely include a number of applications that are used in *engineering* or *biology*, which are of little use for us.

For preliminary analysis it was decided to use the most encompassing key-words. Because our research focuses on the applications and use of discrete choice models in economics and annex sciences we decide to use keywords associated with discrete choice modelling. As said previously, this will inevitably include into our results some of the application fields that are of no interest for us. We expect to identify the keywords for those fields and exclude them on the latter stage of the analysis. For our first research we decide to focus on the *Logit* model. The *logistic regression* is a state of art tool in most interpretable discrete choice analysis tasks, which implies that it should be mentioned at least in the abstract of the relevant works. Unfortunately, such approach may miss some of the most advanced and recent works, where the advanced ML techniques are used without making any link to historical modelling techniques. At this point we can do very little, but to assume that such emerging studies are in minority and we should be able to capture the potential application domains regardless. At the first stage, this results in a query of type:

Logit OR Logistic Regression OR MNL

This query covers most of the potential model names and their abbreviations (ex: *Multinomial Logit* or *Mixed Logit*). This query alone results in 413037 matches in WoS database.

Among the others restriction imposed onto our search we include only *Articles* and exclude:

- Early access entries
- Proceedings papers
- Book chapters and Books
- Data papers
- Retracted papers and those that are with expression of a concern
- Articles announced for year 2023

With restrictions we obtain 380212 entries for the time of data collection procedure - 23 may 2022. This is an important precision, because the given above numbers include the publications made in first months of year 2022. Finally, we prefer to exclude the beginning of the year 2022 and focus our attention on the general trends in the literature for the last decade: from 2011 to 2021 (resulting in 283183 entries)².

Then, using developed toolset for automated data extraction from the WoS we proceed with the data collection. The implemented toolset uses R interface to the Selenium instance running inside a Docker

²Query link example for 2021

container for automated interactions with JavaScript powered web-interface of WoS. Due to limitation of WoS interface, which makes queries larger than 100000 entries impossible due to internal server error we prefer to collect data year by year. In order to limit ourselves to the trends in the recent literature, it was decided not to go beyond the 2011 year point in the preliminary exploration step. The number of entries for each year is given in Figure A.1. As we can see the overall yearly volume of publications increases with time, becoming in 2021 nearly four times the amount of 2011.

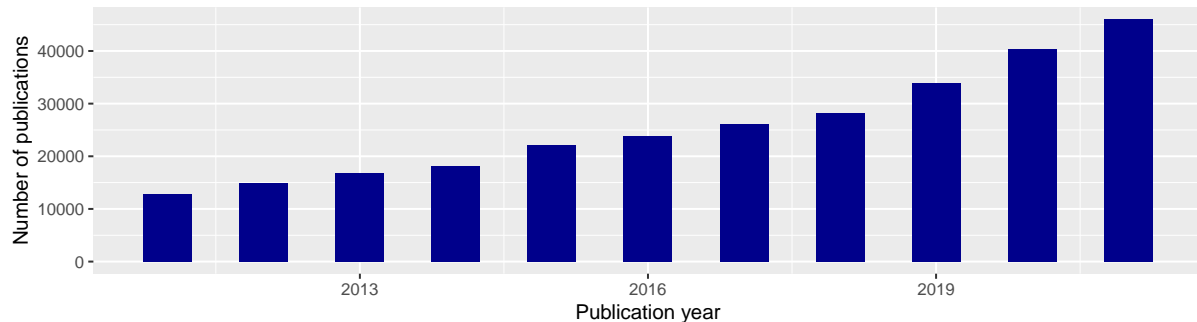


Figure A.1.: Number of publications by year (2011 - 2021)

The *VOSviewer* software is used to perform the analysis due to computational complexity of the tasks. Among the considered alternatives for this task an attempt was made to use *bibliometrix R* package, although due to the memory saturation³ this tool was discarded. The resulting datasets size, for data stored in WoS database compatible `.txt` files, varied from 79.2Mo (for 2011) to 314.5Mo (for 2021). The combined database amounts for 1.9Go of text data. In order to process this volume of information we increase the memory available for *VOSviewer* to 6gb and the stack size to 2Gb.

At this stage we use *VOSviewer* for aggregation purposes, because the software implementation allows us to perform text analysis on collected bibliographic entries. For each of the collected datasets (structured by year) we create a map based on text data. It is important to state at this point that the main interest for us at this point lies not within the textual map analysis and clustering, but in the possibility to analyse and synthesise the collected information. The datasets for each year contain several thousand thousands of keywords to be explored. For aggregated analysis purposes we establish an offset for morpheme appearance at 0.001% of total word count. Meaning, that if an item occurs less than 0.001% of total word count times it's discarded, because of insufficient popularity in the literature and thus insufficient representativeness of the literature state. Otherwise we face memory insufficiency during the analysis stage. Next, the *VOSviewer* further contracts the dataset leaving only 60% of remaining morphemes based on scoring algorithm. This procedure allows us to first of all filter the words that appear often enough to describe the general trends in the literature, as well as to discard at the same time the words that are too general and appear in every paper.

The constructed textual map is stored as `.txt` file and can then be imported and analysed using R software to get total appearances of the given words for the whole decade. This allows us to trace the trends in the keywords popularity, as well as to explore the overall representativeness of the different keywords to further narrow our research. In the Table A.1 we can observe the most popular keywords that appear in through the last decade in association with *Logit* model usage.

As we can see, these words may roughly be divided into two categories: (1) the medical terms and (2)

³On a device limited to 8Gb of RAM.

Table A.1.: Total number of keyword appearances (2011 - 2021)

| Label | Occurrences |
|--------------|-------------|
| model | 44657 |
| survey | 28502 |
| health | 21581 |
| surgery | 19474 |
| mortality | 18356 |
| self | 18332 |
| education | 17549 |
| area | 16052 |
| performance | 15504 |
| complication | 15145 |

Table A.2.: The most occurring keywords by year (2011 - 2021)

| Label | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 |
|--------------|------|------|------|------|------|------|------|------|------|------|------|
| survey | 1306 | 1494 | 1670 | 1747 | 2292 | 2373 | 2715 | 2836 | 3444 | 4113 | 4512 |
| health | 880 | 1095 | 1180 | 1259 | 1577 | 1856 | 1988 | 2232 | 2658 | 3177 | 3679 |
| aor | 278 | 363 | 503 | 640 | 850 | 953 | 1256 | 1459 | 2038 | 2878 | 3607 |
| surgery | 727 | 882 | 1033 | 1149 | 1457 | 1631 | 1855 | 2008 | 2408 | 2923 | 3401 |
| performance | 578 | 665 | 802 | 846 | 1042 | 1181 | 1345 | 1520 | 1958 | 2460 | 3107 |
| self | 849 | 971 | 1174 | 1205 | 1448 | 1538 | 1674 | 1855 | 2163 | 2545 | 2910 |
| education | 780 | 952 | 1003 | 1049 | 1361 | 1494 | 1629 | 1734 | 2160 | 2550 | 2837 |
| complication | 579 | 658 | 767 | 892 | 1135 | 1287 | 1435 | 1550 | 1840 | 2346 | 2656 |
| auc | 195 | 221 | 298 | 350 | 489 | 594 | 730 | 865 | 1342 | 1864 | 2636 |
| prediction | 525 | 621 | 717 | 729 | 877 | 948 | 1107 | 1220 | 1608 | 2062 | 2569 |

statistical terms. From this point it becomes evident that we have no other choice but filtering out all the lexicon associated with biology and medical fields. Even though such filter risks to deprive us of the many applications of the discrete choice models in *Economics of Health*, it will help us to explore all other application fields.

In the Table A.2 we present some trends in the usage of the keywords in the literature. At this stage for more consistency we exclude from our analysis the keywords that appear only for certain years. We consider such drastic changes to be the error in the filtering algorithm and prefer to explore them separately if required. The results in the table are ordered based on the observations for the year 2021. Our finding confirm our hypothesis on the need to exclude the biology related terms: *Health*, *AOR* (which is related to genomics) and *Surgery*.

In addition to this basic analysis we explore the dataset by year separately, focusing our attention on the biology and medicine related clusters identified by the *VOSviewer*. Through this supplementary analysis we discover that such keywords as *Patient* and *Pregnancy* may also be considered as good filters for dimension reduction step in our analysis.

A.3. Advanced analysis

Following the conclusion of our preliminary analysis we decide to apply supplementary filters to our bibliography collection. In order to do so we decide to use the following additional restrictions to our query on WoS⁴:

NOT (Health OR Surgery OR AOR OR Patient OR Pregnancy)

Such refinement result in our dataset significant reduction far beyond the limiting point of the 100000 entries. This allows us to conduct the direct analysis of the complete dataset, without separating it by years as previously. Such analysis introduces new variable, as for example, the average citation year, which may give us indication on the ongoing trend in the different application fields.

The final dataset consists of 65654 items, which are analysed simultaneously. It is important to mention that at this point, due to significant reduction of our dataset base size, we decide to include supplementary observations. Thus, we decide to include the beginning of the year 2022, as well as the publications anterior to year 2011. This shift our lower bound limit to the year 1975, which becomes our new minimum. The inclusion of the start of the first months of the year 2022 induces some potential biases into the replicability of our research, because it makes it more difficult to obtain the same results as new and new publications appear. However, we make an assumption that such new publications should not affect our conclusion in the significant manner.

A.3.1. General information

Once we have excluded most of the biology and biometrics related publications it becomes more interesting in the context of our study to look at the properties of the collected data. We have already delimited the span and respective number of items in our dataset, but it is equally interesting to explore the sources of the publications in our dataset.

As we can see in the Table A.3, the main publishers are: (1) *Elsevier* - regrouping the publications related to economics, management and transportation; (2) *Springer Nature*, which encompasses publications related to ecology an remaining of biology oriented articles; (3) *Wiley* and (4) *Taylor & Francis*. Those key publishers are followed by *Sage*, *Mdpi*, *Emerald Group* and *Oxford University Press*, each amounting for more than 1000 items in the dataset.

The disciplines equally vary significantly. As we can see in the Table A.3b, even though we have excluded a significant number of articles oriented towards biology and natural sciences, our final dataset regroups a lot of publications oriented towards: *Environmental Sciences* and *Studies, Ecology, Geoscience* and *Environmental Occupational Health*. Fortunately, our filter works well enough to push *Economics* to the first place, alongside with tightly related disciplines such as: (1) *Transportation*, (2) *Management*, (3) *Business* and (4) *Sociology*. Because our main research pattern focuses on the statistical tools, we expectedly encounter among the dominating publication domains the *Statistics and Probability*, followed by *Operations Research* and *Computer Science*. Please note, that we show in the corresponding tables only the first and most prominent entries of the corresponding lists.

⁴Query link

Table A.3.: Sources composition

| (a) By publisher | | (b) By discipline | |
|-------------------|-------|--|------|
| Publisher | N | Discipline | N |
| Elsevier | 14576 | Economics | 5276 |
| Springer Nature | 7553 | Environmental Sciences | 4094 |
| Wiley | 7014 | Transportation | 3308 |
| [h] T&F | 5170 | Statistics Probability | 3207 |
| Sage | 2911 | Ecology | 2881 |
| Mdpi | 2417 | Environmental Studies | 2688 |
| Emerald Group | 1493 | Public Environmental Occupational Health | 2553 |
| Oxford University | 1277 | Geoscience Multidisciplinary | 2260 |
| W&W | 882 | Transportation Science Technology | 2259 |
| IEEE | 837 | Management | 2213 |

A.3.2. Keywords

What interests us the most at this stage of the research are the keywords appearing in the literature. Through the thorough keyword analysis we expect to identify the dominating fields and research patterns to further constrain our dataset by discipline. Such strategy should afford us to emerge a set of most prominent and representative works for each discipline / domain for further analysis. In our textual corpus the algorithm detects 836807 keywords, based on both title and abstract fields. Once again we use only binary word count in order to obtain the most general picture. We impose a limit of occurrences at 42 appearances, which narrows the relevant word count to 5631. This subset is further limited to the 60% of the most relevant resulting in 3379 words.

The *VOSviewer* map is represented in the Figure A.2. At this stage we can clearly distinguish three dominating clusters among and two lesser ones. In red we may distinguish the cluster regrouping advanced *ML* terms (such as *SVM*, *ML model*, *Accuracy*, *Recall*) alongside with *Geoscience* and *Bioscience* specific terminology (ex: *Landslide*, *Species*, *Natural Hazard*). In green we can see the drastically reduced subset of the keywords that are related to *Biology* and *Veterinary* fields: *Disease*, *Protein*, *BMI*, etc. The most interesting for us in the context of this study is the blue cluster, focusing on the *Economics* and *Social Sciences*. We can clearly distinguish a number of economic (*Market*, *Firm*), transportation (*Lane*, *Traffic*), finance and marketing terms, the last two categories being closely related to general economics terminology. Finally, we may equally consider as our center of interest the lesser yellow cluster, which is positioned midway between economics and biology fields. This is explained by the nature of the cluster - it regroups mainly *Sociology* and *Psychology* related terms (ex: *Abuse*, *Suicide Attempt*, *Parent*). Please note, the because of the lack of controls in terms of color choice, which is automated by *VOSviewer* software, we cannot guaranty the consistency in cluster colouring in this work.

The most occurring keywords are presented in the Table A.4. We present the *Average citations* and the number of *Occurrences* in our dataset for each of the words. At this point we underline that the keywords related to remains of *Biology* and *Health* related disciplines are still on the first place, regardless of imposed filters. This indicates, that a further analysis will be required with additional filters and refinement.

The density map allows us to further explore the cluster separation in our dataset (Figure A.3). On this

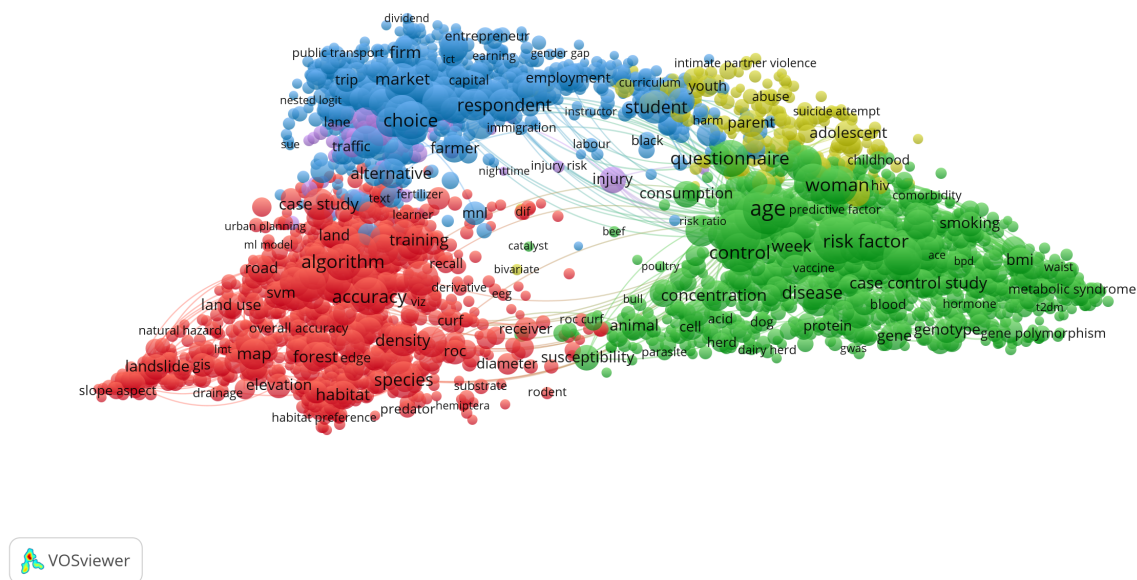


Figure A.2.: Keyword network map (1975 - 2022)}

Table A.4.: The most occurring keywords (1975 - 2022)

| Label | Cluster | Occurrences | Avg. Citations |
|------------------------------|---------|-------------|----------------|
| age | 2 | 10261 | 25 |
| risk | 2 | 8605 | 25 |
| logistic regression analysis | 2 | 7982 | 24 |
| association | 2 | 7940 | 23 |
| risk factor | 2 | 4497 | 27 |
| control | 2 | 4421 | 26 |
| odds ratio | 2 | 4323 | 30 |
| woman | 2 | 4267 | 25 |
| choice | 3 | 4122 | 27 |
| subject | 2 | 4010 | 30 |

figure there are several points of interest for us. First of all, we can see that the *Biology* related cluster has two separate gravitation centers: (1) one focused on the *Risk Factor*, while (2) being focused on the *Age* and *Control*. This can be explained by through the multitude of facets that exist in the clinical and medical studies. We can assume that what we observe here is caused by the presence in our dataset of both: theoretical biology studies, and economics and sociology related studies (ex: *Age* is a typical variable in most of the econometrics studies, which explains the behaviour of individuals).

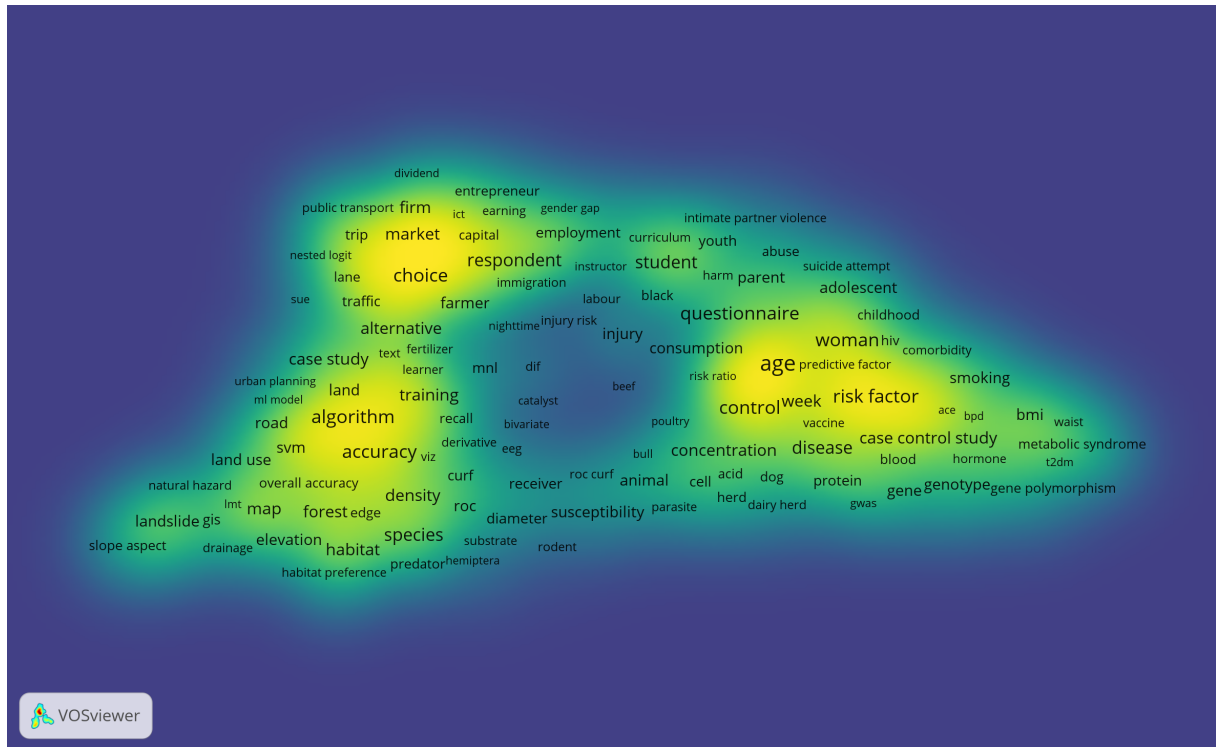


Figure A.3.: Keyword density map (1975 - 2022)

No we can focus on the different cluster more in detail. We have already outlined them and confirmed our hypothesis through the density map exploration. Now comes the time to dwell deeper into the cluster composition. Let's take a look at the most prominent keywords for each of the clusters:

1. Statistics and ML: accuracy, algorithm, species, dataset, distance, classification
2. Medicine and Health: age, risk, logistic regression analysis, association, risk factor
3. Economics and Choice Modelling: choice, logit model, policy, student, preference
4. Sociology and Psychology: disorder, parent, adolescent, drug, alcohol
5. Transportation research: injury, vehicle, crash, traffic, intersection

We are mostly interested by the cluster referring to economics. It comprises the economics related terms and potential economics applications. In the Table A.5 we offer a short list of the most occurring keywords from this cluster. In fact, the keywords included in this cluster may be used for a future refinement of our dataset. Using the detected keywords we will be able to impose additional constraints to our search query and thus refine our search by *Application Field*.

Thus, among the potential topics of interest we encounter:

- Individual choice modelling in general
- Policy making
- Preference studies

Table A.5.: Keywords related to Economics (1975 - 2022)

| Label | Occurrences | Avg. Citations |
|-------------|-------------|----------------|
| choice | 4122 | 27 |
| logit model | 3374 | 19 |
| policy | 3335 | 18 |
| student | 2722 | 15 |
| preference | 2707 | 20 |
| respondent | 2503 | 19 |
| attribute | 2354 | 24 |
| market | 2231 | 17 |
| household | 2217 | 17 |
| firm | 1971 | 23 |
| attitude | 1869 | 18 |
| income | 1869 | 14 |
| demand | 1801 | 21 |
| alternative | 1733 | 33 |
| company | 1689 | 14 |

Table A.6.: Keywords related to Sociology and Psychology (1975 - 2022)

| Label | Occurrences | Avg. Citations |
|------------|-------------|----------------|
| disorder | 1241 | 26 |
| parent | 1208 | 22 |
| adolescent | 1153 | 22 |
| drug | 846 | 21 |
| alcohol | 662 | 26 |
| youth | 648 | 22 |
| depression | 636 | 28 |
| girl | 622 | 23 |
| boy | 605 | 24 |
| partner | 573 | 20 |

- Market analysis
- Attitudes assessment
- Demand modelling (Aggregated demand modelling)
- Modelling of economic agent's behaviour: individuals (students, respondents), households, firms, companies

To the traditional purely economics problematic, we may add *Sociology*, *Psychology* and *Transportation* topics. Those are domains which contrary to biology and geoscience studies have a closely related methodology to our main topic: discrete choice analysis of behaviour.

The cluster combining the first two topics (Table A.6) focuses primarily on the *causal effect detection*. The articles included into our sample address topics of various *Disorders* and *Depression*, as well as *Disability*, *Violence* and *Peer* effects among the individuals.

The last one, *Transportation* related cluster (Table A.7), focuses on *Crash* detection and related policies

Table A.7.: Keywords related to Transportation (1975 - 2022)

| Label | Occurrences | Avg. Citations |
|-----------------|-------------|----------------|
| injury | 1282 | 29 |
| vehicle | 1280 | 24 |
| crash | 908 | 22 |
| traffic | 529 | 22 |
| intersection | 373 | 18 |
| fatality | 352 | 27 |
| injury severity | 335 | 35 |
| collision | 333 | 25 |
| lane | 330 | 22 |
| pedestrian | 308 | 23 |

and *Traffic* analysis. We can assume that those research is mostly oriented on public policy proposals and adoption, which makes this cluster potentially interesting for us.

A.3.3. Co-occurrences

Another possibility to explore the keywords for topic relevance detection can be achieved through co-occurrences network analysis. This procedure allows to go a bit deeper than simple textual mapping performed previously. Unfortunately, in contrast to the text mapping, only the keywords explicitly defined in the paper are analysed. In this case, to perform the analyse of co-occurrences we decide to use *all keywords* available, in order to obtain the most general picture possible. The total number of keywords in this case amounts to 163265 words. As you can see the number is lower than in previous analysis step, because the abstracts are not taken into account. As always, we define the minimal occurrences limit at 0.01% (at least 16 occurrences), which brings up 4917 keywords. For simplicity this number is once again limited at 1000 most relevant items. The resulting co-occurrence map is presented in Figure A.4.

Identically to previously performed analysis, we detect several cluster: (1) in green - the *Biology* and *Epidemics* related terms; (2) in red - the terms related to *Economics* and *Marketing*; (3) in blue - the *Sociology* and *Psychology* disciplines; (4) in violet - highly technical cluster regrouping advanced modelling techniques, standing for *Statistics* and *ML* fields and applications (as well as some part of *Geoscience* discipline); and finally (5) in yellow - keywords related to *Geoscience* and *Ecology*. We do not include the light-blue cluster into our analysis, because it regroup the statistical terms related to discrete choice modelling in general, as well as the statistical practices in usage of Logit types models.

Once we have the understanding of the general principles in the resulting clusters and the mapping criteria, we can proceed with further analysis. The Figure A.5 adds an overlay containing information on average citation year. This is a rather complex metrics, but it can offer us some basic understanding of current trends in the literature. In yellow we can see the keywords related to the emerging, or highly discussed nowadays topics. While in blue are colored the more persistent thematics.

For some more precision concerning the *Economics* discipline we offer a closer look at the related cluster in Figure A.6. The cluster has several major gravity centers within itself. On the frontier with the *Statistics* and *ML* related cluster we have the *Model* term, which is rather expected behaviour. Another

gravity point is defined simultaneously by three keywords: *behaviour*, *performance* and *impact*. We can assume that those are the typical targets of the different empirical and theoretical studies. Finally, two more points remain: (1) *Management*, regrouping the managerial science; and (2) *Determinants*, which is another potential objective of many research papers.

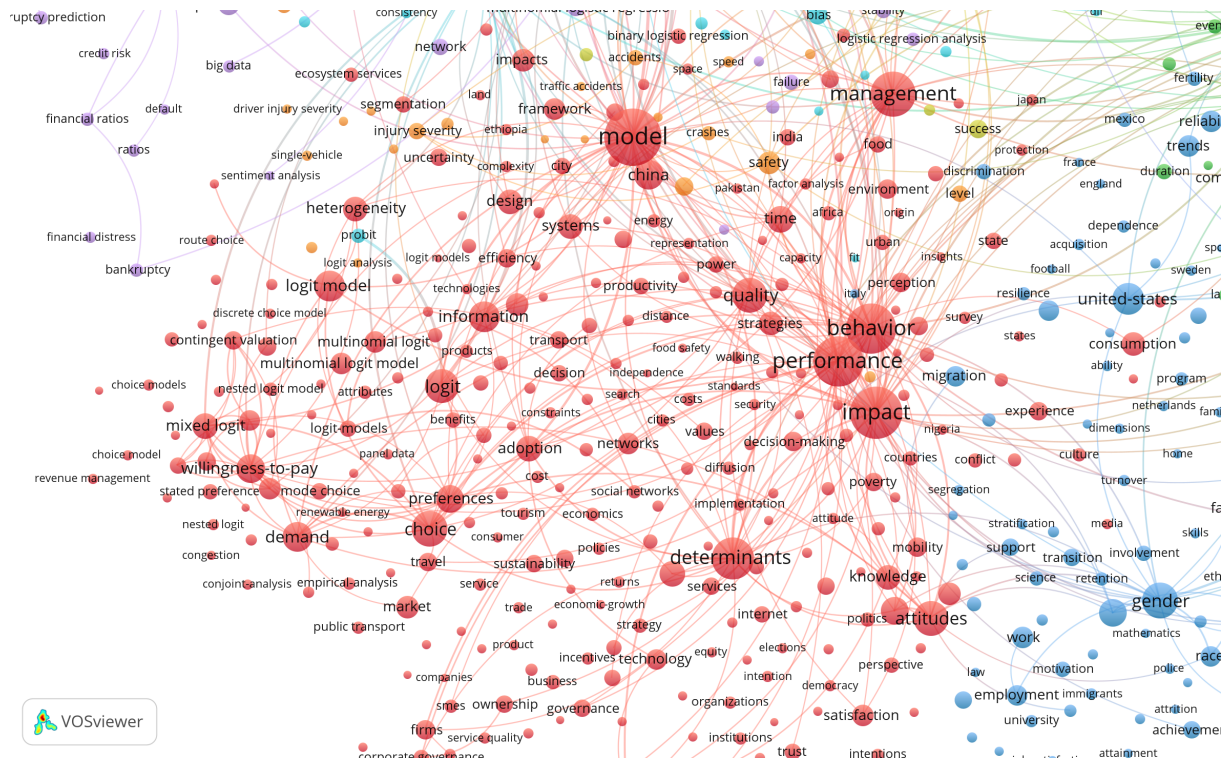


Figure A.6.: Co-occurrences map, all keywords - focus on Economics (1975 - 2021)

A better understanding may be achieved through identification of other prominent keywords in this cluster. The results of such filtering are presented in the Table A.8. This allows us to capture some additional terms, which are now more closely related to field studies. Those are the keywords, that should guide us in the further narrowing of our scope and restricting our dataset to specific topics.

As we can see the keywords presented in the Table A.8 can be regrouped to outline more general topics. For example:

- *Model* and *Performance*, which underline the topic of modelling in general, as well as the particular target in prediction tasks
- *Impact* and *Determinants*, which represent another facet of modelling objectives, focusing on the explication and causal effects understanding
- *Behaviour*, *Attitudes* and *Willingness to Pay (WTP)* - those regroup the topic of understanding of the individual behaviour, which is rather common in choice modelling
- *Management*, which stands aside from other *Economics* related disciplines, although the explored questions and used techniques are closely related
- *Demand*, which unites the market analysis in general

A.3.4. Citations

Finally, before proceeding with subsetting our dataset by topic based on the identified keywords, we would like to analyse the most cited documents in our bibliography. Because of the particular focus of

Table A.8.: The most occurring keywords in Economics related cluster (1975 - 2021)

| Label | Occurrences | Avg. Citations |
|--------------------|-------------|----------------|
| model | 3051 | 24 |
| impact | 2564 | 18 |
| performance | 2349 | 23 |
| behavior | 2340 | 22 |
| management | 1900 | 21 |
| determinants | 1658 | 17 |
| choice | 1200 | 23 |
| attitudes | 1145 | 19 |
| quality | 1144 | 19 |
| logit | 1114 | 29 |
| logit model | 926 | 22 |
| information | 922 | 26 |
| demand | 822 | 24 |
| china | 821 | 23 |
| willingness-to-pay | 768 | 21 |

our bibliometrics study on the final stages of this research we seek to analyse the most prominent and representative studies by domain. This means, that once we arrive at a sufficiently granular level of the dataset (though division by domain) we should be able to identify the most cited articles and analyse them. In order to offer a consistent analysis we should first of all focus on the most cited works, which *are common across all the topics*. This section serves us to perform this exact task.

For general analysis we explore the citations count on the single document level. Thus we will be able to exclude those most cited works from our future analysis. The total number of documents in our collection is at 65654 articles. We define the minimal citation number limit at 0.1% level of all documents (rounded to 66 citations), which drastically reduces our document selection to 5741 works. The most relevant 1000 works, based on weighted link strengths, are selected and the main cluster containing 971 document is explored.

The Figure A.7 offers an overview of the citation map using a density representation. This map allows us to detect the most prominent clusters and dependencies among the cited works. In the center we encounter the biggest cluster of *Biometrics* (*Biology* related modelling) articles, which focus on different ecological and environmental questions. For example, while Friedman (2001) focuses on technical aspects and proposes a gradient boosting method for model estimation, Allouche, Tsoar, and Kadmon (2006) focuses on more applied question related to accuracy of specie distribution models. Dormann et al. (2013) explores the ways to combat the collinearity, and Firth (1993) proposes a methodology for bias reduction in maximum likelihood estimates.

The “branches” descending from the central cluster are more discipline specific. On the left side we encounter the cluster related to *Geoscience*: Ayalew and Yamagishi (2005) describes a GIS-based logistic regression for landslide detection. In the upper part of the figure we encounter more advanced *ML* techniques in application to the engineering and technical disciplines: Chen et al. (2014) uses Deep-Learning techniques for classification of hyperspectral data. Finally, the cluster representing the most interest for us is rightmost branch: McFadden and Train (2000) introducing the Mixed MNL models

for discrete response data analysis, which is one of the key works in *Choice Modelling*.

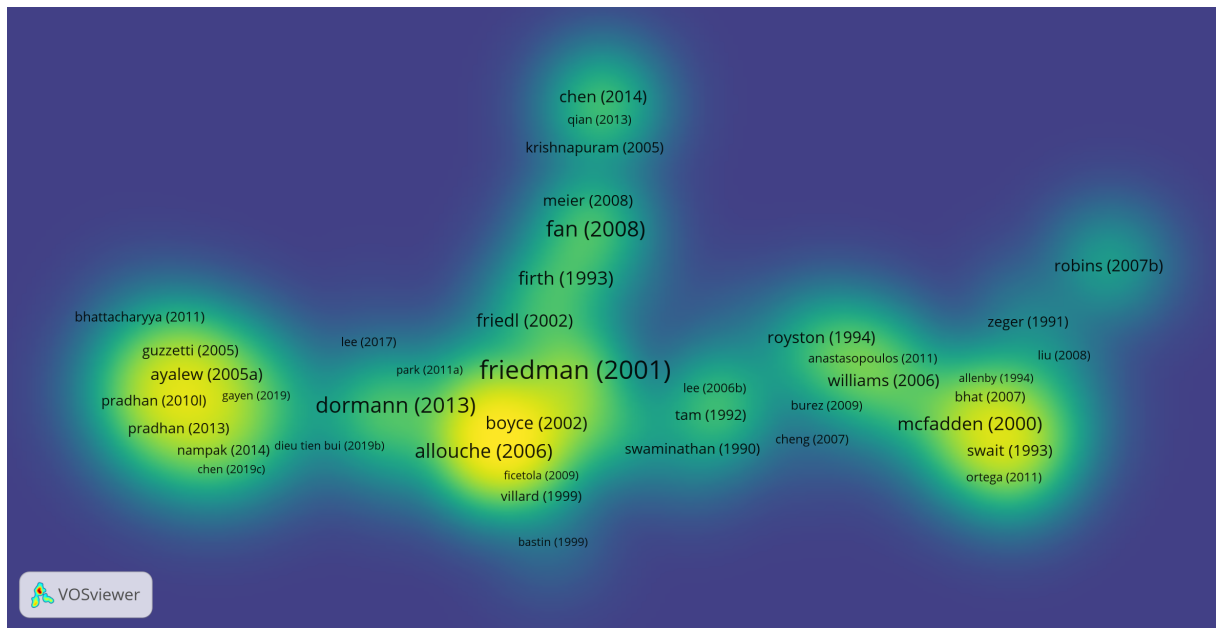


Figure A.7.: Citation map on document level (1975 - 2021)

Let us focus on economics related cluster as depicted on Figure A.8. As we can see the works of McFadden and Train (2000), Albert and Chib (1993), Hausman, Hall, and Griliches (1984) and Greene and Hensher (2003) for the gravity center on the figure. Those are the most cited works in this cluster.

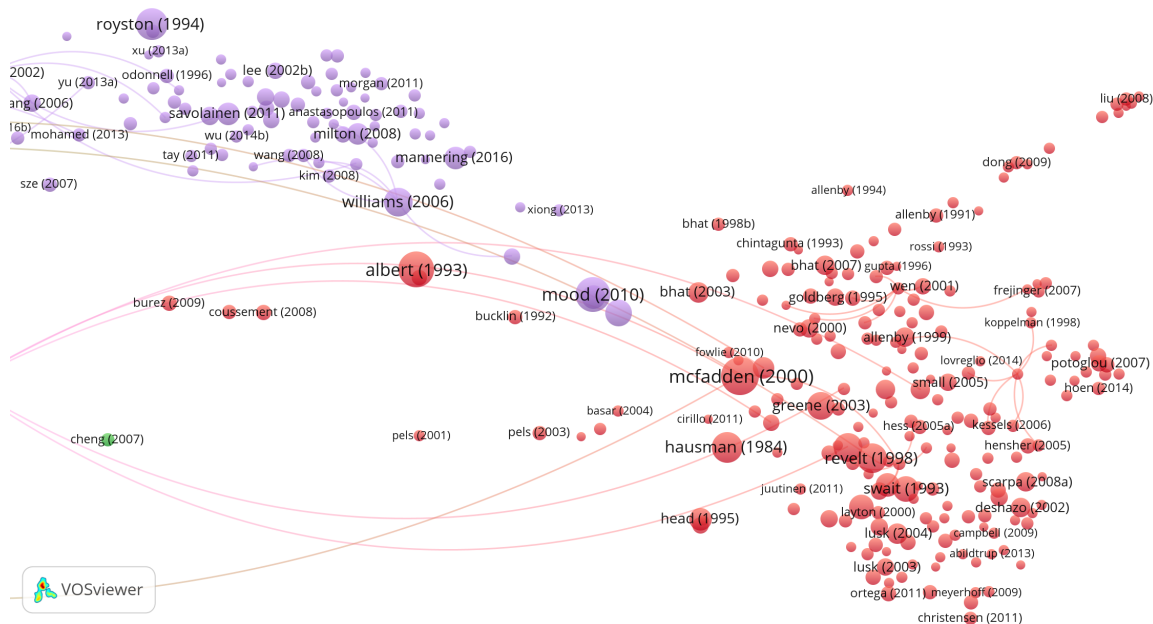


Figure A.8.: Citation map on document level - focus on Economics (1975 - 2021)

To better understand their nature we should probably explore the topics addressed by those works. In the Table A.9 we offer a precise list of top 10 most cited works of this cluster. Most of the works on the list are theoretical or methodological, which perfectly explains their high citation score. The works are mostly consecrated to the discussion of the advanced modelling techniques, that become more and more popular nowadays. Thus, becoming more and more cited in the recent applied studies.

Among the technical topics, one of the key concepts is the *Mixed Logit* (or *Mixed MNL* in some more

Table A.9.: Top 10 most cited works in Economics cluster (1975 - 2021)

| Label | Citations | URL |
|-----------------|-----------|---|
| Mcfadden (2000) | 1994 | "https://doi.org/10.1002/1099-1255(200009/10)15:5<447::aid-jae570>3.0.co;2-1" |
| Albert (1993) | 1737 | https://doi.org/10.2307/2290350 |
| Hausman (1984) | 1206 | https://doi.org/10.2307/1910997 |
| Revelt (1998) | 1102 | https://doi.org/10.1162/003465398557735 |
| Hensher (2003) | 1040 | https://doi.org/10.1023/a:1022558715350 |
| Greene (2003) | 901 | https://doi.org/10.1016/s0191-2615(02)00046-2 |
| Swait (1993) | 740 | https://doi.org/10.2307/3172883 |
| Boxall (2002) | 690 | https://doi.org/10.1023/a:1021351721619 |
| Train (1998) | 607 | https://doi.org/10.2307/3147053 |
| Head (1995) | 532 | https://doi.org/10.1016/0022-1996(94)01351-r |

precise cases). This is an advanced modelling technique allowing the introduction of heterogeneity into the Logit (or MNL) coefficient estimates and thus bypassing some of the technical limitations of the baseline model. McFadden and Train (2000) offers a general overview of the Mixed MNL modelling of the discrete response data; David A. Hensher and Greene (2003) describes the Mixed Logit from the state of practice perspective. Some of the other studies embed similar discussion into more applied work: Revelt and Train (1998) analysing household appliance choice, or Brownstone, Bunch, and Train (2000) illustrating the usage of Mixed Logit models with mixed *Stated Preferences* (SP) and *Revealed Preferences* (RP) data. Finally, some of the works represent *state of art* for some of the disciplines, as for example the article of Head, Ries, and Swenson (1995) analysing the industrial location choice. Or Allenby and Rossi (1998) describing marketing models of consumer heterogeneity.

We also encounter a number of other rather advanced theoretical topics. For example, Albert and Chib (1991) describes a framework for Bayesian Analysis of Binary and Polychotomous data. Hausman, Hall, and Griliches (1984) offers a discussion on MNL model specification testing and validation. Boxall and Adamowicz (2002) and later Greene and Hensher (2003) describe a Latent Class model for discrete choice analysis, which incorporates some of the ideas of semi-parametric estimation techniques. Lusk and Schroeder (2004) and Brownstone, Bunch, and Train (2000) explore the different implications and usage of data from different sources. Bhat (2001) and Bhat (2003) offers a discussion on the *Maximum Likelihood* estimation numerical implementation, with help of quasi-random or Halton sequences.

A.4. Analysis by subdomain

All this information makes it easier for us to proceed. In the previous section we have identified a set of topics and potentially associated with these topics keywords. Now it is time to use those keywords to restrict our dataset: we seek to separate our main dataset into several smaller ones with help of those identified terms. This way we should be able to explore each of the available dimensions separately. We can combine identified keywords, which theoretically should identify the separate application fields, in order to proceed.

At this point we prefer to retain the following keywords:

- *Policy*

- *Demand*
- *Household*
- *Firm or Company*
- *Market* (we prefer to separate this term from *Marketing*)
- *Preference or Attitude*

In addition we have observed separate clusters for:

- *Sociology and Psychology*
- *Transport* (standing out for *Transportation*)
- *Marketing*

We refine the search adding the identified field specific keywords. For example, to search for keyword *policy* we transform our query:

AND (Policy)

From this point we will proceed with separate analysis of each of the available fields for us. In this document we offer an overview only of a fraction of our findings.

A.4.1. Policy

We can assume that *policy* related group focuses on evaluation of public policies, their introduction and analysis. As presented on the Table A.10 we can see from the most occurring keywords the greatest part of such studies rely on the WTP (*Willingness*) and *Preference* examination. Some of the studies involve *Attribute* analysis, which is rather typical for DCM. Other keywords are mostly related to the particular topics (and consequently agents) addressed: *Farmer, Firm, Association*, etc. Or some key concepts through which the policy is analysed (ex: *Price*). Finally, we encounter some of the technical terms as well: *Originality Value, Design Methodology* and *Predictor*.

In the Table A.11 we offer an overview of the most cited works. During this table construction we filter out the references already presented previously in general overview. This allows us to focus primarily on the literature particularly specific to the explored field. As expected the most cited works belong mostly to the period before 2010: the older is the article, the more chances are for it to be cited. Nevertheless, we may assume that in nowadays literature the cited articles should be closely related to the explored problematic. Meaning, that the topics discussed in the literature list presented in Table A.11 are primarily the main topics of interest in the explored domain.

Even after filtering out all the previously discussed theoretical works we still encounter the article of DeShazo and Fermo (2002), which offers an overview of complexity and consistency arbitrage in the choice experiments. This work demonstrates that the excessive complexity of the choice setting risks to introduce some additional biases into the estimates. The rest of the works on the list are mostly applied ones. Even though all of them focus on public policy exploration, either to elaborate a new policy or to assess the impact of an already introduced one. The domains of policy application differ in a significant manner: we encounter the topics varying from transportation (Savolainen et al. 2011) to the market analysis (Goldberg 1995) or location choice (Bhat and Guo 2007). Let's explore those topics one by one.

Table A.10.: Policy adoption or assessment related keywords (1975 - 2022)

| Label | Occurrences | Avg. Citations |
|-----------------------------|-------------|----------------|
| preference | 803 | 19 |
| willingness | 599 | 20 |
| attribute | 581 | 23 |
| farmer | 516 | 15 |
| firm | 494 | 23 |
| association | 445 | 19 |
| student | 432 | 15 |
| woman | 429 | 16 |
| price | 415 | 20 |
| originality value | 398 | 9 |
| design methodology approach | 394 | 9 |
| child | 382 | 17 |
| user | 362 | 17 |
| predictor | 355 | 24 |
| consumer | 343 | 23 |

Table A.11.: The most cited documents related to Policy (1975 - 2022)

| Label | Citations | Cluster | URL |
|---------------------|-----------|---------|---|
| Savolainen (2011) | 542 | 3 | https://doi.org/10.1016/j.aap.2011.03.025 |
| Deshazo (2002) | 387 | 6 | https://doi.org/10.1006/j.eem.2001.1199 |
| Bhat (2007) | 380 | 10 | https://doi.org/10.1016/j.trb.2005.12.005 |
| Krueger (2016) | 343 | 21 | https://doi.org/10.1016/j.trc.2016.06.015 |
| Gellrich (2007) | 342 | 20 | https://doi.org/10.1016/j.agee.2006.05.001 |
| Dell'olio (2011) | 321 | 12 | https://doi.org/10.1016/j.tranpol.2010.08.005 |
| Small (2005) | 311 | 16 | https://doi.org/10.1111/j.1468-0262.2005.00619.x |
| Goldberg (1995) | 310 | 10 | https://doi.org/10.2307/2171803 |
| Scarpa (2010) | 294 | 7 | https://doi.org/10.1016/j.eneco.2009.06.004 |
| Birol (2006a) | 284 | 14 | https://doi.org/10.1016/j.ecolecon.2006.06.002 |
| Parkes (2002) | 282 | 16 | https://doi.org/10.1080/0042098022000027031 |
| Tyrinopoulos (2008) | 270 | 12 | https://doi.org/10.1016/j.tranpol.2008.06.002 |
| Prishchepov (2013) | 247 | 20 | https://doi.org/10.1016/j.landusepol.2012.06.011 |
| Hackbarth (2013) | 246 | 19 | https://doi.org/10.1016/j.trd.2013.07.002 |
| Banfi (2008) | 241 | 7 | https://doi.org/10.1016/j.eneco.2006.06.001 |

Speaking about transportation we encounter several articles related to the topic. Savolainen et al. (2011) performs “The statistical analysis of highway crash-injury severities”, which is mostly a literature review summarising existing statistical methodology for analysis of motor-vehicle injury severities. Among the data specificity authors identify: *underreporting of crashes, ordinal nature of data, fixed parameters, omitted variable bias, small sample size, endogeneity, within-crash correlation and spatial and temporal correlation*. Among methodological models popular within the explored domain they detect: *binary outcome models* (Bayesian hierarchical binary Logit/simultaneous binary Logit, bivariate/multivariate binary Probit), *ordered outcome models* (copula-based multivariate approach, bivariate ordered probit, heterogeneous choice models, generalized logit models, Bayesian ordered probit/mixed generalized ordered logit) and *ordered multinomial discrete outcome models* (multinomial logit models, sequential logit and probit models, Markov switching multinomial logit, nested logit model, mixed logit models), as well as *other methods* (neural networks). Another example, Krueger, Rashidi, and Rose (2016) focuses on an applied study of preferences for shared autonomous vehicles with help of a stated choice survey. The modelling strategy includes a construction of Mixed Logit model and elicitation of the WTP, making the principal focus on the *Value of Time* (VOT) case. dell’Olio, Ibeas, and Cecin (2011) examines the quality of service desired by public transport users, putting accent on the valuation of various attributes by users. Small, Winston, and Yan (2005) exploits both SP and RP data to study commuters’ preferences for speedy and reliable highway travel using a Mixed Logit model. The conclusions are in this case once again based on the WTP derivation: VOT and *Value of Reliability* (VOR). Tyrinopoulos and Antoniou (2008) analyses public transit user satisfaction by means of factor analysis and ordered logit modeling. In the case of ordered logit the analysis is performed through direct coefficient comparison (the explanatory variables are binary in this case, which makes such comparison feasible). Finally, Hackbarth and Madlener (2013) explores consumers’ preferences for alternative fuel vehicles. The analysis is performed by means of WTP elicitation with multinomial logit model and a mixed (error components) logit model for attribute improvements.

As we can see, in most case the data analysis for policy recommendations relies on the WTP estimation. Many other studies in the group are closely tied to this metric as well. Scarpa and Willis (2010) analyses the WTP for renewable energy using both conditional and mixed logit models. Birol, Karousakis, and Koundouri (2006) uses choice experiment to explore the preferences in wetland attributes. Authors use conditional logit model, a random parameter logit model, a random parameter logit model with interactions and a latent class model to derive the WTP. Then the obtained values are exploited to provide a cost-benefit analysis, considering different management strategies for the wetland. Banfi et al. (2008) provides a discussion on the WTP for energy-saving measures in residential buildings. A fixed-effects logit model is implemented to estimate the effects over a dataset collected by means of discrete choice experiment.

Several studies include spatial dimensions. Bhat and Guo (2007) analyses the impacts of built environment characteristics on household residential choice (and auto ownership levels). The key focus in this study is made on the causality detection and identification. The process modelling takes form of a joint mixed multinomial logit-ordered response structure, which is used primarily to analyse the identified effect directly. Gellrich et al. (2007) performs a spatial analysis of agricultural land abandonment and natural forest re-growth using multivariate statistical models based on geo-physical and socio-economic variables. Ordinary logistic model and auto-Logistic model are used in this case for effects estimation, which are later analysed without additional transformations.

Table A.12.: Preference or Attitude related keywords (1975 - 2022)

| Label | Occurrences | Avg. Citations |
|------------------------------|-------------|----------------|
| logistic regression | 1281 | 19 |
| attribute | 962 | 22 |
| consumer | 678 | 23 |
| questionnaire | 577 | 17 |
| choice experiment | 541 | 25 |
| market | 498 | 17 |
| product | 490 | 20 |
| logistic regression analysis | 471 | 24 |
| price | 450 | 21 |
| heterogeneity | 447 | 24 |
| mixed logit model | 438 | 24 |
| alternative | 431 | 24 |
| distribution | 420 | 28 |
| participant | 414 | 14 |
| student | 401 | 18 |

Among sufficiently differing research publications we encounter the paper of Parkes, Kearns, and Atkinson (2002), where the individual neighbourhood dissatisfaction was analysed. The modelling techniques were limited to the basic logistic regression model in this case. And the analysis was principally based on the direct effects comparison. Another interesting study is the work of Goldberg (1995), where the demand in the automotive industry sector is modelled under product differentiation in an oligopolistic market. Here the discrete choice model is implemented on the demand side in conjunction with the aggregate demand derivation. The estimation results are then used in counterfactual simulations to investigate several trade policy issues.

A.4.2. Preferences or attitudes

The preferences and attitudes are studied in conjunction, because it's assumed that those keywords are closely enough related. Table A.12 summarise the most occurring keywords in our subsample. As one can see, we encounter some evident modelling related keywords such as: *logistic regression*, *logistic regression analysis*, *mixed logit model* and *distribution*. Those are accompanied by the specific terminology related to the discrete choice experiments and data collection procedure: *attribute*, *alternative*, *questionnaire*, *choice experiment* and *participant*. This is rather important observation, because it means that most studies focused on the preference or attitude exploration are typically based on the discrete choice experiments or similar techniques (ex: survey based on questionnaire). Finally, we encounter the particular topic specific keywords: *market*, *product* and *price*. Those are the keywords pointing out the domains and applications, which are the most focused (or dependent) on the preference exploration.

Among the most cited works, as shown in the Table A.13, we encounter several previously seen (in the previous section consecrated to the public policies). This observation can be explained by the fact that many papers address several topics. What is more our selection of the keyword on the preselection and domain delimitation steps is far from perfect. This means that several keywords may unite

Table A.13.: Most cited works concerning Preference or Attitudes (1975 - 2022)

| Label | Citations | Cluster | URL |
|-------------------------------|-----------|---------|---|
| Deshazo (2002) | 387 | 6 | https://doi.org/10.1006/jeem.2001.1199 |
| Bhat (2007) | 380 | 13 | https://doi.org/10.1016/j.trb.2005.12.005 |
| Haboucha (2017) | 348 | 8 | https://doi.org/10.1016/j.trc.2017.01.010 |
| Krueger (2016) | 343 | 8 | https://doi.org/10.1016/j.trc.2016.06.015 |
| Dell'olio (2011a) | 321 | 12 | https://doi.org/10.1016/j.tranpol.2010.08.005 |
| Lusk (2003) | 316 | 1 | https://doi.org/10.1111/1467-8276.00100 |
| Small (2005) | 311 | 3 | https://doi.org/10.1111/j.1468-0262.2005.00619.x |
| Caussade (2005) | 310 | 3 | https://doi.org/10.1016/j.trb.2004.07.006 |
| Janssen (2012) | 302 | 1 | https://doi.org/10.1016/j.foodqual.2011.12.004 |
| Azucena Vicente-Molina (2013) | 287 | 7 | https://doi.org/10.1016/j.jclepro.2013.05.015 |
| Birol (2006a) | 284 | 2 | https://doi.org/10.1016/j.ecolecon.2006.06.002 |
| Parkin (2008) | 251 | 13 | https://doi.org/10.1007/s11116-007-9137-5 |
| Hackbarth (2013) | 246 | 4 | https://doi.org/10.1016/j.trd.2013.07.002 |
| Choo (2004) | 237 | 13 | https://doi.org/10.1016/j.tra.2003.10.005 |
| Ortega (2011) | 233 | 1 | https://doi.org/10.1016/j.foodpol.2010.11.030 |

works belonging to different domains and topics at the same time. For example, in the Table A.13 we encounter previously seen works on transportation of Krueger, Rashidi, and Rose (2016), dell'Olio, Ibeas, and Cecin (2011), Small, Winston, and Yan (2005) and Hackbarth and Madlener (2013). As well as some theoretical works and works focused on the spatial analysis: Bhat and Guo (2007), DeShazo and Fermo (2002) or Birol, Karousakis, and Koundouri (2006). We can observe that most of these studies speak about preferences and focus their analysis on the WTP exploration. This is rather expected target indicator to be exploited in preference or attitude analysis, because it provides a clear monetary value to the different attributes in the choice context.

As usual we encounter a number of other works related to the transportation research. Haboucha, Ishaq, and Shiftan (2017) investigating user preferences regarding autonomous vehicles, where acceptance are quantified through random utility models including logit kernel model taking into account panel effects. The resulting policy implications are analysed based on the effect and elasticity estimates. Parkin, Wardman, and Page (2008) identifies the determinants of bicycle mode share using census data. At the core of the study lies the basic logistic regression model, which is then used to provide an aggregate level predictions of transport usage for further analysis. The intermediary analysis in this case is based on the elasticities. Choo and Mokhtarian (2004) analyses the role of the attitudes and lifestyle on vehicle type choice. The multinomial logit model with IIA property is used to derive conclusions on impact of the considered variables on the vehicle type choice. <https://doi.org/> The estimated coefficients are directly analysed (in terms of sign and significance) and compared when possible to provide some insight into the policy implications.

Another interesting topic we encounter among the listed articles is the food and alimentary consumption (and preferences) modelling. For example, Lusk, Roosen, and Fox (2003) analyses the demand for beef from cattle administered growth hormones in several countries (France, Germany and United Kingdom). The authors use conditional logit model with variable interactions for effect identification. The final conclusions are based on the WTP values analysis. Janssen and Hamm (2012) addresses the issue of product labelling in the market of organic food. Thus the consumer preferences for different

organic certification logos are analysed based on the derived WTP values. The random parameter logit models are used for effect identification in this case. The estimated effects magnitude is equally taken in consideration when performing analysis in the study. Ortega et al. (2011) models heterogeneity in consumer preferences for several food safety attributes with several econometrics models, including latent class and random parameters logit. The final result and conclusion rely on the derived WTP values for corresponding attributes.

Finally, we encounter two more studies, which are more difficult to attribute to one of the groups above. First comes the work of Caussade et al. (2005), assessing the influence of design dimensions on stated choice experiment estimates. As one can see, this is a theoretical work oriented to ameliorate the existing practices in construction and design of discrete choice experiments. The authors consider the effects of *number of available alternatives, the number of attributes, the number of levels for those attributes, the range of attribute levels and the number of choice situations presented to each respondent*. To identify the effects of complexity authors use heteroskedastic logit model with the scale parameter specified as a function of the design dimensions. The results imply that all the dimension affect the choice variance (consistency), although no systematic effect on WTP estimates is found. Next we have the work of Vicente-Molina, Fernández-Sáinz, and Izagirre-Olaizola (2013), where the factors impacting the pro-environmental behaviour are explored. To estimate the effects a multinomial ordered logit model is applied, the focus is made on the influence of the covariates on the environmental performance probability. Both effect estimates and the derived elasticities are used in this case to conclude.

A.5. Conclusion

In conclusion, this thorough exploration of research procedures and literature review methods provides a solid foundation for understanding the complexities of discrete choice analysis, particularly in the context of economics. The proposed bibliometric study addresses the ambiguity in terminology across different fields and disciplines, offering a systematic approach to filter and analyze the extensive literature. Recognizing the challenges of balancing general and specific keywords, the study strategically employs multiple stages to refine the dataset. Despite potential limitations due to the exclusive use of the WoS database, the study aims to capture the most important works within the scientific community.

The choice of keywords, focusing on the choice modeling, has its own limitations, as some of the less popular branches working on behavioural studies get excluded. We follow the best practices encountered in literature on bibliometric studies with the usage of VOSviewer for data analysis part. The subsequent examination of publishers, disciplines, and citation patterns contributes to a comprehensive understanding of the landscape of discrete choice modeling literature.

The exploration of the most cited documents adds another layer to the analysis, identifying common themes across various topics and guiding the further subsetting of the dataset by specific keywords. The density representation of citation clusters provides visual insights into prominent themes, ensuring a nuanced approach to subsequent analyses. The strategic refinement of searches based on identified keywords for each field enhances the granularity of the analysis.

This methodologic approach to literature review and bibliometric analysis lays the groundwork for a nuanced and comprehensive exploration of discrete choice analysis in economic applications. The

findings from these stages will undoubtedly contribute to a deeper understanding of the applications, trends, and influential works in the field, providing valuable insights for researchers and practitioners alike.

B. Extracting economic information from Neural Networks

This appendix extends the notions of *Machine Learning (ML)* techniques and models introduced in the first chapter. A broader overview of the various ML techniques related to the Neural Networks construction is provided. Those advanced and less conventional modeling techniques, often overlooked by economists due to perceived limitations in offering insights into effects estimation. ML models, viewed by economists as inefficient for the purposes of effect identification excel in predictive power while suffering from the lack of interpretability. With a focus on prediction, models generally follow three steps: observation, model construction, and prediction; skipping the interpretation stage. Despite sharing assumptions of independence and identical distribution with Econometrics and *Social and Human Sciences (SHS)*, ML scientists prioritize functional form flexibility with large samples.

This document returns to the concepts of artificial neuron and perceptron, introduced in the body of the thesis. However, this time a more detailed description of the associated estimation techniques and algorithms is provided. From Stochastic Gradient Descent, to Adaline and, finally, the Backpropagation algorithm.

Despite the evolution of NN architectures for more flexible functional forms, economists find limited utility in their predictive power. As the focus is made on effects exploration, causal relationships, and deriving meaningful indicators. However, recent work has explored extracting key metrics, shedding light on the potential relevance of NN models for economists, considering their similarity to Discrete Choice Models in adopting a *softmax* output layer. At the end of this section a detailed list of economic indicators extractable from the NN-based models is provided.

B.1. Statistical and Machine Learning perspective

This group of models focuses on more advanced and atypical modelling techniques rarely implemented by the economists in their studies, as usually this family is perceived as not offering enough insight when it comes to the effects estimation. The *Machine Learning* (ML) techniques are usually viewed by economists as some black boxes, which do not provide any information about the underlying process. It is quite easy to comply with their position, as even though the most advanced techniques perform better in terms of predictive power, they rarely offer any insight into the modelling process.

There exist two possible approaches to presentation of the *Neural Networks* (NN). In the literature focused on *statistical learning* and *data analysis* we may encounter the introduction of NN through more simple statistical models (Hastie, Tibshirani, and Friedman 2009). Another representation may be encountered in the community focused on *informatics* and ML, where the authors adopt algorithmic approach. The ML paradigm differs from the statistical analysis by its main focus: the prediction. In other words, we may say that ML focuses on the result, rather than on the process itself. The simplest structure for the majority of the ML algorithms may be summarised to three main steps: (1) observation, (2) model construction, and (3) prediction.

B.2. Introduction to Neural Networks

Most of the models implemented in ML community rely on the assumptions of *independence* of observations and their *identical distribution*. As one may remark those assumptions are identical to the ones encountered in *Econometrics* and *Social and Human Sciences* (SHS) field. Nevertheless, there is a major difference in the paradigms: while econometricians extend their models to tackle various biases in those two key assumptions, the ML scientists focus on the functional form flexibility in presence of large samples.

The learning problem may hence be formalised: Considering input space $\mathcal{X} \subset \mathbb{R}^d$ and an output space \mathcal{Y} . And given example pairs $(x, y) \in \mathcal{X} \times \mathcal{Y}$ are *identically* and *independently* distributed (iid.) with respect to an unknown but fixed probability distribution \mathcal{D} . Assuming that only N pairs of (x_i, y_i) generated from \mathcal{D} are observed. The *aim* is to construct a function $f : \mathcal{X} \rightarrow \mathcal{Y}$, which predicts an output y for a given x with a minimal error.

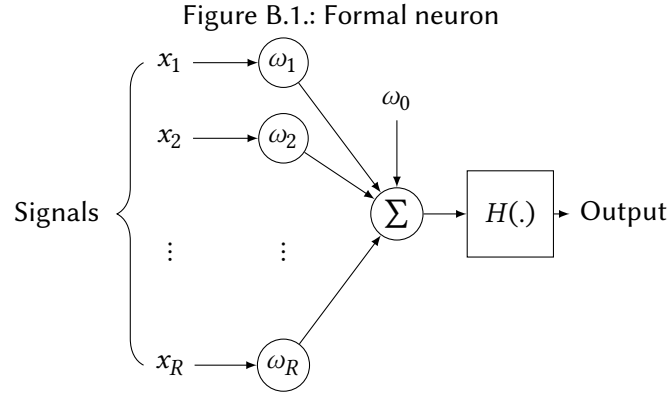
The notation used in this work complies with the previously introduced convention, as expressed in the first chapter. For convenience purposes we repeat the notation convention in Table B.1.

B.2.1. Artificial Neuron and Perceptron

Speaking about the particular implementation of the learning algorithms under the form of a NN, we can trace the history to Ramon y Cajal (2002). Nobel prize laureate in 1906 in biology and neuroscience, he remains known as the first one to represent the biological neurons' anatomy. Grace to this particular step in biology domain, the scientific community obtained a new dream - the possibility to artificially reconstruct the neural structure and hence the brain itself. It's in the work of McCulloch and Pitts (1943) that the first mathematical formalisation of a neuron appears (Figure B.1). Keep in mind, that for simplicity we omit the observation index i so $\mathbf{x} = (x_1, \dots, x_R)$ Afterwards, many various learning rules were proposed.

Table B.1.: Notation
[H]

| Notation | Definition |
|---|---|
| \mathcal{X} | Input space |
| \mathcal{Y} | Output space |
| $(\mathbf{x}_i, y_i), i \in \{1, \dots, N\}$ | Observation i |
| $\mathbf{x}_i = (x_{i1}, \dots, x_{iR})$ | Explicative variables vector of size R |
| y_i | Outcome variable |
| $\mathcal{S}_N = \{\mathbf{x}_i, y_i\}$ | Sample of N observations |
| \mathcal{D} | Probability distribution |
| $f : \mathcal{X} \rightarrow \mathcal{Y}$ | Function mapping \mathcal{X} to \mathcal{Y} |
| $\mathcal{F} = \{f : \mathcal{X} \rightarrow \mathcal{Y}\}$ | Class of functions |
| $\mathcal{L}(f) = \mathbb{E}[l(f(\mathbf{x}), y)]$ | Generalisation Loss (Error) |
| $\hat{\mathcal{L}}(f(\mathbf{x}), \mathcal{S}_N) = \hat{\mathcal{L}}(\omega)$ | Empirical Loss (Error) |
| ω | Parameters of prediction function |



A most simple formal neuron may defined with a prediction function h_ω , which is linear:

$$h_\omega : \mathbb{R}^d \rightarrow \mathbb{R} \quad (\text{B.1})$$

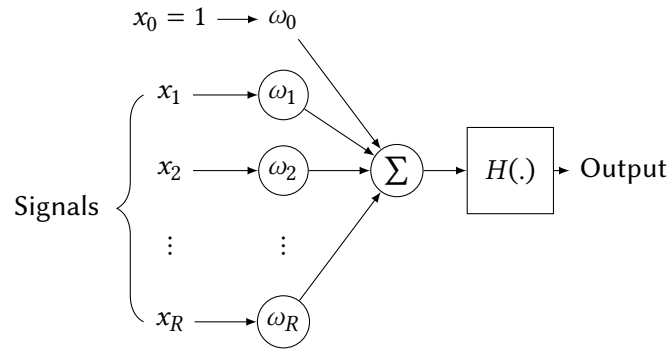
$$\mathbf{x} \mapsto \langle \hat{\omega}, \mathbf{x} \rangle + \omega_0 \quad (\text{B.2})$$

Assuming ω_0 to be included in the vector ω and $x_0 = 1$, we can rewrite the formal rule. The changes may summarised in graphical form as in Figure B.2.

$$h_\omega : \mathbb{R}^d \rightarrow \mathbb{R} \quad (\text{B.3})$$

$$\mathbf{x} \mapsto \langle \hat{\omega}, \mathbf{x} \rangle = \omega \mathbf{x} \quad (\text{B.4})$$

Figure B.2.: Formal neuron (alternative representation)



Later this model was readapted and tested by Rosenblatt (1958). The linear part of the perceptron was identical to the one proposed previously, but the learning rule was optimised. The model opted to find the best set of parameters $\omega = \{\omega_0, \dots, \omega_R\}$ through minimisation of the distance between misclassified examples to the decision boundary. We may define the objective loss function for the simple *Perceptron* model as:

$$\hat{\mathcal{L}}(\omega) = - \sum_{i \in N} y_i(\omega \mathbf{x}_i)$$

B.2.1.1. Gradient descent

The learning algorithm attempts to minimise the given objective function. The most traditional update rule is the *gradient descent* algorithm. The main idea of the given procedure is to modify the parameters accordingly to the observed gradient for a given set of parameters.

$$\forall t \geq 1, \omega^{(t)} \leftarrow \omega^{(t-1)} - \eta \nabla_{\omega^{(t-1)}} \hat{\mathcal{L}}(\omega^{(t-1)})$$

For further use in the algorithms we may immediately note the update part as Algorithm 1.

Algorithm 1 Gradient descent

Choose randomly an example $(\mathbf{x}^{(t)}, \mathbf{y}^{[t]}) \in \mathcal{S}$
if $y\omega^{(t)}x^{(t)} < 0$ **then**
 $\omega^{(t)} \leftarrow \omega^{(t-1)} + \eta \nabla_{\omega^{(t-1)}} \hat{\mathcal{L}}(\omega^{(t)})$
end if

The proof of convergence of such approach was provided by Novikoff (1962). Partial derivatives in this particular case are:¹

$$\nabla_{\omega} \hat{\mathcal{L}}(\omega) : \frac{\delta \hat{\mathcal{L}}(\omega)}{\delta \omega_r} = - \sum_{i \in I} y_i \mathbf{x}_i$$

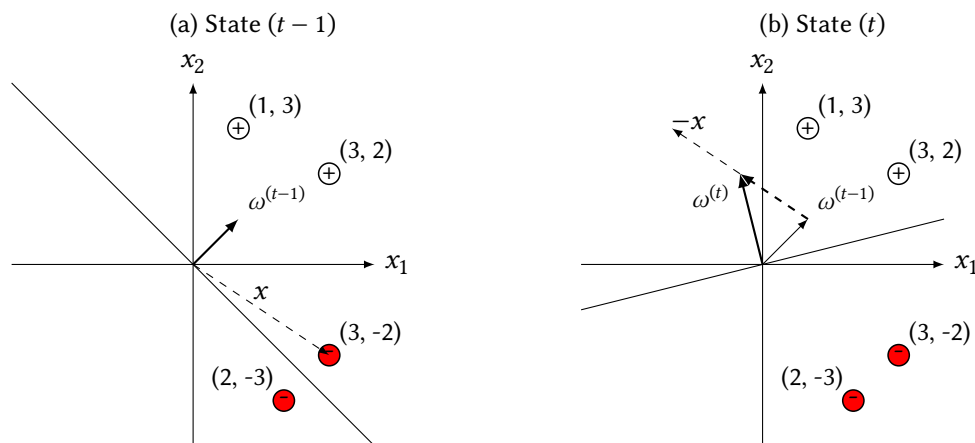
The resulting update rule may be represented as:

¹Particular case for ω_0 is given by $-\sum_{i \in N} y_i$

$$\forall (\mathbf{x}_i, y_i) \in \mathcal{S}_{\mathcal{N}}, \text{ if } y_i \omega \mathbf{x}_i \leq 0 \text{ then } \omega \leftarrow \omega + \eta y_i \mathbf{x}_i$$

Where η stands for the *learning rate* hyperparameter, externally defined by the scientist. The lower η will be, the slower will be the *learning* process, and the more precise results may be obtained.

Figure B.3.: Example: gradient descent update (assuming $\eta = 0.5$)



B.2.1.2. Perceptron algorithm

The simple perceptron algorithm may be summarised in the form of Algorithm 2.

Algorithm 2 Simple Perceptron

Require: $\mathcal{S}_{\mathcal{N}}, \eta > 0, T > 0$

$\omega^{(0)} \leftarrow \mathbf{0}$

$t \leftarrow 1$

while $t < T$ **do**

 Choose randomly an example $(\mathbf{x}^{(t)}, \mathbf{y}^{[t]}) \in \mathcal{S}$

if $y \omega^{(t-1)} \mathbf{x}^{(t)} < 0$ **then**

$\omega^{(t)} \leftarrow \omega^{(t-1)} + \eta y^{(t-1)} \mathbf{x}^{(t)}$

end if

$t \leftarrow t + 1$

end while

Ensure: $\omega^{(t)}$

▷ Initialize weight vector

▷ Initialize epoch count

We can imagine a simple code in R to execute this algorithm:

```
# Perceptron function
perceptron = function(S, eta = 0.01, epoch = 100) {
  w = rep(0, ncol(S) - 1)
  t = 1
  while (t < epoch) {
    S_t = as.matrix(S[sample(1:nrow(S), 1), ])
    x_t = S_t[, 1:(ncol(S) - 1)]
    y_t = S_t[, ncol(S)]
    if (y_t * (w %**% x_t) <= 0) {
      w = w + eta * y_t * x_t
    }
  }
}
```

```

    }
    t = t + 1
  }
  return(w)
}

```

B.2.2. Adaline

The next stage of the NN development was marked by the introduction of *Adaptive Linear Neuron (Adaline)* by Widrow and Hoff (1960). The main difference from the simplistic *perceptron* was the introduction of quadratic loss function. The optimisation in this case was performed through minimisation of the *Mean Square Error (MSE)*. As one may remark, this procedure is quite close by its nature to the one we observe in econometric implementation of well known *Ordinary Least Squares (OLS)* linear model. The prediction function remains linear:

$$h_{\omega} : \mathbb{R}^d \rightarrow \mathbb{R} \quad (\text{B.5})$$

$$x \mapsto \hat{\omega}, x \quad (\text{B.6})$$

The loss function in this case is given by:

$$\mathcal{L}(\omega) = \frac{1}{N} \sum_{i=1}^N (y_i - h_{\omega}(x_i))^2$$

Consequently, the derivatives change as well and the new weights' update rule is given by:

$$\forall (x_i, y_i) \in \mathcal{S}_{\mathcal{N}}, \text{ if } y_i(\omega x_i) \leq 0 \text{ then } \omega \leftarrow \omega + \eta(y_i - h_{\omega}(x_i))x_i$$

B.2.2.1. Adaline algorithm

Consequently, we may easily write the Algorithm 3 for this new procedure:

Algorithm 3 ADALINE

Require: $\mathcal{S}_{\mathcal{N}}, \eta > 0, T > 0$

$\omega^{(0)} \leftarrow \mathbf{0}$

$t \leftarrow 1$

while $t < T$ **do**

 Choose randomly an example $(\mathbf{x}^{(t)}, \mathbf{y}^{[t]}) \in \mathcal{S}$

 Calculate $h_{\omega^{(t-1)}}(x^{(t)}) \leftarrow \omega^{(t-1)} x^{(t)}$

$\omega^{(t)} \leftarrow \omega^{(t-1)} + \eta(y^{(t)} - h_{\omega^{(t-1)}}(x^{(t)}))x_i$

$t \leftarrow t + 1$

end while

Ensure: $\omega^{(t)}$

- ▷ Initialize weight vector
- ▷ Initialize epoch count

We can imagine a simple code in R to execute this algorithm:

```

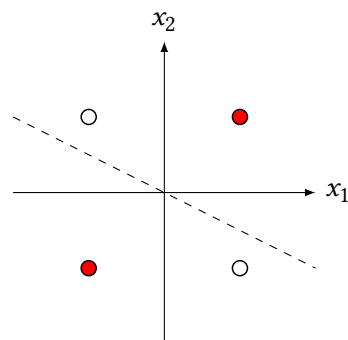
# Adaline function
adaline = function(S, eta = 0.01, epoch = 10) {
  w = rep(0, ncol(S) - 1)
  t = 1
  while (t < epoch) {
    S_t = as.matrix(S[sample(1:nrow(S), 1), ])
    x_t = S_t[, 1:(ncol(S) - 1)]
    y_t = S_t[, ncol(S)]
    h_w = c(w %*% x_t)
    if (y_t * h_w <= 0) {
      w = w + eta * (y_t - h_w) * x_t
    }
    t = t + 1
  }
  return(w)
}

```

B.2.3. Multilayer Perceptron

With the developments and improvements of simple models some of their drawbacks became apparent (Minsky and Papert 1969). Most of them propose a linear (or sigmoid in case of Logit) separations, whereas in real world linearly separable problems are few. More elaborate learning algorithms required more complex logical rules, as for example XOR (exclusive OR², see Figure B.4) or parity rules. The XOR problem is a classic example in the field of machine learning and artificial intelligence that highlights the limitations of linear models.

Figure B.4.: XOR problem



The circuit theory was poorly developed to solve such complex problems, and a single-layer perceptron can only learn linearly separable functions, to which XOR does not belong. This situation resulted in active search for non-linear models and the specific learning techniques.

²The exclusive OR (XOR) is a binary operation that takes two binary digits (0 or 1) as inputs and produces a single binary output. The XOR operation returns 1 if the inputs are different and 0 if they are the same.

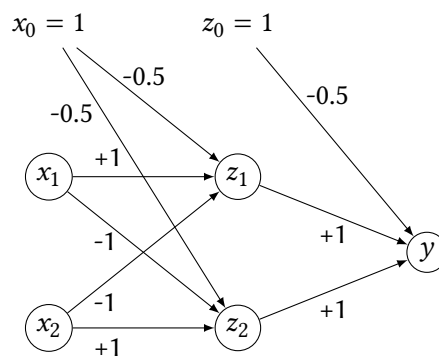
Figure B.5.: XOR problem solution



The XOR problem can be successfully solved by introducing a hidden layer in a neural network, where the hidden layer corresponds to a transition into a latent space. This additional layer allows the network to learn non-linear representations, capturing the complexity of XOR relationships.

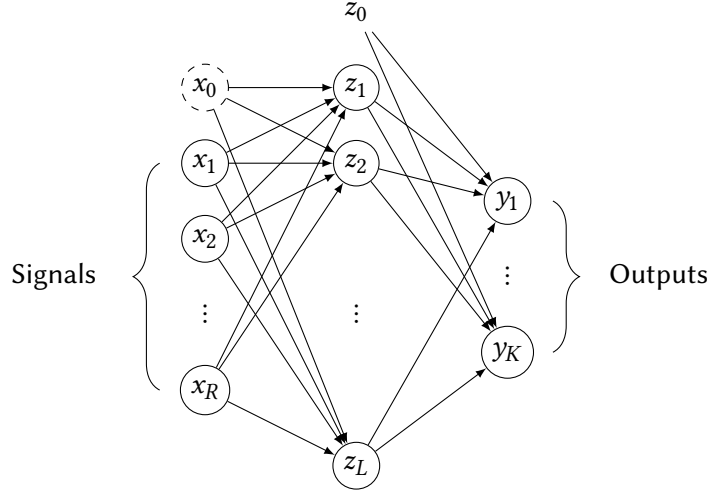
The invention of Neural Networks, also known as *MultiLayer Perceptron (MLP)*, which were afterwards further extended to *Deep Neural Networks (DNN)*, may be associated with the work of Rumelhart and McClelland (1987). This work is considered to be the first introduction of the *backpropagation* estimation algorithm to the wide public, although there are ongoing debates about who was the first to invent it. The main idea behind MLP was the possibility to combine simple neurons into a complex system, feeding the outputs of some neurons to other. For example, we can see how this approach resolves the XOR problem on Figure B.6.

Figure B.6.: MLP solution for XOR



The example of solution for XOR problem perfectly illustrates the main idea behind the MLP. We can see how the first layer of neurons serves for basis transformation, while the second one performs the classification in the new system of coordinates. The representation of MLP may be generalised, for example the case of 2 layer MLP may be represented as in Figure B.7. In this case we face K -class classification problem ($y_k \in \{1, 0\}, k \in \{1, \dots, K\}$), where hidden layer is composed of L neurons.

Figure B.7.: Multilayer Perceptron



In order to formally describe this new model we will need to add index identifiers to our existing notation. For an observation $x_i, i = 1, \dots, R$ taken as input, we can define the element z_l of the hidden layer as:

$$\forall l \in \{1, \dots, L\}, z_l = H^{(1)}(\omega_l^1 x_i) = H^{(1)}\left(\sum_{r=0}^R \omega_{lr}^{(1)} x_{ir}\right)$$

Where $\omega_l^{(1)}$ is the vector of weights associated with element l of hidden layer. The superscript (1) indicates that this vector belongs to the first matrix of weights. Assuming that the elements of hidden layer are not simply linear but undergo some sort of transformation $H(\cdot)$ as well.

The same procedure applies for the output layer, which takes the vector z as input. The Figure B.7 illustrates the general case of K -class *mono-label* classification, where each element is associated with an indicator vector:

$$\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, y = k \Leftrightarrow y : \forall j \in \{1, \dots, K\} = 1, y_{j \neq k}$$

This vector corresponds to the output layer on Figure B.7. We can express $y_k, k \in \{1, \dots, K\}$ as:

$$\forall k \in \{1, \dots, K\}, y_k = H^{(2)}(\omega_k^2 z) = H^{(2)}\left(\sum_{l=0}^L \omega_{kl}^{(2)} z_l\right)$$

Or, in a more complete form as:

$$y_k = H^{(2)}\left(\sum_{l=0}^L \omega_{kl}^{(2)} \times H^{(1)}\left(\sum_{r=0}^R \omega_{lr}^{(1)} x_{ir}\right)\right)$$

Once the model defined we may see the similarities with some other statistical techniques. We can observe, that the resulting model is a further development of *Generalised Linear Models (GLM)* and *Generalised Additive Models (GAM)* denoted *Projection Pursuit Regression (PPR)*. s class of models was

proposed by Friedman and Stuetzle (1981) as a method of non-parametric multiple regression. The idea was identical to the one behind MLP: to project the input data in the optimal direction before applying smoothing functions. The formal representation of PPR may be written as:

$$y_i = \sum_{l=1}^L g_l \left(\sum_{r=1}^R \omega_{lr} x_{ir} \right)$$

B.2.3.1. Backpropagation algorithm

Even though the ideas similar to the ones described by Rumelhart and McClelland (1987) appeared before, the NN design lacked the estimation algorithm. Even though it was proven, that such construct may easily resolve XOR problem and other similar questions, it was rather difficult to estimate. The *Backpropagation* algorithm extended the *Gradient descent* predecessor using the chain rule for partial derivation.

To define the *Backpropagation* term, we first should look at what *Propagation* is. As one may have guessed, the *Propagation* refers to the procedure of estimating y , given the weights ω and the inputs x . The objective of this procedure is to observe the *error* associated with current network configuration, which will later will be reused for weights update. Given the multi-class problem, there exist many different measures for error generalisation, but here we will use one of the simplest ones. Let's define $l(h(x), y)$ as:

$$\forall(x, y), l(h(x), y) = \frac{1}{K} \sum_{k=1}^K (h_k(x) - y_k)^2$$

The weights are then updated following the same rule as in the *gradient descent* algorithm:

$$\omega_{ab} \leftarrow \omega_{ab} - \eta \frac{\delta l(h(x), y)}{\delta \omega_{ab}}$$

Because in the case of MLP we face multiple layers, it was proposed to use Leibnitz's *chain rule* for derivation³. Using this rule we can obtain:

$$\frac{\delta l(h(x), y)}{\delta \omega_{ab}} = \frac{\delta l(h(x), y)}{\delta z_a} \frac{\delta z_a}{\delta \omega_{ab}} = y_a \frac{\delta z_a}{\delta \omega_{ab}}$$

Now, let's observe what happens in the case of MLP with only one hidden layer, as the one depicted in Figure B.7. If the unit $a = k, k \in \{1, \dots, K\}$ is in the output layer, we may write:

$$y_k = \frac{\delta l(h(x), y)}{\delta z_k} = H'(z_k) \times (h_k(x) - y_k)$$

In the case, when unit $a = l, l \in \{1, \dots, L\}$ is in the hidden layer, we can write:

³Chain rule expresses the derivative of two differentiable functions f and g such that $\forall x, h(x) = f(g(x))$. The resulting derivative is defined as $h'(x) = f'(g(x))g'(x)$.

$$\text{gamma}_l = \frac{\delta l(h(x), y)}{\delta z_l} = H'(z_l) \sum_{k \in Af(l)} y_k \times \omega_{kl}$$

Here we assume that $Af(k)$ represents the set of units in the output layer with $k \in \{1, \dots, K\}$.

B.2.3.2. MLP Algorithm

The resulting Algorithm 4 for MLP with two layers of weights (*omega* et *gamma*) may be defined as:

Algorithm 4 MLP (with 1 hidden layer)

Require: $\mathcal{S}, \eta > 0, T > 0, L > 0$

$\omega^{(0)} \leftarrow \mathbf{0}$

▷ Initialize weight vector

$\gamma^{(0)} \leftarrow \mathbf{0}$

$t \leftarrow 1$

▷ Initialize epoch count

while $t < T$ **do**

 Choose randomly an example $(\mathbf{x}^{(t)}, \mathbf{y}^{[t]}) \in \mathcal{S}$

Propagation part:

 Calculate $z^{(t)} = f_{\omega^{(t-1)}}(x^{(t)}) \leftarrow \omega^{(t-1)} x^{(t)}$

 Calculate $h_{\gamma^{(t-1)}}(z^{(t)}) \leftarrow \gamma^{(t-1)} z^{(t)}$

Backpropagation part:

$\gamma^{(t)} \leftarrow \gamma^{(t-1)} + \eta y^{(t)} \frac{\partial H^{(1)}}{\partial \gamma^{(t-1)}}$

$\omega^{(t)} \leftarrow \omega^{(t-1)} + \eta \frac{\partial H^{(2)}}{\partial \omega^{(t-1)}} \sum y^{(t)} \frac{\partial H^{(1)}}{\partial \gamma^{(t-1)}}$

$t \leftarrow t + 1$

end while

Ensure: $\omega^{(t)} \gamma^{(t)}$

The implementation in R will be:

```
# MLP function
# For one hidden layer and binary output
# Assuming linear transformations on all layers and linear loss
mlp = function(S, hidden = 10, eta = 0.01, epoch = 10) {
  # Initialisation of all the weights to 0 will prevent learning
  w = matrix(1, ncol = ncol(S) - 1, nrow = hidden)
  g = rep(1, hidden)
  t = 1
  while (t < epoch) {
    S_t = as.matrix(S[sample(1:nrow(S), 1), ])
    x_t = S_t[, 1:(ncol(S) - 1)]
    y_t = S_t[, ncol(S)]
    z_t = c(w %*% x_t)
    h_t = c(g %*% z_t)
    if (y_t * h_t <= 0) {
      g = g + eta * y_t * z_t
      w = w + eta * y_t * outer(g, x_t)
    }
  }
  t = t + 1
}
```

```

}
return(list(w, g))
}

```

Assuming the case of multi-class classification with y_i being a vector as described previously, we can rewrite the R code as:

```

# MLP function
# For one hidden layer and multiple class output
# Assuming linear transformations on all layers and linear loss
mlp_mult = function(S, hidden = 10, class = 2, eta = 0.01, epoch = 10) {
  # Initialisation of all the weights to 0 will prevent learning
  w = matrix(1, ncol = ncol(S) - class, nrow = hidden)
  g = matrix(1, ncol = hidden, nrow = class)
  t = 1
  while (t < epoch) {
    S_t = as.matrix(S[sample(1:nrow(S), 1), ])
    x_t = as.matrix(S_t[, 1:(ncol(S) - class)])
    y_t = as.matrix(S_t[, (ncol(S) + 1 - class):ncol(S)])
    z_t = as.matrix(c(w %**% x_t))
    h_t = as.matrix(c(g %**% z_t))
    if (t(y_t) %**% h_t / class <= 0) {
      g = g +
        eta *
        matrix(
          rep(y_t, each = hidden),
          ncol = hidden,
          byrow = TRUE
        ) *
        outer(rep(1, class), z_t)
      w = w + eta * y_t * outer(rep(1, hidden), x_t) %**% g
    }
    t = t + 1
  }
  return(list(w, g))
}

```

B.2.3.3. Universal approximator

After the invention of MLP the scientific community was focused on this class of models because of the advantages and flexibility it offered in comparison with more simplistic models such as OLS, GLM and even GAM. The main advantage was the possibility to approximate *any* function f , without imposing any additional restrictions and supposition. This allowed to bypass the limitation of the more simple models, for which it was necessary to introduce prior assumptions concerning the defined functional form. Cybenko (1989) was among the first to demonstrate this property for MLP with

sigmoid activation functions. Later, Hornik (1991) demonstrated that the results are not limited to some specific activation functions, but can be generalised for the whole family of the *feed-forward* MLP architecture.

B.2.4. Convolutional Neural Network (CNN)

The CNN is presented in the body of the thesis, thus it is omitted in the Appendices.

B.3. NN in Choice Analysis

There exist different approaches to introduction of NN base techniques into classical choice modelling. While in the main body of the thesis an overview of usage of ML models in general in choice analysis is offered, here a focus is made on the NNs usage.

There exist a number of studies focusing on the general comparison of ML methods in application to economic problematic, where among other tools NNs make appearance. For example, Hagenauer and Helbich (2017) examine individual travel mode choice. The authors use an extensive Dutch travel diary data spanning 2010 to 2012, augmented with variables encompassing the built and natural environment, along with weather conditions. Their work assesses the predictive efficacy of seven chosen ML classifiers for travel mode choice analysis. Results indicate that the random forest classifier outperforms all other examined classifiers, including the commonly employed multinomial logit model. Trip distance emerges as the most pivotal variable in this case study, although variable importance varying across classifiers and travel modes. Meteorological variables attain utmost importance for the support vector machine, while temperature assumes particular significance in predicting bicycle and public transport trips. Their findings demonstrate the possibility to explore the variable importance from the ML models in the context of economic studies. At the same time, the diversity of obtained results outlines the necessity of analyzing variable importance concerning diverse classifiers and travel modes for a more complete understanding of people's travel behavior.

However, there exist studies much more oriented towards the NNs exclusive usage. As the DNNs started gaining attraction in choice analysis due to their high predictive capability, S. Wang, Wang, and Zhao (2020) decide to explore in more in detail their capacities in providing some sensible economic indicators. The extent to which economic information can be interpreted from DNNs remaining uncertain even today. This work establishes that DNNs can furnish economic information on par with classical DCMs. The extracted economic information includes: choice predictions, probabilities, market shares, substitution patterns, social welfare, probability derivatives, elasticities, marginal rates of substitution, and heterogeneous willingness to pay. This information derived from DNNs may be unreliable in cases of small sample sizes, owing to three challenges associated with automatic learning: sensitivity to hyperparameters, model non-identification, and local irregularities. To illustrate the strengths and challenges, authors estimated DNNs using stated preference data from Singapore and revealed preference data from London. The extracted economic information from DNNs was compared with that from DCMs. The study revealed that aggregated economic information, is more reliable than disaggregated information at the individual observation or training level. Additionally, a larger sample size, hyperparameter tuning, model ensemble techniques, and effective regularization substantially enhance the reliability of economic information extracted from DNNs.

The attempts to merge the two approaches with addition of some transitional steps are not that rare in the literature. For example, among those works we may cite the recent publications of Sifringer, Lurkin, and Alahi (2020). The authors operate under assumptions that in the realm of DCA, inaccuracies in model specifications can result in limited predictability and biased parameter estimates. This paper introduces a novel approach to estimating choice models by segmenting the systematic utility specification into two components: (1) a knowledge-driven segment and (2) a data-driven one that learns a novel representation from available explanatory variables. This formulation enhances the predictive capabilities of standard DCMs without compromising their interpretability. The effectiveness of this approach is demonstrated by enhancing the utility specification of Multinomial Logit (MNL) and Nested Logit (NL) models with a non-linear representation derived from a Neural Network (NN), resulting in new choice models named Learning Multinomial Logit (L-MNL) and Learning Nested Logit (L-NL) models. Utilizing multiple publicly available datasets based on revealed and stated preferences, the study illustrates that the proposed models surpass traditional ones in terms of predictive performance and parameter estimation accuracy.

Another work with relatively similar interpretation is the publication of S. Wang et al. (2021). This study employs statistical learning theory to establish a framework examining the trade-offs between estimation and approximation errors, as well as prediction and interpretation losses in DNN. Interpretability in DNN-based DCM is implemented through metrics measuring the difference between true and estimated choice probability functions. Unlike traditional choice models relying on parameter estimation and manually crafted utility specification, DNN-based models rely on function estimation and automatic utility specification. The study uses statistical learning theory to upper bound the estimation error of prediction and interpretation losses in DNN, elucidating why DNN does not suffer from overfitting. Simulations comparing DNN to the Logit in three scenarios reveal DNN's superior performance in both prediction and interpretation, with larger sample sizes enhancing DNN's predictive power. In terms of results the DNNs proved to be more predictive and interpretable than plain Logit, unless complete knowledge about the choice task is available, and the sample size is small.

Yet another publication by Han et al. (2022) focused on the automated utility specification as ways to improve the model quality. Utility misspecification can result in biased estimates, inaccurate interpretations, and limited predictability. The solution to this problem involved a formulation comprising two modules: (1) a NN (TasteNet) learning taste parameters as flexible functions of individual characteristics, and (2) a MNL model with utility functions defined by expert knowledge. Taste parameters learned by the NN were then incorporated into the choice model, establishing a link between the two modules. This approach extended the L-MNL model (Sifringer, Lurkin, and Alahi 2020) by enabling the NN part to learn interactions between individual characteristics and alternative attributes. Furthermore, it formalized and reinforces the interpretability condition, demanding realistic estimates of behavior indicators (e.g., value-of-time, elasticity) at the disaggregated level, crucial for model suitability in scenario analysis and policy decisions. Employing a unique network architecture and parameter transformation, the model integrates prior knowledge and guides the neural network to produce realistic behavior indicators at the disaggregated level.

As it could be observed, most of the methods implement some adjustment to the existing ML techniques to comply with the classical choice theory and thus be *explicative* at least in some degree. Or incorporation of some ML driven components of the more classical theory driven choice analysis models.

B.3.1. CNN design for MNL imitation

The first ideas to adapt the MNL structure and inscribe it into the NN framework was published by Bentz and Merunka (2000), alongside with popular at the time comparisons on NNs capabilities to more basic Logit models (Abdelwahab and Abdel-Aty 2002; D. A. Hensher and Ton 2000) At the time this solution was proven to be inefficient, as it combined both the disadvantages of the choice models rigidity, alongside with the inconveniences of the lower explainability and high data availability requirements of the ML.

Nevertheless, for a better understanding we are obliged to start with the most basic model bundling together the concepts of CNN and MNL models. The transition from MNL to CNN is rather straightforward in this case. The MNL model already has the latent components integrated into its structure, which are the deterministic utilities V_j . Those deterministic utility elements may be viewed as elements of a hidden layer. However, in order to respect the structure of the classical MNL model with attribute and characteristic specific effects there is a need to impose further restrictions on the hidden layer generation functions. It is unreasonable in this case to use a fully connected network structure, because it will lead to the situation where all the inputs are involved in construction of each of the hidden layer elements.

There are different options to impose those restrictions over the model. One of the simplest ones is to use the convolutions to calculate the alternative specific utilities. And while the focus is made on the conditional Logit, where only alternatives' attributes play some role in utility computation, this is rather simple. In the case of complex mixes of individual characteristics, attributes and environmental effects the NN internal design might appear rather cumbersome.

Here is a relatively short example formalising the resulting NN structure. The designed CNN consists of two transformation layers. The first one is 1D convolutional layer⁴ with a linear activation function each. It takes as input the dataset in a *wide* format⁵, and produces a single value as an output for each alternative. This is effectively an equivalent of computation of deterministic utilities V_j in the context of classic MNL model. Thus we can interpret the elements of the last network layer, preceding the softmax transformation (z_k), as deterministic utilities V_j , assuming $j = k$. For simplicity, let's assume that each alternative is described by a set of d attributes, and no individual characteristics are simplified. This means that convolutional layer have a size and stride both set to d , assuming that all the alternative specific attributes are grouped by alternative. The second layer is a restricted softmax transformation layer, which directly applies softmax transformation over the inputs, without any supplementary permutations. The only restriction imposed for this layer is the absence of the weights to calculate for the algorithm. In the baseline MNL model the V_j are typically not subject to further transformations for the purposes of choice probability calculation.

The vector of inputs issued from the dataset transformed into the "wide" format can be represented as⁶:

⁴Meaning that each convolution produces a unique value.

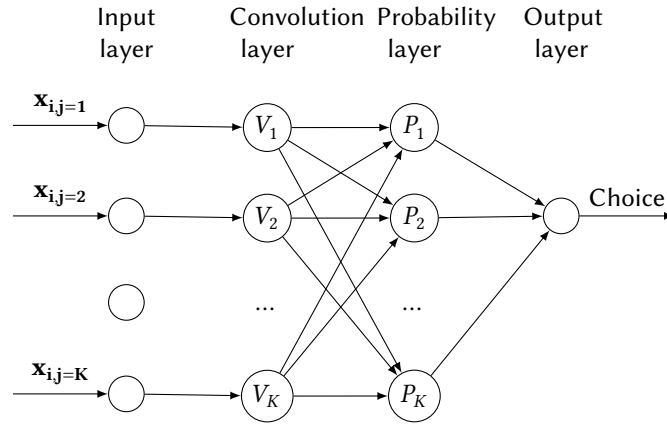
⁵Sometimes this data format is referred as *flattened* format. It assumes that all alternative specific attributes are input into the model in a single vector format.

⁶Although here we focus on a linear version of the model representation with a 1D convolution layer, it can as well take form of a 2D convolution layer. A matrix of $\mathbf{x}_{i,j}$ with K rows is constructed in this case, where each line corresponds to the alternative specific utility part.

$$\mathbf{X}_i = \{\mathbf{x}_{i,j=1}, \mathbf{x}_{i,j=2}, \dots, \mathbf{x}_{i,j=K}\}$$

Where $j = k \in \{1, \dots, K\}$ and each element $\mathbf{x}_{i,j}$ is a vector of d attribute values corresponding to the alternative j . This simple transformation allows us to use 1D convolutions alongside this vector with both size and stride equal to d , Ensuring that the input vector \mathbf{X}_i with dimensions of $d \times K$ will be transformed into a vector of relative deterministic utilities \mathbf{V} of size K . Each element of this vector is thus composed of a linear combination of single 1D convolution across $\mathbf{x}_{i,j}$. The graphical representation may be depicted in figure B.9.

Figure B.8.: Convolution Neural Network design



However, such simple interpretation poses some difficulties once there are individual characteristics involved, or once there are effects varying across alternatives (Conditional Multinomial Logit). To incorporate the variation between the alternative specific effects, as well as their interactions, we may adopt a more complex approach. Within the convolution on per-alternative basis we should include the effects structure information. This means that instead of simple concatenation of the alternative specific vectors $\mathbf{x}_{i,j}$, those should be remapped into a higher dimensionality space. Hence we introduce the new element $\hat{\mathbf{x}}_{i,j}$, which contains additional information on the coefficients varying between the alternatives. Let assume that an element $x_{i,j,r} \in \mathbf{x}_{i,j}$ is an attribute which has per-alternative effects. Assuming \mathbf{v}_j is a vector of zeroes and ones corresponding to the alternative index, such that for alternative j :

$$\mathbf{v}_j : v_k = \begin{cases} 1 & k = j \\ 0 & k \neq j \end{cases}$$

Thus to a vector

$$\mathbf{x}_{i,j} = \{x_{i,j,1}, \dots, x_{i,j,r}, \dots, x_{i,j,R}\}$$

Corresponds a new vector:

$$\hat{\mathbf{x}}_{i,j} = \{x_{i,j,1}, \dots, \mathbf{v}_j \times x_{i,j,r}, \dots, x_{i,j,R}\}$$

This procedure may be performed for multiple elements of $\mathbf{x}_{i,j}$ and performs more complex structures, tying together the values of several alternatives, while leaving some other out. This leaves us with the final vector of type:

$$\hat{\mathbf{X}}_i = \{\mathbf{x}_{i,j=1}, \mathbf{x}_{i,j=2}, \dots, \mathbf{x}_{i,j=K}\}$$

Finally, the size and strides of convolution window should be adjusted accordingly. As this method ensures that size of $\mathbf{x}_{i,j} \forall j$ is identical, the size of convolution window corresponds precisely to the the length of the $\mathbf{x}_{i,j}$ vector.

B.3.2. Alternative Utility Specific DNN (ASU-DNN)

A logical extension of this baseline model was introduced by S. Wang, Wang, and Zhao (2020) (and S. Wang et al. (2021)). The shortest description of the new method may be seen as:

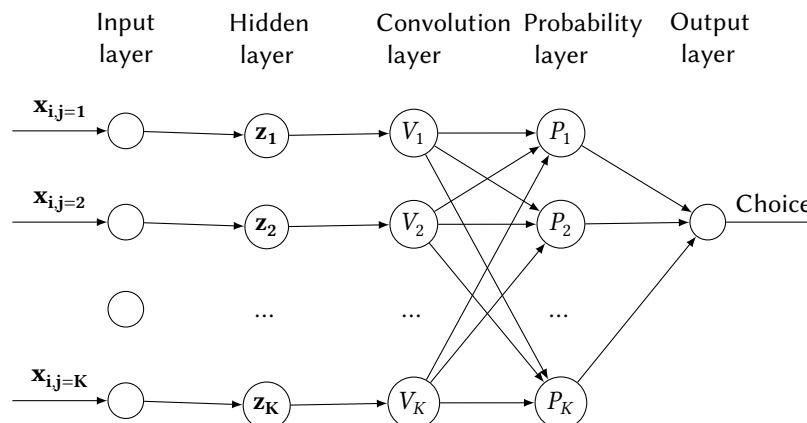
It could be considered as a stack of fully connected subnetworks, with each computing a utility score for each alternative.

The key modification involved addition to the Fully-connected Deep Network layers between inputs and deterministic utilities, thus ensuring that the $V_j = \mathcal{F}(\mathbf{x}_j)$, where \mathcal{F} reflects the FDN transformations. The FDN layer in the original paper are assumed to have ReLU activations.

Another change involves the relaxation of the convolution layer restrictions, as nothing in the work of S. Wang, Wang, and Zhao (2020) indicates on the usage of convolution layer in strict sense, but rather a dimension reduction of the hidden layers.

The resulting model thus respects the baseline utility theory, assuming that V_k is independent from the $\mathbf{j} \neq \mathbf{k}$. The alternative models are equally possible, but are deprived of logical interpretation. For example, one may introduce a Fully-connected Deep Network prior to convolution layers across all alternatives' inputs, which would imply that different alternatives' attributes influence the alternative specific deterministic utilities V . Or one could imagine the introduction of several fully-connected layers after the convolution and the probability computations, which would be nearly identical from the interpretation point of view.

Figure B.9.: Convolution Neural Network design



B.3.3. Extracting interpretable information from NN

We have observed the history of the NN's evolution and development. Now, more complex NN architectures emerge, allowing to estimate even more flexible functional forms. However, for economists and econometricians the predictive power of such models is of little use. The economics focuses on the exploration of effects, causal relationships and derivation of other indicators (Michaud, Llerena, and Joly 2012), which limits their interest to the unexplainable modelling techniques.

Extracting interpretable information from NNs is crucial for understanding the decision-making processes of these complex models. NNs, especially DNNs, are known for their ability to capture intricate patterns in data, but their lack of interpretability has been a challenge. Several methods and techniques are popular in ML sphere for enhancement of the interpretability of NNs⁷:

- Feature Importance Analysis (Li et al. 2017), focusing on identification of the most important features in a neural network's decision-making process. This is typically achieved with analysis of how the model's predictions change when features are modified.
- Activation Visualization (Kahng et al. 2018), exploring the activations of neurons in different layers of a neural network provides insights into what each layer has learned.
- Saliency Maps (Nagsubramanian et al. 2018), which are generated by computing the gradient of the output with respect to the input. Thus reflecting which input features contribute most to the prediction.

Those are only a fraction of the available methods used by data scientists to obtain some interpretations of their models in depth behaviour. Among other techniques we may cite: Layer-wise Relevance Propagation, usage of interpretable architectures and activation maximisation. At this point the transition to the interpretable NNs for economists is rather straightforward, as using the knowledge of economics and theoretical assumptions on human behaviour it is possible to implement an interpretable architecture for choice analysis.

Consequently, the NN models may be interesting for economists. Even though it is impossible to directly obtain information about confidence interval for weights, there is still the possibility to derive some key metrics. The recent work of S. Wang, Wang, and Zhao (2020) (and S. Wang et al. (2021)) shed some light on the procedures of extracting such information. The key idea of the work is that DCM are extremely similar to the ML models with *softmax* output layer. Here we can see the *softmax* transformation:

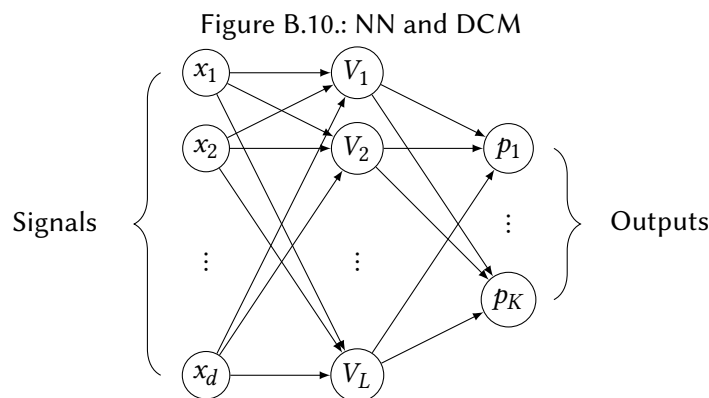
$$y_k = \frac{e_j^z}{\sum_{l=1}^L e_l^z}$$

From the theoretical point of view, this transformation is identical to one observed in *Multinomial Logistic Regression (MNL)* and an entire family of other choice models. This allows us to treat the entries to such *softmax* transformation as deterministic utilities ($V_k \forall k \in \{1, \dots, K\}$). The required NN structure is represented on Figure B.10.

⁷The references provided in this part of the appendix should be viewed as examples of the described methods.

Table B.2.: Extracting information
[H]

| <i>Economic information</i> | Formula |
|--|--|
| Choice probability of class k | $p_k(x_i)$ |
| Choice prediction | $\arg \max_k p_k(x_i)$ |
| Market share | $\sum_i p_k(x_i)$ |
| Substitution between alternatives k_1 and k_2 | $p_1(x_i)/p_2(x_i)$ |
| Social welfare | $\sum_i \frac{1}{\alpha_i} \log(\sum_k e^{V_{ik}}) + C$ |
| Probability derivative of alternative k with respect to x_{ij} | $\delta p_k(x_i)/\delta x_{ij}$ |
| Elasticity of alternative k with respect to x_{ij} | $\delta \frac{p_k(x_i)}{\delta x_{ij}} \times \frac{x_{ij}}{p_k(x_i)}$ |
| Marginal rate of substitution between x_{ij_1} and x_{ij_2} : | $-\frac{\delta p_k(x_i)}{\delta x_{ij_1}} / \frac{\delta p_k(x_i)}{\delta x_{ij_2}}$ |
| Value of Time (assuming x_{ij_1} is time and x_{ij_2} a monetary value): | $-\frac{\delta p_k(x_i)}{\delta x_{ij_1}} / \frac{\delta p_k(x_i)}{\delta x_{ij_2}}$ |



Where the deterministic utilities, associated with k class V_k may have non-linear form. Compared to simple MNL model, such approach allows V_k to be extremely non-linear, but may limit the results interpretation. The work of Q. Wang et al. (2020) proposes us a solution, which may be rather limited for now, but which may be further extended. The proposed metrics may be summarised in the Table B.2.

It is important to remember, that even though it is possible to extract some useful information from the NN models, the implications behind such metrics differ from the typical DCM paradigm. This idea may be summarised as follows (Q. Wang et al. 2020):

... the interpretation of DNNs is a prediction-driven process: the economic information is generated in a post-hoc manner after a model is trained to be highly predictive. This prediction-driven interpretation takes advantage of DNNs' capacity of automatic feature learning, and it is also in contrast to the classical DCMs that rely on handcrafted utility functions. This prediction-driven interpretation is based on the belief that "when predictive quality is (consistently) high, some structure must have been found"

B.4. Conclusion

In conclusion, this appendix delves into advanced and less conventional modeling techniques, often overlooked in economic studies due to perceived limitations in offering insights into effects estimation. ML techniques, notably NN, are sometimes regarded as black boxes by economists, focusing on predictive power rather than revealing the underlying process. The ML paradigm emphasizes results over processes, with models relying on assumptions of independence and identical distribution similar to those in Econometrics and SHS. However, the emphasis in ML is on functional form flexibility with large samples.

The formalization of the learning problem involves constructing a function that predicts an output with minimal error based on observed pairs generated from an unknown distribution. The appendix introduces artificial neurons, perceptrons, Adaline optimization, Multilayer Perceptron, and an initiation to the Backpropagation algorithm. While convolutional neural networks are omitted from the appendices, the discussion extends to the implementation of these methods for economic choice analysis. Economists prioritize exploring effects, causal relationships, and deriving indicators, limiting their interest in less explainable modeling techniques. Despite this, recent work demonstrates potential applications of NN models for economists, offering insights into key metrics even though direct information about confidence intervals for weights may be elusive. The convergence between Discrete Choice Models and ML models with softmax output layers opens avenues for further exploration and integration of these methodologies.

C. Independence from Irrelevant Alternatives

In this appendix we attempt to shed some light on the history and implications of the *Independence from the Irrelevant Alternatives (IIA)* hypothesis. It is one of fundamental properties in the context of choice theory and social choice theory. The concept assumes that the introduction to the choice set or removal of irrelevant alternatives should not impact the relative preference orderings among the remaining alternatives. In other words, if a certain alternative is chosen over another in a given set of options, the addition or removal of additional alternatives should not alter this preference relationship.

This principle has implications in various fields, including economics, political science, and operations research. In most of those cases the stability of preferences and choices is crucial for modeling and predicting individual and collective decision outcomes. Among the most popular presentations in various handbooks on economics and econometrics one may encounter the famous *red and blue bus paradox*. At the same time the origins of the concept remain relatively obscure, especially in the context of the choice analysis studies. Nevertheless, this is a concept with a long and interesting history, taking its roots in studies of individual decision making in votes. What is more, the restrictions imposed by the IIA are sometimes more severe than described in some of the handbooks and may not be as easy to detect and bypass.

C.1. Traditional formulation of IIA

In the econometrics course one typically encounter the IIA during the discussion of the Discrete Choice Models (DCM) and similar techniques. Moreover, the discussion of the IIA is never initiated before the introduction of the Multinomial Logistic regression concept, which heavily relies on this hypothesis. For example, in Greene and Hensher (2003) we may encounter the definition of the introduction of the IIA through the MNL model derivation, where the IIA property is derived from the models' architecture itself. In this section we are going to trace some of the considerations described in the handbook of Greene and Hensher (2003).

The reader is placed in the typical context of the DCM framework. We consider that individuals are (i) fully rational and (ii) respect the utility maximisation principle. In other words, the individuals are supposed to always choose the alternative (k) presenting the highest utility ($U_k > U_j, \forall j \neq k$) from a given set of alternatives ($\Omega = \{1, \dots, K\}$). Through imposition of additional restrictions on the utility form and error distributions the MNL model is derived. More precisely, the utility (U_k) is assumed to regroup two elements: (1) a deterministic part V_k and (2) a random term ε_k .

$$U_k = V_k + \varepsilon_k$$

We will not introduce the complete proof of the IIA derivation, but will rather synthesise it to the bare minimum. An identical presentation may be encountered in Luce (1957). Consider the MNL model, where the probability to choose the alternative k by an individual i is given as:

$$P_{ik} = \frac{e^{V_{ik}}}{\sum_j e^{V_{ij}}}, i \in \{1, \dots, N\}, k \in \{1, \dots, K\}$$

Consequently, the probability odds of choosing an alternative k over an alternative $l \in \{1, \dots, K\}$ is defined as:

$$\frac{P_{ik}}{P_{il}} = \frac{\frac{e^{V_{ik}}}{\sum_j e^{V_{ij}}}}{\frac{e^{V_{il}}}{\sum_j e^{V_{ij}}}} = \frac{e^{V_{ik}}}{e^{V_{il}}} = e^{V_{ik} - V_{il}}$$

As one can see, this probability ratio depends only on the deterministic utilities of two alternatives in question. In traditional DCM (ex: McFadden (1974)) the deterministic utilities V_{ik} are assumed to be independent of the other alternatives, which leads us to the IIA definition. The probability ratio should not be affected by addition or subtraction of *irrelevant alternatives* from the set Ω .

Greene and Hensher (2003) proposes as well another understanding (or rather another consequence of the IIA presence in the DCM context) speaking about the cross-elasticities of the different alternatives. For us, it is one of the many straightforward transitions from the IIA concept to the more familiar economic indicators. The cross-marginal effects of changes in probability to choose an alternative k with changes of some attribute of the alternative l : z_l , may be written as¹:

¹here we ignore the individual index i for simplicity

$$\frac{\partial P_k}{\partial z_l} = \frac{\partial P_k}{\partial V_l} \times \frac{\partial V_l}{\partial z_l} = -P_k P_l \beta_{zl}$$

Where β_{zl} stand for corresponding effect in the linearly defined deterministic utility V_l for the corresponding variable z_l . The chain rule is used for the derivation procedure, one may find more about this in Greene and Hensher (2003). The transition to the cross-elasticity once we have defined the marginal effects is simple. The cross-elasticities represent the changes in the attributes of one alternative on the choice probabilities of other alternatives. Under IIA the inclusion or exclusion of an irrelevant alternative should not impact the relative odds between the remaining alternatives. This relationship might be verified:

$$\frac{\partial P_k}{\partial z_l} \times \frac{z_l}{P_k} = -P_l \beta_{zl} z_l$$

The cross-elasticity in this case is not influenced by any external variables, but fully depends on the variables describing the alternative l : the attribute z_l and the probability P_l .

At this point we fully understand the links between the MNL and IIA, but there is little insight on how to bypass it, neither there is evidence on how other statistical models behave. Greene and Hensher (2003) provides an extensive discussion on how to overcome the limitations imposed by the IIA in a traditional MNL, which we are going to see later. However, there is little information on whether the IIA restrictions remain on other, more complex statistical or Machine Learning (ML) models. What is more, in the handbook, as well as in many other books providing a discussion on presence of IIA in MNLs, we find little insight into the links between IIA and the behavioural theory. Yes, obviously, we encounter some examples (ex: the famous “red and blue bus” paradox), but no clear anchoring in the decision theory. The following sections of this document will provide more information about these two points.

C.2. History and ambiguity of the IIA

In the book of Greene and Hensher (2003) in the part consecrated to IIA description, we encounter references for Luce (1957) work. The work on “*Games and Decision*” describes some of the key utility properties, where we encounter the IIA as well. But Luce was not the only one to describe the IIA property of rational decision making, it equally appears in the works of Arrow (1951) and latter, under different angle in the work of Radner and Marschak (1954). The history may be traced even further to J. F. Nash (1950), Savage (1951) and further till Daunou (1803) and Condorcet (1785), as summarised in the article of Gensch and Ghose (1997). One of the interesting reference in this context is the production of Ray (1973), which contrasts the different formulations of the IIA rule. Let us see what are these definitions, their common points and differences. Here we follow the explanations on the terms differences offered by McLean (1995) and later revisited by Patty and Penn (2019).

C.2.1. IIA(A) by Arrow (1951)

In this first case the authors focus on political science and voting rule, leading to a definition of IIA in this restrictive context:

Let $\{R, \dots, R_n\}$ and $\{R', \dots, R'_n\}$ be lists of orderings, $C(S)$ and $C'(S)$ be social the social choice functions. If for all i (individuals) and all x, y in given S $xR_i y$ if and only if $xR'_i y$, then $C(S)$ and $C'(S)$ are the same.

In other words, Arrow states that choices made from fixed S should be independent from alternatives from outside. Arrow devised this rule in the context of the collective choice rules exploration, defining IIA as one of the requirements for such choice rules. In contemporary literature we may encounter some other interpretations of this statement. For example, Patty and Penn (2019) writes:

Choice rule cannot respond to members' preferences for irrelevant alternatives not contained in some set S , when making judgement over the alternatives in S

C.2.2. IIA(RM) by Radner and Marschak (1954)

Following the ideas expressed by Arrow (1951), Radner and Marschak (1954) developed their own IIA axiom as a fundamental principle in the context of consumer choice theory and utility theory. This IIA axiom posits that the introduction or removal of an alternative should not affect the relative *preference ranking* of the remaining alternatives in a choice set.

If x is an element of choice set S and belongs to S_1 , constrained in S , then x is also an element of the choice set of S_1 .

The focus in this case shifts to choice sets and rankings. In other words, the probability of choosing one alternative over another should not be influenced by the presence or absence of irrelevant alternatives. We can rewrite this statement using the same notation as for the Arrow's definition as:

$$\{x \in C(S) \text{ and } x \in S_1 \subset S\} \implies x \in C(S_1)$$

C.2.3. IIA(L) by Luce (1957)

Completing the works of previous authors, Luce's work laid the foundation for the study of preferences and choices in the context of mathematical models, and choice modelling in general.

Let $P_s(x)$ be the probability of x being chosen from S , and let $P(x, y)$ be the probability of x being chosen from $\{x, y\}$. For $x, y \in S \subset T$, a finite set: if $P(x, y) \neq 0, \forall x, y \in T$, then for any $S \subset T$, such that $x, y \in S$

Here the focus shifts to the probabilistic definition of the choice, and preference ordering by consequence. The likelihood of selecting one option over another should not be influenced by the presence or absence of irrelevant alternatives. The mathematical formulation is rather simple, as given two sets of alternatives, one containing another, the probability ratios should hold:

$$\frac{P(x, y)}{P(y, x)} = \frac{P_s(x)}{P_s(y)}$$

C.2.4. Contraction consistency by J. F. Nash (1950)

In the fundamental contribution of Nash the notions similar to the IIA principle may also be observed. Denoted as *contraction consistency*, the property assumes that if a set S having a solution point is contained in another set T , the the solution is not affected by the elements in T and not in S .

If set T contains a set S and solution point $c(T)$ for given individuals u_1 and u_2 is in S , then
 $c(T) = c(S)$

In other words it implies that a reduction of set T to any set $S : \{c(T), \dots\}$ should not affect the equilibrium. As we can see, this definition is very similar to other definitions of the IIA, and especially to the one proposed by Radner and Marschak (1954). Later we may encounter a closely similar formulation in the definition of the *Sen's property- α* for choice sets (Sen 1983).

C.2.5. Criticism of the minimax decision theory by Savage (1951)

Contrary to the previously introduced works, which mainly focused on the IIA description, the contribution of Savage (1951) in reference to the work of Von Neumann and Morgenstern (1947) was among the first to observe the eventual inconsistency of IIA. Everything started with a criticism of minimax optimisation related to its application in situations of uncertainty.

There are cases in which if S is enlarged a new minimax regret solution is obtained and which differs from the previous one, yet is constrained in the original S .

We encounter here a description of the particular case of violation of the IIA property, encountered when a minimax decision rule is applied. Savage (1951) emphasizing the importance of probability distributions and subjective probability assessments in decision-making under uncertainty. Effectively, this only indicates the differences between the existing decision making algorithms and theories, but it is an important historical point. This case helps us to understand the in-depth nature of the IIA property and the cases when it may not be respected.

C.3. Linking the IIA with reality

In more recent literature we encounter less theoretic considerations about the nature of the IIA. As science links all the different disciplines; different understandings and interpretations of the IIA and different translation of its impacts on the reality might be encountered.

Blau (1971) states, that weakening the strict pairwise definition of the IIA(A) to m -IIA(A) may potentially make such IIA(A) hypothesis more realistic in the context of the individual choice. The proposition is to detect the independence from the irrelevant alternative not in pairwise context, but rather in some m constrained minimal sets.

Gensch and Ghose (1997) points out, that not only the different formulations of the IIA exist, but that the same type of IIA may have different types of consistency: (1) individual vs. aggregated level; and (2) single pair vs whole choice set IIA respect. The work assumes, that it is impossible to meet IIA at aggregate level without perfect choice homogeneity on individual level.

This way Saari (1999) indicates that IIA by its nature requires the aggregation rule in collective decision making to ignore whether individuals' preferences are transitive or not. The assumption is made, that

pairwise ranking ignores information about any other ranking, which may potentially lead to biases in modelling procedure.

David A. Hensher, Rose, and Greene (2005) relates the IIA to the models' errors distribution, pointing out that *Identical and Independent Distribution (IID)* of the residuals is equivalent to IIA behavioural assumption. This point actually explains the appearance of the Mixed MNL (MMNL) models as one of the potential solution of the IIA non-respect.

Cato (2013) speaks about the IIA in the context of the “*connectedness*” introduction, extending IIA to the subsets:

... given set Ω collection of subsets of set of alternatives, social ranking over $S \in \Omega$ depends on the individual ranking over menu S

Sen (2014) states in his exploration of the IIA, that it ignores the context of choice situation and thus is not realistic enough to be used as one of the key determinants of the rational behaviour. As an example he proposes the situation, where an individual, constrained with the social norm will never take the last apple from the basket. But this same individual, will eagerly take the same apple, were we to add one more to this same basket, which creates a rationality paradox, violating the IIA assumption.

Finally, Benson, Kumar, and Tomkins (2016) propose an extremely interesting study of whether or not the IIA hypothesis should be accounted for in the MNL models or not. The authors propose a battery of statistical tests to verify, whether the IIA is violated often enough in the traditional Choice Modelling (CM) datasets, to worry about it.

From the list of given above studies exploring the IIA property in CM, we may observe, that questions about IIA arise mostly in the situations where rational behaviour cannot be assumed. For example, real-world preference data, preference experiments or biological behavioural studies, all of them question the presence of the IIA. Among the most popular types of IIA inconsistency, we usually encounter the case of *context dependent preferences*, which arises in economics, psychology and neurobiology (Chung et al. 2017).

C.4. Mitigation of the IIA inconsistency

On this point we may switch to the possible solutions to the IIA non-respect in the Discrete Choice Models (DCM). Mitigating the IIA inconsistency in choice models is a critical challenge in discrete choice analysis. Researchers and practitioners employ various strategies to address this issue and enhance the realism of choice behavior. This section may be roughly divided into two interconnected part: (1) the methods for IIA validation and violation detection, and (2) the techniques (classes of models even) that allow to address the issue, was the inconsistency discovered.

C.4.1. IIA tests and validation

We start with the exploration of the potential solutions for IIA verification. Testing and validating the IIA assumption is a crucial step in assessing the reliability of discrete choice models. The main approaches have been proposed by the IIA discover or themselves, because of the definitions they used. It may appear quite logical to just verify the existing setup against the criteria given in definition.

David A. Hensher, Rose, and Greene (2005) describes quite well the two main approaches implemented in econometrics, when testing for the presence of IIA(L). It is quite natural, that because of the nature of the IIA(L) it can be extended to any other probabilistic models of the same class (ex: those, where the *softmax* transformation is used, such as described by S. Wang, Wang, and Zhao (2020) or earlier by ..). Consequently, all the considerations listed below may be extended over a larger class of models.

Starting in the historical order, we may look into the implications of the Luce's definition of the IIA(L). The most natural solution will be to test the probability ratio observed in one sample, over another ratio obtained for another sample. To formalise this approach we may say, that we compare the ratio P_x/P_y from $S = \{x, y\}$, against another ratio P'_x/P'_y from $S' = \{x, y, z\}$. This is a rather simple idea of how to perform a test of IIA hypothesis, but no test statistics was initially provided. Latter, the ideas of this test were adopted by McFadden in his works (McFadden 1987).

Before proceeding with the description of more tests, we should focus our attention at so called "*Anna Karenina principle*". This is an extremely well suited concept in the discussion about the IIA problematic. The initial phrase of the Leo Tolstoy may be altered to meet our problematic: "*there is only one way to be rational, but there are many ways to deviate*". We may understand in manner, that there exist many potential sets and subsets of alternatives, which may or may not present the IIA: "*more there are sets, more there are ways to deviate from IIA*". To illustrate this, we may take the simple case of the transportation mode choice as an example. Imagine that we attempt to model the choice of *taking a bus* against *not taking a bus*. In this case there are no other potentially relevant alternatives, that we may include into the choice set, which would still be mutually exclusive and consistent with those, already present in it. By the design itself the model does not allow for IIA to be violated (in this case, speaking about IIA has no sense at all). Now imagine we change the experimental design and decide to model the choice of *taking a bus* against an alternative of *taking a car*. The potential alternatives which may be or not relevant are many: *taking a plane, renting a boat, driving a bike* and so on. Consequently, to verify the validity of IIA, we should iteratively test our initial probability ratio, obtained for starting set of two alternatives, against all the other potentially relevant or irrelevant alternatives². And if adding *taking plane* to our choice set does not affect the probability ratio for the two initial alternatives, it does not mean that adding the choice of *driving a bike* will not do otherwise. Nevertheless, many test of IIA rely on assumption that a more general choice set Ω hold correct model specification.

Considering the described above complexity of the IIA test construction, the latter tests focused mostly on another perspective. Those more precise tests appeared after the work of Luce (Luce 1957). Among them, we may identify several, that were among the first ones to emerge in the works of:

- Horowitz (1982)
- Hausman, Hall, and Griliches (1984)
- Small and Hsiao (1985)

The key idea behind the majority of these tests was quite similar to those, described by Luce (1957). The main divergence among the tests was rather the concept of reducing the existing choice set, instead of its extension. Now, the ratio P_x/P_y from $S = \{x, y, z\}$, was compared against another ratio P'_x/P'_y from $S' = \{x, y\}$. In order to obtain a valid test statistics, such comparison was performed indirectly. Authors proposed to compare the differences in parameter estimates for various model specifications and not the probability ratios directly. Such approach managed to partially reduce the magnitude of problem's

²In this particular case, by design we primarily assume that all other alternatives are irrelevant

complexity, because now the initial choice set was assumed to be either complete, or containing some of the irrelevant alternatives. In practice when such tests for IIA consistency are performed, they are performed only against one of the alternatives. Meaning that only one of the alternatives is excluded, and not each and every one iteratively. Usually there is simply no possibility to test all the imaginable specifications of the choice set extensions or contractions to provide a sufficiently consistent answer. Here we are going to present the most popular among those tests, which is implemented in the majority of statistical software packages: the test of Hausman-McFadden (Hausman, Hall, and Griliches 1984) and McFadden test (McFadden 1987).

C.4.1.1. Hausman-McFadden test

This test is based on the observation that under IIA, parameters for choosing a subset of alternatives can be estimated using a MNL model either on that subset or on the entire set, leading to identical estimates. In case IIA does not hold, the parameter estimates for the full set become inconsistent, while those for the subset remain consistent when properly selected. This test involves conducting two MNL model construction and assessing the differences in parameter estimates. In a procedural manner the test might be written as:

1. Estimate the *model* (MNL type model is assumed) on a set of alternatives Ω
2. Re-estimate the *model* over a subset of alternatives $\Omega' \in \Omega$
3. If the estimates are different, the IIA does not hold

C.4.1.2. McFadden test

Another test attributed to McFadden focuses on the inclusion of cross-effects into the model in order to validate that IIA holds.

1. Assume the choice set of alternatives is Ω' and an alternative $K + 1 \in \Omega$ is added to this set
2. Estimate the *model* over a given subset of alternatives Ω' adding cross-effects to explicative variables set
3. If the cross-effect variables are significant the IIA does not hold

C.4.2. IIA treatment

As we can see, the testing of the IIA hypothesis is a rather complex task. Because of the complexities in definition of the correct test statistics and the overall ambiguity in the IIA validation and verification, around the years 2010 the test of IIA were disregarded as something unreliable (Fry and Chong 2005; S. Cheng and Long 2007). It was demonstrated that the traditional IIA tests are not efficient in small samples and are rather unreliable in general.

Anna Karenina principle has yet another interpretation in the context of DCM modelling. There is only single behaviour pattern, which is considered to be *rational* in the traditional behavioural sciences. However, there are many ways to deviate from this *rational* behaviour. The most straightforward translation of this idea is linked to the *behavioural model* concept itself and the rationality definition. If one remembers the traditional behaviour theories based on the expected utility, such as Nash's theory (J. F. Nash 1950) or McFadden's modelling techniques (McFadden 1974), usually impose several restrictions

on the individual's behaviour for it to be considered rational and be easily identifiable. This leads us to the situation, where violation of any of the imposed restriction renders the individuals irrational, each in different manner. Obviously, there is no possibility to account for all the potential biases at the same time, the usual procedure incorporates testing and mitigation steps. Such approach assumes that when a specific model is applied to address the IIA inconsistency, the researcher knows exactly what biases he / she attempt to regulate.

The main strategies to the mitigation of the IIA inconsistency were summarised in Haynes, Good, and Dignan (1988). There exist only two of them, which are:

1. Explicitly model the *alternatives substitutability* within the systematic component of utility V (through a Nested Logit, for example)
2. Implicitly model the *alternatives substitutability* through error structure ε (using Mixed MNL model types)

C.5. Conclusion

In conclusion, this appendix provides insights into the history and implications of the IIA hypothesis, a fundamental concept in choice theory and social choice theory. The IIA assumes that the introduction or removal of irrelevant alternatives should not impact the relative preference orderings among the remaining alternatives. In econometrics courses, IIA is introduced in the context of discrete choice modelling, particularly as a part of MNL regression. Recent literature discusses diverse interpretations and impacts of IIA across disciplines.

The appendix delves into theoretical considerations, definitions, and historical formulations of IIA. Different formulations of IIA are introduced, considering individual and aggregated level consistency and single pair against whole choice set IIA understanding. Mitigating IIA inconsistencies in DCM involves two interconnected parts: methods for IIA validation and violation detection and techniques to address inconsistencies when detected. IIA testing is acknowledged as a complex task, with traditional and widely available tests criticized for unreliability and inefficiency in small samples.

Addressing IIA inconsistencies involves testing and mitigation steps. The main strategies, explicitly modeling alternatives substitutability (e.g., Nested Logit) and implicitly modeling it through error structures (e.g., Mixed MNL models), are outlined. The conclusion underscores the ongoing challenges and evolving perspectives in understanding and dealing with IIA in the realm of discrete choice analysis.

D. Research practices: Unstructured interviews

This appendix presents the study of research procedures in application to the discrete choice analysis and the choice experiments in particular. Through a semi-structured interviews we collect information on the individual habits and customs of data analysis and research procedures construction among scientists. The resulting information is then processed for a better understanding and generalisation of the adopted scientific procedures. A particular attention is paid to the applied studies and the research focused on the applied problematic.

The work starts with the presentation of motivation for this particular study. The second part focuses on the methodology of unstructured and semi-structured interviews in application to the particular use-case. The final section described the data collection, as well as a part of obtained results, completing the study. The elements drawn from the collected data have a particular impact on the final version of the performance comparison framework proposed in the thesis.

D.1. Motivation and research objectives

The understanding of the research practices is extremely important in the context of the *meta-* and *inter-*disciplinary studies. Depending on the application case, discipline and individual background the research practices vary. While some of those changes are dictated by the established conventions in the community and verified research practices, another part might be dictated by conveniences or personal preferences. The last counterpart is the most difficult to explore and analyse, because contrary to the well documented research practices, which are rather uniform within a single domain, the individual preferences and arbitrary decisions made throughout research procedure are rarely presented in publications.

Speaking about the existing literature on *research procedures* it is rather scarce. While there exist a multitude studies on research procedures in the context of clinical studies and clinical procedures in working with animals, the general literature on research procedures is rather rare. As one of the rare examples we can cite the work of (Seroussi 1995) focuses on the research procedures introduction for students. In the work the approach to solving problems with heuristic hypotheses involves making initial approximations of unknown values, which are then refined during the problem-solving process, is criticised. Among the more recent studies we may encounter the work of Paul et al. (2021) and focusing on scientific procedures for systematic literature reviews, which is a rather specific use-case. Numerous articles offer guidance on literature reviews, but few provide a clear and trustworthy protocol that researchers can follow confidently. The article concludes with examples of systematic literature reviews featured in the inaugural special issue.

The main reasons behind this study concerns the exploration of scientific habits and conventions present among economists working with *Choice Modelling* (CM) and *Choice Experiments* (CE) in particular. This major aim may be separated into two sub-objectives, each as important as another.

Primarily, the scope of the study is delimited by the research field, potential applications and treated questions. As specified previously the *epistemology models* lying a the heart of the research may differ between the research fields. What is more, the different level of experience will also affect the perception and comprehension of the framework.

The first objective is to uncover the *Hidden Aspects of CM and CE Scientific Procedures*. Here the primary focus is made on the understanding of the research habits and unspoken conventions among the practising researchers. The possibility to gain insight into the genuine scientific procedure employed, rather than the version eventually presented in research publications, is particularly important. The goal is to pinpoint the different stages, functional components, and overall structure of CM and CE tasks. A better understanding of those elements may simplify the task of performance analysis framework creation, better adjusting the framework's elements for real-life use-cases.

The second objectives concerns the *highlighting of Critical Stages* of the research procedures. The identification of the key stages within the research procedure, which pose the most difficulties to the researchers may prove itself extremely valuable. It is assumed that knowing such stages in the research procedure may help to adjust the proposed performance analysis framework for best efficiency.

Furthermore, among the most valuable comments and observations, we should emphasize the importance of precisely defining the target audience for the resulting tool. Who will benefit from this framework? The task of creating the framework should be perceived as a product for the purposes of

this work. While the research question primarily focuses on the comparability of models and their performance in discrete choice models, it is crucial to acknowledge that the ultimate product intended for the audience is the framework itself. This framework aims to streamline operational decision-making processes, typically necessitating substantial expertise. Therefore, it is imperative to gain a thorough understanding of the requirements of the target audience, which, in turn, entails the appropriate delineation and characterization of said target audience.

D.2. Methodology

For the purposes of this study it was decided to conduct a series of *semi-structured interviews*. Which represent a mix of classical in science *structured interviews* and more predominant in journalism *unstructured interviews*. For a better understanding of the applied methodology there is a need to address first both of the extreme cases existing in interview methodology.

The *unstructured interviews* are one of the possible tools for conducting qualitative analysis. While the unstructured interviews have a set of criticised limitations, they still remain a valuable source of information in certain application cases. In particular the unstructured interviews are critiqued for their relatively low reliability and poor external validity in comparison to structured interviews. However, they have several advantages in terms of costs and feasibility face to particularly restricted or heterogeneous target population. The work of Chauhan (2022) offers an interesting discussion on the usefulness of the unstructured interviews in some particular cases. Another work presenting the advantages of the unstructured interviews is the publication of Mueller and Segal (2015), with a particular focus on the unstructured interviews. This less rigid approach offers many chances to collect comprehensive information and a relatively detailed account of the individual's experience, rather than solely concentrating on the individual's issues or symptoms in the context of clinical studies. While the majority of research on interviews and selection leans towards structured interviews for their perceived higher validity and reliability, this critical review, guided by relevant authors, aims to highlight the advantages of unstructured interviews. It commences by defining and distinguishing both interview approaches, then proceeds to elucidate the merits and advantages unique to unstructured interviews. These include enhanced face-validity, positive responses from both interviewees and interviewers, comparable validity levels, and increased practicality across various real-world organizational contexts. One of the definitions offered in the Encyclopedia of Quality of Life and Well-Being Research is as follows:

Unstructured interviews (Sanchez 2014) involve a complex interaction between researchers and interview subjects undertaken for the purpose of collecting data pertaining to cognitive processes, social worlds, and experiences. Unlike structured interviews, yet similar to natural conversations, researchers ask questions that are largely unscripted.

In this case, interviewers utilize a conversational approach, adapting to their role within the field setting. While the unstructured interviews do not follow a set of predetermined questions, the more traditional in literature *structured interviews* are characterised by a strict predefined set of questions. A classic example of a structured interview is a *survey*, which usually also imposes strict limitation on the possible answers.

In between the two options exist a number of mixed approaches: *semi-structured interviews*. The latter employ a written guide that directs the conversation toward specific subjects or matters. This topic

guide is prepared in advance, and interviewers generally maintain their focus on the designated topics.

Semi-structured interview(s) (Magaldi and Berler 2018) is an exploratory interview used most often in the social sciences for qualitative research purposes or to gather clinical data. While it generally follows a guide or protocol that is devised prior to the interview and is focused on a core topic to provide a general structure, the semi-structured interview also allows for discovery, with space to follow topical trajectories as the conversation unfolds.

The limitations of the study induced this information collection to be either *unstructured* or *semi-structured*, due to the heterogeneity of the target population. The relatively restricted profile for the subjects as well as time and effort limitations further restrict the methodology. For the purposes of the study we devise a relatively flexible survey guide, which could be potentially adjusted depending on the individual profile, background and experience of the interviewed person.

At this point we are going to present the methodology of interviewees sample composition. The next part will present the survey *as-is*, without any modification or corrections.

D.2.1. Target audience

For the purposes of sample constitution we focus on the researchers and engineers working or having worked with the discrete choice models directly or implementing classification techniques, which could be considered equivalent to the choice modelling. The preference is given to the in person interviews, while virtual interviews are reserved for particular cases.

D.2.2. Survey

Prior to each interview a survey sample in pdf format was sent to each subject. The survey text was equally presented to the subject during the in-person interviews in paper version. For virtual interviews, the pdf file was resent to the participant directly during the interview via the video-conference tool. For this purpose zoom was primarily used, while some subjects preferred to communicate over MS Teams for convenience reasons.

D.2.2.1. Cultural and background differences

- Do you use classification / discrete choice modelling techniques in your research ?
- How did you know about the discrete variable models?
- What discrete (classification) models are familiar to you?
- What (econometric) models can you implement without any particular difficulty?
- Are you familiar with *Discrete Choice Experiments (DCE)* concepts?
 - How did you know about them?
 - Did you undergo some specific training?

D.2.2.2. Research question and problematic definition

- Was the research project initialised on a request from firm / policy maker/ public entity?
 - How precisely was the request formulated?
 - Were the target indicators imposed in the demand?

- Was the project performed in answer to a research project demand / proposition?
- Was the project mounted to obtain a particular sponsorship?
- Was the project performed without any particular dedicated financial aid?
- What was the next research stage once the main research question was identified?

D.2.2.3. Target indicators definition

- Were the target indicators imposed by the research question / problem definition?
- Were the target indicators imposed by the underlying theoretical assumptions?
- Were the target indicators delimited by the known / traditional modelling techniques?

D.2.2.4. Modelling techniques

- Which modelling technique was used ?
- How the modelling technique was selected?
- Was the modelling approach already well known / tested?
- Has the modelling approach impacted the choice of the research topic / question?
- Was the implementation of a particular modelling approach a *selling point* for the research project?
- Was the modelling technique defined by theoretical assumption?

D.2.2.5. Hypothesis / Theoretical assumptions

- What behavioural / decision making theories underlined the modelling?
- The choice of theoretical assumptions was performed:
 - based on personal expertise and experience?
 - after a literature review?
- How the literature review was performed?
- What factors impacted the literature choice?
 - personal experience
 - recommendations of colleagues
 - ... (something else?)

D.2.2.6. Data collection (and experimental design)

- Did the study include a data collection step or an open dataset / dataset from another project was used?
- Was the data shared with other projects or used for other studies?
 - Intentionally (it was collected for other applications). How it has affected the experimental design?
 - Non-intentionally. Did you wish some changes in the experimental design were made?
- Were the *Hypothetical Biases (HB)* taken into account when collecting the data?
 - Which ones?
 - How exactly they were mitigated?
- Was the *statistical efficiency* taken into account during data collection stage?

- Was the data collection separated into several steps?
 - Which ones?
 - What were the reasons for performing such separation?

If working with DCE data :

- What techniques were implemented during *Experimental Design (ED)* engineering?
 - Was any specific software used?
 - What ED was used? (factorial, fractional, efficient, Bayesian, ...)

D.2.2.7. Encountered difficulties and identified limitations

- Which part of the project was the most difficult?
 - Problematic and question definition, indicators identification
 - Theoretical assumptions and literature review (protocol)
 - Experimental design (*if working with DCE*)
 - Data collection
 - Statistical modelling and data analysis
- *What may be wrong with such representation partitioning?*
- What were the difficulties exactly?
- What potential errors and biases were discovered (assumed) after:
 - Theoretical limitations
 - Data collection
 - Model selection and data analysis
 - Results presentation
- Were there any biases / pitfalls / limitations you wish you would know about prior to starting the study?

D.3. Results

The data collection was performed in the period starting with June 2022 to March 2023. The sample of interviewed individuals comprised the members of following research institutions:

1. In France
 - University Grenoble Alpes (UGA)
 - Grenoble INP
 - INRAE
 - CNRS
 - University Paris Cité
2. In Canada
 - University of Montréal
 - Polytechnic of Montréal
 - University Laval
 - University of Quebec

The participants were firstly contacted by mail to determine their will to participate in the interviews.

The diffusion channels differed in order to increase the participation rate. For the researchers in close contact with research supervisors the message bore a more personal character. The more generalised and formal letter was transmitted to all other identified potential participants:

Message example (original in French): Dans le cadre de ma thèse, j’explore les performances des modèles de choix discrets et la classification dans les études des choix individuels. Comme résultat final, nous envisageons une démarche permettant de comparer les procédures d’analyse des données en utilisant ce type de modèles, mais qui pourra être étendue en dehors de ce champ. Nous sommes actuellement en train de réviser les besoins potentiels des utilisateurs potentiels afin d’ajuster la méthodologie proposée. Je cherche par ailleurs des interlocuteurs pour mieux comprendre les procédures d’analyse des données adoptées dans les études appliquées.

Message example (in English): In the context of my thesis, I am exploring the performance of discrete choice models and classification in individual choice studies. As a final outcome, we envision an approach that allows for comparing data analysis procedures using these types of models, but that can be extended beyond this field. We are currently revising the potential needs of potential users to refine the proposed methodology. Furthermore, I am seeking contacts to better understand the data analysis procedures adopted in applied studies.

The final sample comprises 23 individuals: 6 from France and 17 from Canada (QC). As expected the final sample is relatively small, due to the limitations imposed by the specificity of target audience and the relative complexity of the study. Because of relatively small sample associated with relatively large number of research entities of affiliation the descriptive statistics incorporating information on the subjects’ occupation is not provided. It suffices to mention that most of the subjects had a grade of professor by the time they were interviewed.

The analysis of unstructured interviews is a nuanced process that involves delving into rich qualitative data to extract meaningful insights. Due to the unstructured nature of the collected data, as well as the limited sample size and relatively high heterogeneity in the participants’ profiles the analysis part was rather complex. The analysis step could not be performed with desired precision and usage of advanced statistical methods.

The initial step involved transcribing and familiarizing oneself with the entire dataset to identify recurring themes or patterns. The transcription of the results was performed on individual basis with a separate file for each participant. All the answers were stored in markdown format in order to simplify the access, no personally identifiable information was stored in the process. To this day the analysis remains incomplete, due to the complexity of the collected data. For the purposes of framework structure identification the confirmatory analysis was performed focusing on the general topics addressed by the subjects.

D.4. Conclusion

In conclusion, this appendix provides a comprehensive exploration of research procedures within the context of discrete choice analysis, particularly through the lens of choice experiments. The study employs semi-structured interviews to gather insights into the diverse habits and customs of data analysis among scientists, with a specific emphasis on applied studies and real-world problematic.

The motivation, methodology, and data collection processes are detailed, offering a transparent view

of the research procedures under investigation. Understanding the variations in research practices is crucial in the realm of interdisciplinary studies, where conventions, preferences, and individual decisions play integral roles. This work also sheds light on the scarcity of literature addressing research procedures, emphasizing the need for more comprehensive insights, that will bridge the applied studies with the fully theoretical counterpart.

The analysis of unstructured interviews is acknowledged as a nuanced and ongoing process, impacted by the unstructured nature of the data, limited sample size, and participant heterogeneity. Despite challenges, this study contributes valuable insights to the proposed performance comparison framework and highlights the intricacies of research practices within the discrete choice analysis domain.

E. Software packages

This appendix presents two ‘R’ packages developed for discrete choice modeling and performance evaluation as a part of the PhD work. The first package, ‘dcesimulatr’, offers a flexible and controlled environment for simulating Discrete Choice Experiments. While currently supporting a minimal set of behavioral theories and experimental design configurations, the package aims to expand its functionalities in the future. The package was inspired by leading packages like ‘biogeme’ (Python) and ‘apollo’ (R) and is available on GitHub. The second package, ‘performancer’, is a collection of functions designed for assessing the performance of Discrete Choice Models. It enables users to calculate various performance metrics commonly used in classification and discrete choice analysis tasks. It consolidates a diverse set of metrics explored during the PhD project, providing a comprehensive solution for efficient and thorough performance evaluation in discrete choice modeling.

E.1. ‘dcesimulat_r’: A DCE simulation toolset

The package `dcesimulatr` provides a flexible controlled environment for discrete choice experiment simulation. At this moment, the package supports only minimal number of behavioural theories and preset experimental design configuration. We hope that in near future, the number of available functionalities will increase.

The ‘`dcesimulatr`’ package¹ is a tool for simulating *Discrete Choice Experiments (DCE)* written in R. It provides a flexible and controlled environment for conducting simulations of discrete choice experiments, a widely used methodology in economics and marketing. In its logic the package follows the leading packages in the field, among which the `biogeme` (Python) and `apollo` (R).

The functionality is built up around the *Agent Based Simulation (ABS)* paradigm, allowing the user to define a series of individuals with particular behavioural rules set and combine them into population. The resulting population is used to compute the associated choice probabilities for the corresponding choice sets of alternatives. The choice sets might be supplied either externally, as a separate dataframe, or declared internally. The internal choice set declaration only support the most widespread experimental design layouts for now.

Overall, ‘`dcesimulatr`’ facilitates the simulation of discrete choice experiments, offering a range of functionalities for designing experiments, generating populations, and simulating decision processes within a controlled environment.

E.2. ‘performancer’: Performance estimation functions collection

The second package created during PhD completion is tool designed for assessing the performance of *Discrete Choice Models (DCM)*. The package ‘`performancer`’ provides a number of function to easily calculate many of the performance metrics used in classification and discrete choice analysis tasks. The main theoretical reference for this implementation is the handbook “*Evaluating Learning Algorithms : A Classification Perspective*” by Japkowicz and Shah (2011) ([Cambridge University Press](#)).

The package offers a wide range of functions that facilitate the evaluation of model predictions and classification results using various metrics. The focus of these functions is on quantifying the accuracy and reliability of predictions in the context of binary and categorical classification problems.

Key functionalities include computing widely used performance metrics, among which: accuracy, binary and categorical crossentropy, class ratios and diverse confusion matrix based metrics. As well, as some less common metrics including: Cohen’s Kappa, *Kullback–Leibler Divergence (KLD)*, Scott’s Π and *S*-coefficient. Finally, the package also offers functionality for precision-recall (PR) coordinates computation. The later might be used for *Receiver Operating Characteristic (ROC)* curves construction for sensitivity analysis and algorithms’ performances evaluation.

Because most of those metrics might be encountered within other software solutions, this is no more than a convenience software, regrouping most of the metrics explored during the PhD.

¹‘`dcesimulatr`’, version 0.1.1, is available on GitHub.

F. Synthesis in French

Abstract

Cette thèse est une étude interdisciplinaire de la modélisation discrète des choix individuels économiques, abordant à la fois les techniques d'économétrie et d'apprentissage automatique (ML) appliquées à la modélisation de choix individuelle. La problématique découle de points de contact insuffisants entre les utilisateurs (économistes et ingénieurs) et les analystes des données, qui poursuivent différents objectifs, bien qu'utilisant des techniques similaires. Pour combler cet écart interdisciplinaire, ce travail propose un framework unifié pour l'analyse des performances du modèle. Il facilite la comparaison des techniques d'analyse des données sous différentes hypothèses et transformations. Le framework conçu convient à une variété de modèles économétriques et ML. Il aborde la tâche de comparaison des performances du point de vue de la procédure de recherche, incorporant toutes les étapes affectant potentiellement les perceptions des performances. Pour démontrer les capacités du framework, nous proposons une série de 3 études appliquées. Dans ces études, la performance du modèle est explorée face aux changements de: (1) la taille et l'équilibre de l'échantillon, résultant de la collecte de données; (2) les changements de la structure des préférences au sein de la population, reflétant des hypothèses comportementales incorrectes; et (3) la sélection du modèle, directement liée à la perception des performances.

Introduction

Avec le développement des dispositifs informatiques nous pouvons constater l'augmentation de la disponibilité des données, ainsi que de nouvelles méthodes d'analyse. En particulier, les avancées en apprentissage statistique (Hastie, Tibshirani, and Friedman 2009) et en science des données (Donoho 2017) des dernières décennies ont entraîné la propagation des techniques d'apprentissage automatique (*Machine Learning* ou *ML*) dans l'application à la résolution de problèmes économiques. Les modèles les plus gourmands en ressources des décennies précédentes peuvent être exécutés en quelques minutes, et la recherche actuelle se concentre de plus en plus sur le big data et l'automatisation de l'analyse. En économie une attention particulière est accordée à des tâches d'évaluation des politiques ou pour la modélisation du comportement des agents économiques. Malheureusement, le nombre de stratégies d'analyse de données disponibles peut rendre difficile la sélection de la solution optimale pour les non-experts (Athey and Imbens 2019).

Par exemple, même le cas en le plus banal de l'analyse des données de choix du consommateur binaire peut être abordé avec des outils très variés. Ces méthodes varient des tests de différences d'échantillons de base à l'analyse de régression plus sophistiquée, pour finalement arriver aux classificateurs supervisés complexes avec la mise en œuvre d'algorithmes de renforcement. Chacune des options énumérées a ses avantages et ses faiblesses, et un utilisateur inexpérimenté peut facilement négliger certains de ces éléments (Mullainathan and Spiess 2017). Pour résoudre ce problème, il est important de mieux comprendre les points forts et les points faibles des différents modèles.

Cependant, aborder le problème général de la comparaison des performances des modèles sans contexte particulier serait extrêmement difficile. Dans les disciplines économiques, il existe de nombreux scénarios d'application et de cas d'utilisation, chacun ayant des exigences très spécifiques en termes de sélection d'outils. Un tel travail fondamental nécessiterait une connaissance approfondie à la fois des modèles et des spécificités d'application économique, car l'utilisation d'un modèle ne peut rarement être analysée sans aucun contexte d'application. De plus, chaque année, le nombre de modèles disponibles augmente à mesure que des outils de plus en plus complexes traitant des cas d'utilisation spécifiques émergent, ce qui place la création d'une taxonomie unifiée de tous les modèles disponibles en dehors du cadre de toute étude limitée dans le temps.

Afin de limiter la portée de notre étude, nous concentrerons notre attention sur la famille de modèles de choix discrets dans le contexte des études de choix individuel. Cette limitation établira une base pour la discussion. La modélisation du choix se concentre sur l'exploration de l'analyse du comportement, qu'il s'agisse d'un individu ou d'un autre type de décideur. Elle encadre un nombre relativement limité de techniques, par rapport à toutes les autres familles de modèles disponibles. Celles-ci peuvent être résumées comme les méthodes de classification utilisant la terminologie d'apprentissage statistique (*Statistical Learning* ou *SL*).

À ce stade, il est important de souligner les problèmes clés et les difficultés associés à la tâche de comparaison des performances des modèles interdisciplinaires. Ces problèmes peuvent être séparés en deux groupes principaux : (1) les complexités techniques de la mise en œuvre et de l'utilisation de l'ensemble d'outils disponible ; et (2) les différences conceptuelles imposées par l'hétérogénéité entre les cas d'application, ainsi que les profils divers des utilisateurs.

Tout d'abord, nous partons du principe que les descriptions disponibles de l'ensemble d'outils peuvent

sembler extrêmement complexes pour les utilisateurs non-experts du domaine. En d'autres termes, nous supposons que chaque modèle en dehors du niveau d'expertise de premier cycle ou de cycles supérieurs pourrait nécessiter des efforts d'apprentissage de la part du public cible. Pour justifier cette hypothèse, examinons la présentation de l'un des modèles de base largement utilisé pour l'analyse du choix de nos jours, le *Modèle Logit Multinomial* (MNL) soutenu par le cadre de *Maximisation de l'Utilité Aléatoire* (RUM). Alors que presque tous les manuels disponibles présentent cet outil de manière guidée et accessible (Agresti 2013), le travail original introduisant cet ensemble d'outils (McFadden 1974) est beaucoup plus complexe pour les lecteurs inexpérimentés. Les outils modernes nécessitent une connaissance experte avancée pour être utilisés, et la littérature à jour est principalement orientée vers des chercheurs expérimentés.

Effectivement, il est possible de trouver des notes techniques ou des guides qui tentent de combler le fossé existant entre la littérature scientifique la plus récente et la littérature éducative de base, formant une couche de sources de connaissances *avancées*. Cependant, de tels supports *avancés* fournissent rarement suffisamment d'informations au lecteur et sont généralement biaisés ou incomplets. Une illustration très intéressante de cela peut être tirée des tentatives de mettre en œuvre la méthodologie d'apprentissage automatique dans les études économiques. Par exemple, dans les travaux de Athey and Imbens (2019) ou de Mullainathan and Spiess (2017), nous rencontrons des *directives* pour les économistes sur l'utilisation de l'ensemble d'outils d'apprentissage automatique. Cependant, bien que dans les deux cas les publications fournissent des discussions intéressantes sur l'utilité de l'ensemble d'outils d'apprentissage automatique pour les économistes, les deux passent à côté de la complexité de la courbe d'apprentissage pour les économistes non familiers de ces techniques avancées.

Deuxièmement, nous pouvons observer une ambiguïté extrême et des incohérences dans le vocabulaire variant selon les communautés. Les domaines et les branches différentes de la science, bien qu'utilisant des outils assez similaires, peuvent avoir une compréhension différente des implications théoriques qui les sous-tendent. L'exemple le plus basique dans ce cas serait la distinction entre les tâches de *classification* et d'*analyse de choix discrète* (DCA). Bien que l'ensemble d'outils mis en œuvre pour ces tâches soit généralement très similaires (Agresti 2007; Hastie, Tibshirani, and Friedman 2009), les différences conceptuelles rendent relativement difficile la fusion des connaissances disponibles sur un support commun. Bien que dans les deux cas, les manuels introduisent des concepts relativement similaires, parmi lesquels les régressions logistiques binomiales et multinomiales, la présentation varie considérablement. L'introduction d'autres applications potentielles pour un ensemble d'outils apparemment identique, comme les études des préférences (Förnkrantz and Hüllermeier 2010) ou l'analyse des choix en économie de la santé (Soekhai et al. 2019), ne fait qu'augmenter le nombre de terminologies divergentes.

De plus, non seulement le côté pratique diffère, mais les termes les plus basiques peuvent être compris selon des perspectives différentes. L'une des illustrations les plus parlantes à cet égard est le terme ambigu de *modèle*, qui peut être compris différemment en fonction du contexte. Les modèles théoriques, statistiques, mathématiques, économétriques et économiques apparaissent dans la littérature, et tous peuvent être désignés simplement comme *un modèle* compte tenu de la spécificité du travail. Par exemple, le travail de Sfeir, Rodrigues, and Abou-Zeid (2022) a un terme *modèle* présent directement dans le titre de la publication : "*Gaussian process latent class choice models*" - se référant à la famille des modèles statistiques des choix. Il en va de même pour le travail de El-Badawy, Elharoun, and Shahdah (2021) : "*Captivity impact on modelling mode choice behaviour*". Cependant, dans ce deuxième cas, la

délimitation du terme *modèle* est plus ambiguë, car il n'est pas clair à partir du titre s'il concerne les modèles théoriques de comportement de choix ou l'aspect statistique de la question. Un exemple plus complexe peut être tiré du travail de Lee, Derrible, and Pereira (2018), où différentes configurations de *réseaux de neurones* (NN) sont comparées avec le *modèle* Logit Multinomial (MNL). Alors que le *modèle* MNL est relativement bien délimité dans la littérature, la partie des NN est moins claire dans sa définition en raison d'une structure modulaire particulièrement complexe.

Avec ce travail, nous tentons d'organiser les connaissances existantes de différentes disciplines dans une étude interdisciplinaire. **Le Chapitre 1** de du manuscrit est consacré à la définition de la portée de l'étude. Il passe en revue les divers malentendus qui surviennent lorsque différentes disciplines sont réunies sur un même sujet. Les définitions du vocabulaire utilisé par la suite sont présentées. Cela comprend également un aperçu historique avec un accent sur le développement des outils et de la théorie du choix. **Le Chapitre 2** déplace le focus sur la tâche de la *comparaison de performances*. Nous abordons un par un les principaux éléments de la procédure scientifique qui peuvent potentiellement influencer la performance observée du modèle ainsi que la perception ultérieure de celle-ci. De la question de recherche aux métriques cibles en passant par la collecte de données et la sélection de modèles, nous passons en revue chacune des étapes qui peuvent potentiellement influencer les performances perçues. Cette revue nous amènera à la définition d'un nouveau *framework*¹ de comparaison des performances. Enfin, dans **Le Chapitre 3**, nous proposons une sélection d'études de cas, accompagnées par des réflexions sur leur mise en œuvre par rapport au framework proposé. Il s'agit principalement de communications scientifiques faisant usage du framework de comparaison des performances pour aborder et explorer différentes problématiques en modélisation du choix.

En mettant particulièrement l'accent sur la modélisation du choix individuel, ce travail contribue à la littérature économique, et en particulier à la littérature sur l'économie expérimentale portant sur l'analyse des données d'expérience de choix discrets. Le framework proposé pour l'évaluation et l'analyse des performances des modèles peut être étendu au-delà de la portée directe de cette étude par le biais d'une série de généralisations. Théoriquement, la solution proposée peut être mise en œuvre de manière équivalente dans différents scénarios d'application, notamment : (1) l'économie de la santé avec une utilisation intensive de la modélisation du choix et des expériences de choix discrets ; (2) le marketing avec un accent sur l'optimisation et l'analyse des préférences ; (3) l'économie de l'innovation, en se concentrant sur les préférences individuelles pour les biens et services innovants ; et (4) l'analyse de la prise de décision stratégique dans le contexte de l'économie industrielle, car la modélisation du choix peut être étendue à d'autres sujets.

F.1. Modélisation du choix : À la croisée des disciplines

L'*Analyse de Choix Discrets* (DCA) est un ensemble de techniques quantitatives de recherche utilisées pour analyser et prédire le comportement individuel dans des tâches basées sur le choix (K. Train 2002). Elle est largement répandue dans de nombreux domaines de recherche tels que l'économie (Durlauf and Blume 2010; Athey and Luca 2019), la santé (Mühlbacher and Bethge 2015), le marketing (Coussement, Benoit, and Poel 2010), les transports (Guevara and Ben-Akiva 2013), et les sciences environnementales (Daziano and Achtnicht 2014). Indépendamment du contexte, elle est principale-

¹Bien que en français il existe le terme *cadre*, nous préférons garder l'anglicisme *framework* dans cette synthèse, afin de garder la cohérence dans la terminologie.

ment utilisée pour comprendre les préférences individuelles et les processus de prise de décision, que ce soit le choix d'un mode de transport ou les préférences pour des attributs particuliers parmi les produits disponibles. La modélisation du choix implique généralement la conception d'enquêtes ou d'expériences où les répondants font des choix entre différentes options, et les données collectées sont utilisées pour estimer des modèles qui révèlent les préférences sous-jacentes et les compromis que les individus considèrent lors de la prise de décisions (Ben-Akiva, McFadden, and Train 2019). Cette approche fournit des informations précieuses pour les entreprises (Bode, Macdonald, and Merath 2022), les décideurs politiques (Mihailova et al. 2022), et les chercheurs (Fifer, Rose, and Greaves 2014). Elle les aide à prendre des décisions éclairées, à développer des stratégies efficaces et à comprendre les motivations derrière les choix individuels.

Récemment, la DCA traditionnelle a commencé à adopter certaines des techniques de modélisation complexes du ML (Hillel et al. 2021; Aboutaleb et al. 2021). Cette convergence de méthodologies a enrichi l'ensemble d'outils de la DCA en améliorant ses capacités prédictives et en élargissant son applicabilité (Danaf et al. 2019). Bien que cette fusion de disciplines représente une avenue prometteuse dans les environnements de décision complexes et riches en données d'aujourd'hui, certaines complications se posent. La variété croissante de stratégies d'analyse de données peut poser un défi significatif pour les chercheurs sans expertise dans le domaine, rendant de plus en plus difficile le choix de la solution la plus adaptée. Selon les résultats des entretiens réalisés avec des chercheurs en exercice menés au cours de ce travail de thèse, deux principales stratégies dans le choix des modèles sont soulignées : (1) les chercheurs appliquent les modèles avec lesquels il est intéressant de travailler pour eux ; ou (2) ils utilisent les modèles avec lesquels ils sont suffisamment familiers pour accomplir la tâche donnée. Alors que ce raisonnement s'applique aux chercheurs expérimentés, les novices peuvent être encore plus limités dans le choix de la stratégie de modélisation. Cela les conduit généralement à suivre les stratégies de modélisation les plus courantes, potentiellement sans une compréhension complète des processus sous-jacents. Ce choix est rendu encore plus difficile par la diversité de la littérature scientifique moderne sur la modélisation du choix et les techniques de classification.

La diversité des origines des techniques de modélisation disponibles et les variations de terminologie compliquent davantage le processus de sélection de la bonne option pour les utilisateurs occasionnels. Par exemple, en fonction de la familiarité avec l'un ou l'autre domaine, le scientifique recherchera soit des techniques de *classification*, soit des techniques de *modélisation du choix*. Pour résoudre ce problème, il est nécessaire de mieux comprendre les forces et les faiblesses des différents modèles face à différentes questions économiques. Au cœur de cette tâche se trouve la capacité à comparer et à différencier les approches de modélisation disponibles. En effet, si la comparaison de modèles de choix apparemment similaires est relativement facile en raison de leur structure similaire, la comparaison avec des méthodes complètement différentes est beaucoup plus difficile.

L'évaluation des performances est généralement réalisée dans des articles académiques proposant de nouveaux modèles ou des techniques d'estimation alternatives. Cela lie étroitement le concept de performance au modèle lui-même. Les études appliquées adoptent généralement une approche plus prudente lors de la présentation de la procédure et des résultats de l'évaluation des performances. En général, seul le modèle le plus prometteur arrive à la publication ou à la production. Cependant, certaines œuvres méthodologiques, principalement dans les études orientées économétriques, s'éloignent de ce paradigme et explorent l'élucidation des effets individuels (M. Bierlaire, Bolduc, and McFadden 2008) ou la capacité à dériver correctement certaines métriques composites (Rose and Bliemer 2013).

Cela souligne la disparité présente dans la littérature. De plus, le concept de performance est loin d'être le seul terme ambigu dans la littérature. Les définitions de concepts aussi simples que *modèle*, *ML* ou *procédure scientifique* peuvent être comprises différemment en fonction de la formation du lecteur dans le contexte d'un travail interdisciplinaire. Par conséquent, consacrer l'une des sections introductives à la spécification de la terminologie est essentiel.

Cependant, une telle tâche ne peut être accomplie sans aucune connaissance préalable du domaine d'application. Le domaine d'application ainsi que la littérature associée définiront les principes de base de la spécification de la terminologie. Dans le même temps, les domaines d'intérêt associés influenceront nos définitions, les façonnant et les ajustant. L'état actuel de la littérature existante souligne l'impératif de systématiser minutieusement la terminologie qui sera employée tout au long de ce manuscrit.

Le Chapitre 1 est dédié à l'introduction de concepts fondamentaux et de la terminologie qui sont utilisés dans les sections suivantes de ce manuscrit. En commençant par une introduction de base à la discipline du DCA, ce chapitre établit une base pour une discussion ultérieure sur la comparaison des performances. Les différences des objectifs entre les domaines d'application (économie, gestion, sociologie), ainsi que les différents paradigmes épistémologiques (apprentissage automatique et économétrie), sont soulignés. Les problèmes de construction de la taxonomie des modèles dans le contexte d'un travail interdisciplinaire seront également abordés. Enfin, la complexité de la définition du concept de modèle sera introduite, passant à la présentation des complexités de la tâche de comparaison des performances.

F.2. Un framework universel de comparaison des performances

Dans le Chapitre 1.1, le lecteur est introduit aux concepts fondamentaux et aux défis abordés dans ce travail. Cela a permis de démontrer le point de convergence pour le discours interdisciplinaire, comblant le fossé entre les études axées sur la théorie et celles axées sur les données. Malgré diverses tentatives de concilier ces approches, un consensus sur des stratégies efficaces d'atténuation des écarts reste à atteindre.

Répondre au défi posé par l'incohérence de la base commune à travers divers domaines d'application et les stratégies d'acquisition de connaissances nécessite l'introduction d'une approche accessible à l'utilisateur pour l'exploration des performances des modèles. La littérature actuelle n'offre pas de méthodologie unifiée pour l'évaluation des performances des modèles, expliquée par la compréhension hétérogène des concepts de performance dans différents domaines d'application, comme décrit dans la section 1.4.2. Un nouveau *framework* est conçu pour combler les lacunes et développer une compréhension plus unifiée des performances des modèles à travers différentes disciplines et méthodologies de recherche. Le framework vise à fournir un ensemble d'outils flexible pour évaluer et comparer l'efficacité de différentes techniques de modélisation et leur applicabilité dans des contextes divers. Cette approche tente d'harmoniser les perspectives des études axées sur la théorie et celles axées sur les données, facilitant le dialogue au sein de la communauté interdisciplinaire.

L'analyse des performances des modèles et l'analyse des performances en général servent de base à ce chapitre. L'analyse complète est essentielle car elle englobe l'essence même des objectifs et des méthodologies sous-tendant les tâches d'analyse de données et de modélisation. Il existe plusieurs

études abordant ces questions de manière séparée, mais rarement la discussion tient compte de toutes les dimensions disponibles du problème. Cependant, l'analyse des performances des modèles et leur évaluation ne peuvent pas être adéquatement capturées de manière isolée, l'intégrité de l'étape d'analyse des données et même les objectifs de la tâche scientifique influent sur les perceptions de la performance.

Ainsi, le framework proposé pour l'analyse et la comparaison des performances se base sur la procédure scientifique standard, avec une flexibilité suffisante pour l'étendre à d'autres domaines et disciplines. La procédure adoptée peut être considérée comme assez proche de nombreux articles économiques appliqués et théoriques. Malheureusement, aucun travail que nous connaissions, à l'exception du manuscrit de Williams and Ortuzar (1982), n'aborde le flux de travail scientifique du même point de vue que nous. Alors que de nombreux articles de recherche mettent en œuvre des idées similaires au framework proposé, nous n'avons pas connaissance de travaux qui mette particulièrement l'accent sur la partie systématique des procédures.

En particulier, dans leur travail, Williams and Ortuzar (1982) abordent les *théories comportementales de la dispersion et la mauvaise spécification des modèles de demande de déplacement*. En introduisant un framework de comparaison des performances axé sur les implications politiques, ils illustrent comment la mauvaise spécification dans la génération de l'ensemble de choix peut biaiser les paramètres du modèle dans les modèles de choix de mode. Les idées présentées dans leur travail ont un impact durable sur le paysage de l'analyse du DCA. Gonzalez-Valdes, Heydecker, and Ortúzar (2022) fait référence à ce travail fondamental dans le contexte de la simulation de jeux de données à des fins d'évaluation des performances du modèle ICLC. Vij and Walker (2016) explore les cas où les modèles ICLV sont utiles, illustrant leur point de vue par la simulation. Bahamonde-Birke and Ortúzar (2014) utilise la procédure de simulation proposée pour explorer et tester les capacités des modèles HCM. Cependant, la plupart des travaux se concentrent uniquement sur le concept de l'utilisation de la simulation pour l'analyse des performances, ignorant parfois le concept d'évaluation des performances dans le contexte des implications politiques publiques.

Dans le Chapitre 1.1, nous présentons l'approche proposée en détail, élément par élément, en abordant des dimensions telles que : (1) l'analyse et la comparaison des performances, (2) la sélection des modèles et (3) la gestion des données. Les concepts clés pour comprendre le contenu de ce chapitre ont été présentés précédemment, et en cas de malentendu, il est conseillé de consulter le Chapitre 1. Tous les éléments susmentionnés sont ensuite réunis dans un framework de comparaison des performances (Figure F.1). Le chapitre se termine par une description des cas d'utilisation du framework sur la base de littérature existante.

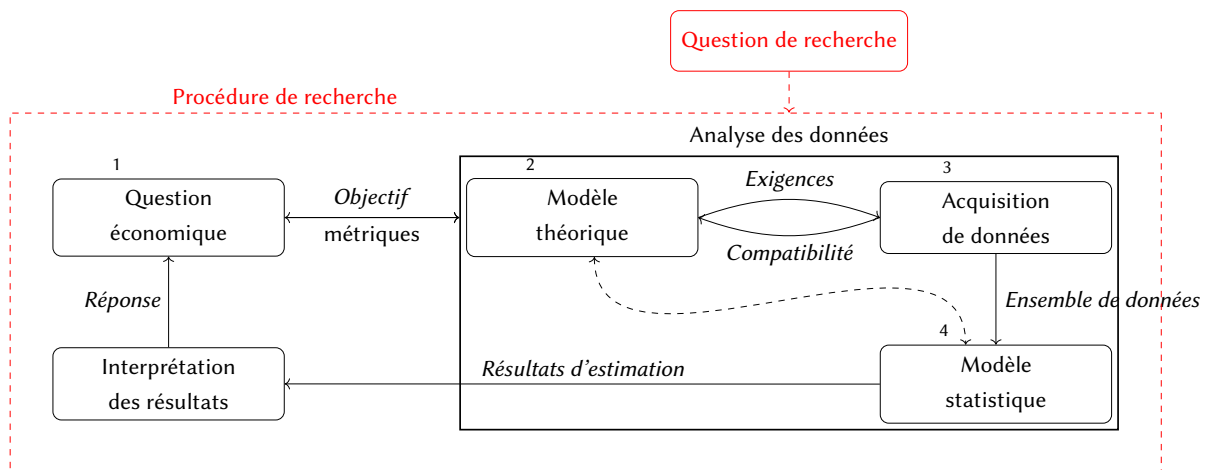


Figure F.1.: Cadre proposé pour la comparaison des performances

F.3. Le framework en action : Études de cas

Dans le chapitre précédent, le lecteur a été introduit au framework de comparaison des performances, apportant une nouveauté et une cohérence en unifiant les approches existantes des études méthodologiques et appliquées basées sur l'analyse du choix. Plusieurs exemples ont été proposés sur la manière dont les études existantes se rapportent à la structure du framework. Dans ce chapitre, une série d'études de cas sont présentées, offrant une compréhension plus approfondie des fonctions et des cas d'utilisation du framework.

Chacune des études de cas se concentre sur un élément différent du framework : (1) les relations avec les hypothèses théoriques et les métriques cibles à la modélisation, (2) les problèmes liés à l'acquisition de données, en particulier la configuration de l'ensemble de données ; et (3) la modélisation statistique. Évidemment, le framework intègre beaucoup plus d'étapes et de phases, dans le temps limité de la thèse, nous avons eu l'opportunité de nous concentrer uniquement sur quelques uns des éléments clés. Tous ces éléments font partie de l'étape de la procédure d'analyse des données, complétant la présentation des cas d'utilisation du framework.

La première étude, disponible dans la section 3.2, combine les modèles économétriques et ML pour la modélisation des préférences de choix des consommateurs, abordant les défis interdisciplinaires. C'est la première production de cette thèse à introduire le framework de simulation et de test théorique. L'adaptabilité du framework aux indicateurs économiques et statistiques est illustrée en faisant référence au travail de Michaud, Llerena, and Joly (2012), les données issues de l'expérience faite en 2012 sont réutilisées dans cette étude de cas. Trois modèles d'économétrie et de ML sont estimés et comparés sur deux ensembles de données synthétiques avec des fonctions d'utilité prédéfinies, simulant des préférences homogènes et hétérogènes.

Bien qu'il s'agisse de la phase initiale d'un projet de thèse plus vaste, cet article met en lumière des idées concernant les différences entre les disciplines tout en restant optimiste quant à la simplicité de la comparaison interdisciplinaire des performances des modèles. La méthodologie implique la génération de jeux de données artificielles avec des structures de préférences hétérogènes et homogènes, explorant les effets de l'hétérogénéité des goûts sur les performances des modèles. Les résultats soulignent que le choix du modèle dépend des hypothèses du chercheur concernant le processus sous-jacent de

génération de données, mettant en avant le compromis entre la complexité du modèle et la précision de l'estimation. Notamment, dans le contexte de la sélection de modèle en présence d'incertitudes concernant les hypothèses théoriques, l'article suggère une préférence pour des modèles plus complexes afin d'atténuer les biais potentiels introduits par une spécification incorrecte du modèle, offrant ainsi un aperçu nuancé du paysage complexe de la sélection de modèle. L'étude reconnaît ses limites, soulignant la nécessité de recherches supplémentaires et d'extensions pour améliorer la généralisabilité et l'utilité pratique du framework.

La deuxième étude, présentée dans la section 3.3, se concentre sur la tâche d'élicitation des consentements à payer (WTP), une métrique largement utilisée pour évaluer les préférences individuelles pour les attributs dans les choix économiques. L'étude étend le framework de comparaison des performances pour organiser les recherches antérieures et évaluer systématiquement les performances des modèles dans la tâche d'élicitation des WTP, en tenant compte des mauvaises spécifications potentielles, des changements de taille d'échantillon et de l'équilibre de l'ensemble de données. Un ensemble de données synthétique est utilisé pour une application pratique, avec des simulations modifiant la taille et la configuration de l'échantillon pour l'estimation du modèle et l'élicitation de la WTP. Les résultats montrent la variabilité des estimations de la WTP selon les configurations.

L'étude poursuit l'exploration des techniques traditionnelles de DCM et les techniques interprétables de ML, reconnaissant la complexité découlant de la multitude de modèles et de techniques d'estimation, posant un défi aux chercheurs pour choisir l'approche la plus adaptée. L'évolution du framework est particulièrement notable, mettant en évidence de légères modifications résultant d'une compréhension affinée des relations entre les éléments. Cependant, la recherche fait face à certaines limites et à des résultats inattendus. Notamment, la part des effets correctement estimés n'augmente pas de manière cohérente avec la taille de l'échantillon, remettant en question les résultats anticipés. De plus, des résultats intrigants émergent concernant l'impact de l'équilibre des classes sur les estimations de la WTP, surtout dans les cas où des ensembles de données déséquilibrés produisent de meilleurs résultats pour certains modèles. Ces divergences nécessitent une exploration approfondie, soulignant les complexités continues dans le domaine interdisciplinaire.

La troisième étude, figurant dans la section 3.4, explore la comparaison entre les modèles économétriques et ML dans le contexte de la modélisation du choix du mode de déplacement domicile-travail. Elle évalue les modèles traditionnels de choix discret par rapport aux approches ML émergentes dans le contexte de l'élicitation des indicateurs économiques, en particulier la WTP. L'étude utilise l'ensemble de données bien connu *swissmetro* et génère des échantillons synthétiques. Elle confronte ensuite les modèles traditionnels de choix discret (MNL et NL) avec des alternatives ML émergentes, dont ASUDNN (S. Wang, Wang, and Zhao 2020), dans la tâche d'estimation de la WTP. En tant que *working paper*, cette étude de cas ne fournit pas de nouvelles preuves sur le sujet, mais sert à illustrer les problèmes associés à l'analyse et à la comparaison des performances des modèles à des fins de sélection de modèle.

Dans son état actuel, l'article offre un aperçu du potentiel du framework dans le domaine des études économiques appliquées et théoriques, explorant divers modèles dans des questions de recherche liées aux transports. De plus, malgré sa nature incomplète, l'étude joue un rôle particulier au sein de cette thèse. Elle sert à illustrer les limitations et les erreurs dans la tâche d'analyse des performances. Les résultats révèlent des considérations cruciales telles que le rôle du choix des hyperparamètres et de l'échelle dans les modèles de NNs, soulignant la nécessité d'un réglage méticuleux et d'un prétraite-

ment des données dans les expériences futures. L'examen des estimations des effets directs met en lumière des biais et des anomalies entre les modèles, indiquant des défis potentiels dans l'application d'algorithmes ML standard aux tâches d'estimation d'effets sur de petits échantillons. Les temps d'estimation pour les techniques liées aux NNs sont paradoxalement élevés, s'écartant des attentes initiales et soulignant l'importance de l'optimisation, des considérations matérielles et des choix logiciels dans la sélection du modèle. L'étude contribue finalement à des perspectives précieuses sur le paysage nuancé des performances des modèles, mettant en garde contre l'application hâtive de techniques avancées à des problèmes en apparence simples. Le travail souligne également l'importance de prendre en compte les caractéristiques spécifiques des données, le réglage des hyperparamètres et l'efficacité computationnelle dans les études empiriques.

Étant donné que tous les articles individuels ont été produits à différentes étapes de maturité de la thèse, il peut y avoir des incohérences dans la vision et la présentation du framework. Une évolution du framework peut être retracée sur ces travaux, car leur ordre correspond à l'ordre chronologique de leur production. Nous préservons également le format original de ces articles, y compris le libellé, les définitions et l'orthographe. Chaque section de ce chapitre commence par un résumé de l'article respectif, suivi directement du contenu. À la fin de chaque travail, nous offrons une courte discussion abordant les éventuels enseignements tirés du travail. Les différences entre les éléments figurant dans l'article et la version finale du framework sont également mises en évidence. Une telle structure devrait aider le lecteur à comprendre les limites du travail, ainsi que sa position dans le contexte de la version finale du framework.

F.4. Conclusion

Les progrès dans l'efficacité informatique et l'augmentation de la disponibilité des données ont popularisé des méthodes d'analyse de données autrefois gourmandes en ressources. Ces changements ont particulièrement impacté les disciplines dépendant de l'apprentissage statistique, parmi lesquelles les études empiriques en économie. Des modèles autrefois exigeants en ressources de calcul peuvent désormais être estimés en une fraction de temps, incitant la recherche actuelle à se concentrer davantage sur le big data et l'automatisation. Bien qu'il existe de nombreuses études visant à combler le fossé interdisciplinaire et à promouvoir des approches innovantes de l'analyse de données, les études appliquées restent souvent contraintes par des techniques de modélisation plus accessibles. La disponibilité de nouvelles approches en analyse de données ne réduit pas la charge associée à la sélection de modèles et conduit à une surcharge de choix pour les utilisateurs non-expérimentés.

Cette étude aborde le problème de l'évaluation des performances des modèles dans le contexte de la modélisation des choix des consommateurs. La modélisation des choix discrets reste une tâche plutôt complexe, et la disponibilité accrue d'outils avancés en ML ne fait qu'augmenter la flexibilité dans le choix des techniques de modélisation. La pléthore de stratégies d'analyse de données disponibles peut être écrasante pour les non-experts cherchant à choisir la solution optimale. Chaque option a ses propres forces et faiblesses, facilitant l'omission d'éléments clés par les utilisateurs inexpérimentés, d'où la nécessité croissante d'une meilleure compréhension des forces et des faiblesses de divers modèles.

La première partie de ce travail, en particulier le Chapitre 1.1, offre un aperçu de l'état actuel des pratiques dans les applications de modélisation des choix en économie, en tenant compte du contexte inter-

disciplinaire. Cela met en évidence les complexités et les divergences qui existent entre les différentes disciplines et contextes d'application. Des applications économiques de base à la recherche sur les transports et les études de préférences, les techniques de modélisation des choix sont largement répandues, entraînant d'importantes différences de pratiques, dictées directement par les cas d'utilisation sous-jacents. Ces différences vont au-delà des conventions et des pratiques établies, touchant également le vocabulaire et la terminologie utilisés par les chercheurs.

L'environnement de recherche rapide ne facilite pas la recherche d'un terrain d'entente entre les disciplines. Chaque année voit émerger de nouvelles approches novatrices d'analyse des données, de leurs combinaisons et de leurs transformations. Cela rend la tâche initialement sous-entendue de la taxonomie des approches de modélisation presque impossible, en raison des limitations naturelles des capacités cognitives d'un seul chercheur. La surveillance constante de la littérature dans plusieurs domaines reste une tâche extrêmement difficile, obligeant à chercher des solutions alternatives.

Dans ce travail, l'une des réponses naturelles au problème est proposée sous la forme d'un framework d'évaluation et de comparaison des performances, dont la construction est détaillée dans le Chapitre 2. Bien qu'il soit presque impossible de réunir toutes les informations croissantes sur différentes techniques de modélisation, il est toujours possible de fournir un ensemble d'outils pour la comparaison et la sélection de modèles à la communauté des utilisateurs appliqués. Bien qu'il ne fournisse pas de réponse au problème d'acquisition de connaissances, il offre un ensemble d'outils pour la sélection et le réglage fin des techniques de modélisation.

La comparaison des performances est au cœur de nombreuses tâches d'optimisation et est largement utilisée dans divers contextes d'analyse statistique. La communauté de modélisation des choix s'est principalement appuyée sur les approches théoriques à l'analyse des données, et plus précisément - les approches économétriques, pour la comparaison des performances des modèles. Malheureusement, l'accent mis sur le modèle, comme dans *modèle statistique* ou *modèle économétrique*, a imposé une série de contraintes aux implications de telles comparaisons. La comparaison de l'ajustement aux données ou des intervalles de confiance pour les estimations pourrait ne pas toujours être optimale, car la performance n'est pas toujours définie dans ces termes.

Cette thèse offre une argumentation sur la raison pour laquelle les perceptions de la performance ne devraient pas se limiter aux indicateurs de performance statistique de base. Bien que des approches alternatives à la compréhension de la performance aient fait leur apparition dans la littérature scientifique depuis des décennies, il n'y avait aucun travail agrégeant ces pratiques et fournissant une procédure complète et conviviale pour leur application. Seules quelques rares publications ont établi le lien entre les estimations du modèle et les implications pour les politiques publiques ou d'autres résultats de l'utilisation des estimations. L'accent est mis en particulier sur ces dimensions précises : le lien entre la question de recherche et la performance des techniques appliquées. En raison de la complexité et des particularités des modèles les plus avancés, l'argumentation plaide en faveur de l'évaluation de la performance non pas sur le modèle, mais sur la procédure d'analyse des données dans son ensemble.

La principale contribution à la comparaison des performances des modèles sous la forme d'un framework d'analyse de performance axé sur la procédure est présentée dans la Section 2.5. Il fournit les lignes directrices à la fois pour les utilisateurs experts et non-expérimentés de l'évaluation et la comparaison des performances des modèles. Plusieurs illustrations de son utilisation sont fournies dans

différents contextes. Tout d'abord, plusieurs études avancées de la littérature sont positionnées selon le framework proposé pour mieux illustrer comment les pratiques existantes dans l'analyse des performances des modèles sont prises en compte par cet ensemble d'outils. Une série d'études appliquées est ensuite réalisée pour illustrer comment les futures études pourraient être guidées par le framework proposé. Une discussion est fournie pour chacun des cas explorés, et l'évolution du framework peut être retracée à travers les illustrations.

Dans le dernier chapitre, le Chapitre 3, le framework de comparaison des performances introduit dans la section précédente est exploré plus en détail à travers une série d'études de cas. Chaque étude de cas explore différents éléments du framework, se concentrant sur les relations entre les étapes de modélisation, les problèmes d'acquisition de données et la modélisation statistique dans le cadre plus large de la procédure d'analyse des données. La première étude combine l'économétrie et les modèles d'apprentissage automatique pour la modélisation des préférences de choix des consommateurs, introduisant un framework de simulation et de test de théorie. La deuxième étude se concentre sur la tâche d'élicitation de WTP, évaluant systématiquement la performance du modèle dans ce contexte en tenant compte des éventuelles erreurs de spécification, des changements de taille d'échantillon et de l'équilibre des données. La troisième étude de cas explore la comparaison entre les modèles économétriques et d'apprentissage automatique dans le contexte de la modélisation du choix de mode de déplacement, utilisant l'ensemble de données *swissmetro* et des échantillons synthétiques pour contraster les modèles de choix discrets classiques avec les alternatives émergentes d'apprentissage automatique dans la tâche d'estimation des WTP. Malgré les incohérences potentielles dans la vision et la présentation du framework en raison des stades de maturation variables entre ces trois articles, le chapitre vise à offrir des aperçus de l'analyse et de la comparaison des performances des modèles pour une sélection efficace, illustrant l'évolution du framework à travers différentes œuvres. Chaque section offre une brève discussion sur les connaissances acquises et les différences par rapport à la version finale du framework.

En conclusion, cette thèse aborde les défis découlant de l'évolution des techniques d'analyse de données, en particulier dans le contexte de la modélisation du choix des consommateurs. Le framework proposé d'évaluation et de comparaison des performances, détaillé dans le Chapitre 2, émerge comme un ensemble d'outils précieux pour une exploration systématique des performances des modèles. En déplaçant l'accent des simples indicateurs de performance statistique à une compréhension complète des implications pour les questions de recherche explorées, le framework offre une approche nuancée de l'évaluation des performances des modèles. Les études de cas dans le Chapitre 3 illustrent davantage l'application du framework dans divers contextes, mettant en lumière les relations entre les étapes de modélisation, les problèmes d'acquisition de données et la modélisation statistique dans le cadre plus large de la procédure d'analyse des données. Malgré la nature évolutive du framework à travers les études, ce travail contribue au discours continu sur l'évaluation des performances des modèles et, par conséquent, sur la sélection des modèles.

Table of contents

| | |
|--|-----------|
| Abstract | i |
| Acknowledgements | ii |
| Introduction | 1 |
| 1. Choice modelling: At the intersection of disciplines | 4 |
| 1.1. Current state of Choice Modelling | 5 |
| 1.2. The specificity of the discrete choice modelling | 6 |
| 1.2.1. Choice Analysis in Economics | 8 |
| 1.2.1.1. Application fields, problematic and research questions | 9 |
| 1.2.1.2. Target metrics of interest | 15 |
| 1.2.2. Discrete Choice Modelling | 16 |
| 1.2.2.1. Data driven approach and classification | 18 |
| A. Artificial neuron and Perceptron | 20 |
| B. Multilayer Perceptron (MLP) | 21 |
| 1.2.2.2. Theory driven approach and Classic Choice Modelling | 24 |
| A. Multinomial Logistic Regression (MNL) | 27 |
| B. Nested Logistic Regression (NL) | 28 |
| C. Mixed Multinomial Logistic Regression (MMNL) | 29 |
| D. Integrated Choice and Latent Variable Models (ICLV) and Hybrid Choice Models (HCM) | 29 |
| 1.2.2.3. Other theory driven approaches and irrational behaviour | 30 |
| A. Expected Utility (EU) | 32 |
| B. Prospect Theory (PT) | 33 |
| C. Decision Field Theory (DFT) | 34 |
| D. Quantum Probability (QP) and Quantum Decision Theory (QDT) | 35 |
| 1.2.3. Concluding remarks | 36 |
| 1.3. Taxonomy issues | 37 |
| 1.3.1. Existing taxonomies | 37 |
| 1.3.1.1. Bottom-up approach | 38 |
| 1.3.1.2. Top-down approach | 40 |
| 1.3.1.3. Alternative representation | 40 |
| 1.4. The vocabulary and terminology | 42 |
| 1.4.1. Models and modelling | 42 |
| 1.4.2. Performance | 45 |
| 1.4.2.1. Estimates and derived indicators | 46 |

| | | |
|-----------|---|-----------|
| 1.4.2.2. | Predictive qualities | 46 |
| 1.4.2.3. | Resource efficiency | 47 |
| 1.4.2.4. | Generalisations | 48 |
| 1.5. | Conclusion | 48 |
| 2. | A universal performance comparison framework | 50 |
| 2.1. | A need for unified methodology | 51 |
| 2.2. | Performance comparison issues | 52 |
| 2.2.1. | Scientific procedures | 53 |
| 2.2.2. | Performance evaluation | 58 |
| 2.2.2.1. | Output based performance metrics | 60 |
| 2.2.2.2. | Direct estimates metrics | 60 |
| 2.2.2.3. | Derived metrics | 61 |
| 2.2.3. | First framework elements | 61 |
| 2.3. | Data constraints and simulation | 63 |
| 2.3.1. | Data acquisition | 64 |
| 2.3.1.1. | Data simulation | 65 |
| 2.3.1.2. | Data collection | 67 |
| 2.3.2. | Experimental design and sources of bias | 68 |
| 2.4. | Models and their capabilities | 70 |
| 2.4.1. | Statistical models | 73 |
| 2.4.2. | Data transformation | 74 |
| 2.4.3. | Algorithms | 75 |
| 2.4.4. | Software choice | 76 |
| 2.4.4.1. | R language | 76 |
| 2.4.4.2. | Python language | 77 |
| 2.4.4.3. | SAS software | 77 |
| 2.4.4.4. | Stata software | 77 |
| 2.4.4.5. | Julia language | 78 |
| 2.5. | Framework presentation | 78 |
| 2.5.1. | Data analysis | 79 |
| 2.5.2. | Complete framework | 80 |
| 2.6. | Existing studies in framework context | 83 |
| 2.6.1. | Applied study procedure | 83 |
| 2.6.2. | Theoretical innovation introduction procedure | 84 |
| 2.6.3. | Theoretical study procedure | 86 |
| 2.7. | Concluding remarks | 87 |
| 3. | Framework in action: Case studies | 89 |
| 3.1. | Introduction | 90 |
| 3.2. | Case 1: Theoretical assumptions | 91 |
| 3.2.1. | Introduction | 91 |
| 3.2.2. | Methodology | 92 |
| 3.2.2.1. | Artificial dataset | 93 |
| 3.2.2.2. | Modelling consumer choices | 94 |

| | | |
|----------|--|------------|
| 3.2.2.3. | Performance measures | 95 |
| 3.2.3. | Results | 96 |
| 3.2.3.1. | Data | 96 |
| 3.2.3.2. | Estimation results | 98 |
| 3.2.3.3. | Performance comparison | 100 |
| 3.2.4. | Conclusion | 101 |
| 3.2.5. | Discussion | 102 |
| 3.3. | Case 2: Dataset acquisition | 104 |
| 3.3.1. | Research question | 104 |
| 3.3.2. | Methodology and context | 105 |
| 3.3.2.1. | Willingness to Pay | 105 |
| 3.3.2.2. | Performance comparison framework | 106 |
| 3.3.3. | Application | 108 |
| 3.3.3.1. | Dataset description | 108 |
| 3.3.3.2. | Simulation | 109 |
| 3.3.3.3. | Estimation | 110 |
| 3.3.3.4. | WTP and model performance | 111 |
| 3.3.4. | Conclusion | 113 |
| 3.3.5. | Discussion | 113 |
| 3.4. | Case 3: Statistical modelling | 115 |
| 3.4.1. | Introduction | 115 |
| 3.4.2. | Performance comparison and model selection | 116 |
| 3.4.2.1. | Model differences | 117 |
| A. | MNL and NL model translation to NN graph | 118 |
| B. | ASUDNN-MNL and -NL versions | 121 |
| 3.4.2.2. | Performance comparison framework | 122 |
| 3.4.3. | Application | 124 |
| 3.4.3.1. | Data simulation strategy | 125 |
| 3.4.3.2. | Model estimation | 126 |
| 3.4.3.3. | Results | 127 |
| A. | Prediction quality | 127 |
| B. | Direct effects | 128 |
| C. | Resource efficiency | 129 |
| 3.4.4. | Conclusion | 130 |
| 3.4.5. | Discussion | 130 |
| | Conclusion | 132 |
| | Glossary | 135 |
| | Acronyms | 135 |
| | Special terms | 138 |
| | Tables | 141 |
| | List of Figures | 141 |
| | List of Tables | 142 |

| | |
|--|------------|
| Index | 144 |
| Bibliography | 146 |
| Appendices | 167 |
| A. Bibliometric study | 167 |
| A.1. Motivation and research objectives | 168 |
| A.2. Data collection and preliminary analysis | 168 |
| A.3. Advanced analysis | 172 |
| A.3.1. General information | 172 |
| A.3.2. Keywords | 173 |
| A.3.3. Co-occurrences | 177 |
| A.3.4. Citations | 179 |
| A.4. Analysis by subdomain | 182 |
| A.4.1. Policy | 183 |
| A.4.2. Preferences or attitudes | 186 |
| A.5. Conclusion | 188 |
| B. Extracting economic information from Neural Networks | 190 |
| B.1. Statistical and Machine Learning perspective | 191 |
| B.2. Introduction to Neural Networks | 191 |
| B.2.1. Artificial Neuron and Perceptron | 191 |
| B.2.1.1. Gradient descent | 193 |
| B.2.1.2. Perceptron algorithm | 194 |
| B.2.2. Adaline | 195 |
| B.2.2.1. Adaline algorithm | 195 |
| B.2.3. Multilayer Perceptron | 196 |
| B.2.3.1. Backpropagation algorithm | 199 |
| B.2.3.2. MLP Algorithm | 200 |
| B.2.3.3. Universal approximator | 201 |
| B.2.4. Convolutional Neural Network (CNN) | 202 |
| B.3. NN in Choice Analysis | 202 |
| B.3.1. CNN design for MNL imitation | 204 |
| B.3.2. Alternative Utility Specific DNN (ASU-DNN) | 206 |
| B.3.3. Extracting interpretable information from NN | 207 |
| B.4. Conclusion | 209 |
| C. Independence from Irrelevant Alternatives | 210 |
| C.1. Traditional formulation of IIA | 211 |
| C.2. History and ambiguity of the IIA | 212 |
| C.2.1. IIA(A) by Arrow (1951) | 212 |
| C.2.2. IIA(RM) by Radner and Marschak (1954) | 213 |
| C.2.3. IIA(L) by Luce (1957) | 213 |
| C.2.4. Contraction consistency by J. F. Nash (1950) | 214 |

| | |
|--|------------|
| C.2.5. Criticism of the minimax decision theory by Savage (1951) | 214 |
| C.3. Linking the IIA with reality | 214 |
| C.4. Mitigation of the IIA inconsistency | 215 |
| C.4.1. IIA tests and validation | 215 |
| C.4.1.1. Hausman-McFadden test | 217 |
| C.4.1.2. McFadden test | 217 |
| C.4.2. IIA treatment | 217 |
| C.5. Conclusion | 218 |
| D. Research practices: Unstructured interviews | 219 |
| D.1. Motivation and research objectives | 220 |
| D.2. Methodology | 221 |
| D.2.1. Target audience | 222 |
| D.2.2. Survey | 222 |
| D.2.2.1. Cultural and background differences | 222 |
| D.2.2.2. Research question and problematic definition | 222 |
| D.2.2.3. Target indicators definition | 223 |
| D.2.2.4. Modelling techniques | 223 |
| D.2.2.5. Hypothesis / Theoretical assumptions | 223 |
| D.2.2.6. Data collection (and experimental design) | 223 |
| D.2.2.7. Encountered difficulties and identified limitations | 224 |
| D.3. Results | 224 |
| D.4. Conclusion | 225 |
| E. Software packages | 227 |
| E.1. ‘dcesimulatr’: A DCE simulation toolset | 228 |
| E.2. ‘performancer’: Performance estimation functions collection | 228 |
| F. Synthesis in French | 229 |
| Introduction | 230 |
| F.1. Modélisation du choix : À la croisée des disciplines | 232 |
| F.2. Un framework universel de comparaison des performances | 234 |
| F.3. Le framework en action : Études de cas | 236 |
| F.4. Conclusion | 238 |
| Table of contents | 241 |