

Essays on

Advanced Discrete Choice Applications

Dissertation (kumulativ)

Eingereicht bei der
Fakultät für Betriebswirtschaft
der Universität Hamburg

zur Erlangung des akademischen Grades einer Doktorin der
Wirtschafts- und Sozialwissenschaften (Dr. rer. pol.) (nach PromO 2010)

vorgelegt von

Dipl-Verk.wirtsch. Frauke Korfmann geb. Seidel
geboren am 14.05.1982
in Görlitz

Schlankreye 4
20144 Hamburg
frauke@korfmann.net
+49 15152917978

Datum der Disputation: 07.03.2018

Vorsitz der Prüfungskommission: Prof. Dr. Henrik Sattler
Erstgutachter: Prof. Dr. Knut Haase
Zweitgutachter: Prof. Dr. Guido Voigt

Table of Contents

- 1 Synopsis 1**
 - 1.1 Introduction 1
 - 1.2 Research Contributions 2
 - 1.3 Declarations on Co-Authorships 4

- 2 Papers 7**
 - 2.1 Exposing Unobserved Spatial Similarity: Evidence from German School Choice Data . . 8
 - 2.2 Students' perceptions, academic departments' image, and major choice in business
administration studies - The example of Hamburg Business School. 31
 - 2.3 Synthetic Data Sets with Non-Constant Substitution Patterns for Fare Class Choice . . 75
 - 2.4 Choice-Based Revenue Management with Flexible Substitution Patterns 100

- A Eidesstattliche Versicherung 131**

1 Synopsis

1.1 Introduction

Modeling choices of discrete alternatives comprises of the utilization of discrete choice models that describe and predict the resulting choice decisions to analyze individual choice processes. In sum, these disaggregate individual choices provide demand for certain products or services (i.e., transportation modes, flight itineraries, school locations, travel destinations). The theoretical modeling framework is provided by random utility theory. Thus, discrete choice models that are derived by assuming that individuals act rationally and make their choice decision under the utility maximization principle are called random utility models (RUM) (Marschak, 1960). Individual utility thereby is a random variable that is composed of observed characteristics of the decision maker, observed attributes of the considered alternatives and a random term. The latter follows a distribution that is specified by the researcher such that individual choice behavior is best reflected by the applied demand model. Thus, we use utility to explain choice decisions of individuals.

Different specifications of the distribution of the random term of utility lead to the derivation of various discrete choice models. A first model of economic decision making that yielded choice probabilities for each alternative in a finite choice set was the multinomial logit model (MNL). It was developed by Daniel McFadden in the late 1960s (McFadden, 2001). In the MNL model, the utility is assumed to follow an identical and independent (iid) type I extreme value distribution. This may, in some choice situations, cause false predictions of demand for choice alternatives since applying an MNL model results in constant demand substitution patterns between available alternatives. Thus, if one alternative is removed from the original choice set demand for this alternative will be allocated to the remaining choice options such that the ratio of choice probabilities between any two of the remaining alternatives is constant. This characteristic of the MNL is known as independence of irrelevant alternatives (IIA) assumption. It is useful in choice situations where demand between choice alternatives is in fact independent. However, if at least some of the alternatives within the choice set are assumed to share common unobserved characteristics, applying an MNL demand model is inappropriate since it might lead to false demand predictions. Therefore, certain choice situations require the application of more advanced discrete choice models that exhibit non-constant demand substitution patterns like, for example, the nested logit (NL) or generalized nested logit (GNL) models.

In addition to modeling and predicting demand in certain choice situations more accurately by applying advanced discrete choice models, recently a new area of research is developing. In choice-based optimization, discrete choice models are integrated into optimization models to represent demand for products or services as well as its effects on optimization outcomes more accurately. In operations research, mathematical model formulations are usually applied to optimize a system where demand is considered to be known in advance. Discrete choice models, on the other hand, help to predict demand for a given system. Thus, by combining both approaches, choice-based optimization models enable researchers to explicitly account for effects of demand endogenization. Such optimization models cover for example the research areas of airline revenue management, network planning in public transportation, health care facility location planning and assortment optimization. Thereby, the direct incorporation of discrete choice probabilities into a mathematical program would result in non-linear models (Müller and Haase, 2016). Therefore, the MNL demand models' constant substitution patterns that are a result of its IIA property are exploited to obtain linear reformulations (Benati and Hansen,

2002; Haase, 2009; Davis et al., 2013). However, since the applicability of the MNL demand model is limited, the incorporation of discrete choice demand models with non-constant substitution patterns into mathematical programs is more desirable.

In this context, the research contribution of this thesis regarding discrete choice applications and their integration into choice-based optimization approaches can be summarized as follows:

- First, for the modeling tasks of school choice, university major choice and fare class choice in airline revenue management approaches are developed to examine individual choice processes and to predict demand for the available choice alternatives. The applied approaches are connected methodologically by covering advanced discrete choice applications that prove to be superior to more simple and, thus, limited approaches. The articles presented in this thesis capture (i) effects of non-constant demand substitution between choice alternatives for the school choice and fare class choice modeling tasks and (ii) the impact of students' perceptions and beliefs on their choice of a university major.
- By further exploiting the modeling task of fare class choice, an approach for choice-based optimization in airline revenue management is developed that allows for the incorporation of a general discrete choice demand model with non-constant substitution patterns. As a result, limitations in optimization models that arose from an integration of deterministic demand figures or simple demand models like MNL can be overcome.

1.2 Research Contributions

During my time at the Institute of Transport and Economics at Hamburg University, my scientific work covered the research areas of discrete choice applications and choice-based optimization. Based on this, four articles covering different areas of advanced discrete choice applications arose. Table 1.1 gives an overview of the articles with authors and current status.

The research provided within this thesis methodologically covers discrete choice demand models that are either applied to collected data or are utilized to generate data with desired characteristics regarding demand substitution patterns. Thereby, the articles of Müller et al. (2012), Seidel (2014) and Seidel et al. (2016) cover a general nested logit (GNL) model for school location choice, a nested logit (NL) model for airline fare class choice and an integrated choice and latent variable (ICLV) model for university major choice. Concerning the specific modeling task, these papers examine the advantages of applying demand models with flexible substitution patterns over discrete choice models that exhibit the more restrictive non-constant substitution patterns. Thus, we obtain insights into individual choice processes as well as demand predictions for choice alternatives by applying the most appropriate demand model according to our assessment of the specific modeling task.

The paper by Müller et al. (2012) examines school choice in the city of Dresden, Germany for secondary schools (German school form Gymnasium) to analyze future effects of school network planning. Since we assume that some of the considered schools share common unobserved characteristics regarding their spatial location, we apply a generalized nested logit (GNL). This type of discrete choice model allows for non-constant spatial substitution between competing school locations while its choice probabilities exhibit an analytical closed-form. Within the spatial choice context of school choice this model type, in contrast to the much simpler MNL model, allows for a reproduction of the decision-making process that is as realistic as possible.

In Seidel et al. (2016) we apply an integrated choice and latent variable (ICLV) model to data on students' major choice decisions that we collected at Hamburg University in 2013. Besides observed factors that influence individual choice decisions, this type of model allows for the incorporation of psychometric factors into the modeling framework. Such psychometric factors are individuals' attitudes,

Article	Authors	Status
Exposing Unobserved Spatial Similarity: Evidence from German School Choice Data (2012)	Sven Müller, Knut Haase, Frauke Seidel	Published in Geographical Analysis
Students' perceptions, academic departments' image, and major choice in business administration studies - The example of Hamburg Business School (2016)	Frauke Seidel, Sven Müller, Knut Haase	Major Revision
Synthetic Data Sets with Non-Constant Substitution Patterns for Fare Class Choice (2014)	Frauke Seidel	Published in Zeitschrift für Verkehrswissenschaft
Choice-Based Revenue Management with Flexible Substitution Patterns (2017)	Frauke Seidel, Sven Müller, Knut Haase	Working Paper

Table 1.1: Research overview

perceptions and beliefs that are assumed to have a considerable effect on the outcome of decision processes. In fact, Vij and Walker (2016) state that such factors play an essential role in decision making and may even override the influence of observable variables on individual choice behavior. By collecting the students' assessments of various psychometric factors that we assume to be related to their major choice decision (i.e., career opportunities, quality of supervision and internationality of courses) the article examines how the perceived image of the offered majors affects students' choice decisions.

The paper by Seidel (2014) focuses on the generation of synthetic demand data with non-constant substitution patterns according to an NL demand model. The considered modeling task is fare class choice in airline revenue management where accurate data is either unavailable or lacking important information (Hess et al., 2010). Therefore, the article focuses on the development of a methodology to generate demand data with non-constant substitution patterns. The obtained results provide the basis for further research towards choice-based optimization in airline revenue management where simplifying assumptions about demand (i.e., the independent demand model) do not allow for the consideration of demand dependencies between choice alternatives.

Extending the research by Seidel (2014), the fourth paper presented in this thesis integrates the before mentioned fare class choice demand model into an optimization model for airline revenue management. The approach by Seidel et al. (2017) is a choice-based optimization model that, like an airline booking system, reacts to passengers' purchase requests by (i) offering available alternatives (a choice set) to arriving passengers and (ii) determining the prices for tickets in the offered fare classes. Thereby, the outcomes of both (i) and (ii) depend on the substitution patterns of the considered demand model.

Either part of the problem, the demand model, and the optimization model, are mutually dependent. The outcome of the optimization model (i.e., chosen alternatives and revenues) depends on the demand for offered fare classes. However, demand depends, amongst other things, on the ticket price, which, at the same time, is a decision variable in our optimization problem. Thus, we face an endogeneity problem that we solve by directly incorporating the choice problem of airline passengers into the mathematical model. By comparing the outcomes of the developed optimization model for both a demand model with constant and a demand model with non-constant substitution patterns, we prove the superiority of our approach. Furthermore, we overcome some of the limitations imposed on airline revenue management optimization models by the simplifying assumptions of the independent demand model (Talluri and Van Ryzin, 2004, p. 33).

1.3 Declarations on Co-Authorships

According to

- §6 Abs. 3 der Promotionsordnung der Fakultät für Wirtschafts- und Sozialwissenschaften vom 24. August 2010 der Universität Hamburg -

the following declarations contain my work contribution to the articles presented within this thesis. The overall contribution is thereby further divided into the following subtasks: concept development, methodological development, literature review, data collection, data analysis, development of modeling script/model formulation, discussion of results, manuscript preparation. For each paper, my share of contribution to the specific subtasks is as follows:

Exposing Unobserved Spatial Similarity: Evidence from German School Choice Data (2012)

- Concept development: no share
- Methodological development: partially
- Literature review: partially
- Creation of modeling script/model formulation: completely for calculation of cross elasticities
- Data collection: no share
- Data analysis: completely for cross elasticities
- Discussion of results: partially in general, completely for discussion of cross elasticities
- Manuscript preparation: predominantly

Students' perceptions, academic departments' image, and major choice in business administration studies - The example of Hamburg Business School (2016)

- Concept development: predominantly
- Methodological development: predominantly
- Literature review: completely
- Creation of modeling script/model formulation: completely
- Data collection: partially
- Data analysis: completely
- Discussion of results: predominantly
- Manuscript preparation: completely

Synthetic Data Sets with Non-Constant Substitution Patterns for Fare Class Choice (2014)

- Concept development: completely
- Methodological development: completely

- Literature review: completely
- Data collection: completely
- Data analysis: completely
- Creation of modeling script/model formulation: completely
- Discussion of results: completely
- Manuscript preparation: completely

Choice-Based Revenue Management with Flexible Substitution Patterns (2017)

- Concept development: completely
- Methodological development: predominantly
- Literature review: predominantly
- Creation of modeling script/model formulation: predominantly
- Data collection: completely
- Data analysis: completely
- Discussion of results: completely
- Manuscript preparation: completely

Bibliography

- Benati, S., Hansen, P., 2002. The maximum capture problem with random utilities: Problem formulation and algorithms. *European Journal of Operational Research* 143, 518–530.
- Davis, J., Gallego, G., Topaloglu, H., 2013. Assortment planning under the multinomial logit model with totally unimodular constraint structures. Department of IEOR, Columbia University. Available at http://www.columbia.edu/~gmg2/logit_const.pdf.
- Haase, K., 2009. Discrete location planning. Tech. Rep. WP-09-07, Institute for Transport and Logistics Studies, University of Sydney.
- Hess, S., Niemeier, H.-M., Forsyth, P., Müller, J., Gillen, D., 2010. Airport Competition: The European Experience. Ashgate Publishing, Ltd., Ch. Modelling air travel choice behaviour, pp. pp. 151–176.
- Marschak, J., 1960. Binary-choice constraints and random utility indicators. In: *Mathematical Methods in the Social Sciences*. Stanford University Press, pp. 312–329.
- McFadden, D., 2001. Economic choices. *American Economic Review* 91 (3), 351–378.
- Müller, S., Haase, K., 2016. On the product portfolio planning problem with customer–engineering interaction. *Operations Research Letters* 44 (3), 390–393.
- Müller, S., Haase, K., Seidel, F., 2012. Exposing unobserved spatial similarity: Evidence from German school choice data. *Geographical Analysis*. 44, 65–86.
- Seidel, F., 2014. Synthetic data sets with non-constant substitution patterns for fare class choice. *Zeitschrift für Verkehrswissenschaft* 85 (1), 32–55.
- Seidel, F., Müller, S., Haase, K., 2016. Students' perceptions, academic departments' image, and major choice in business administration studies - the example of hamburg business school. Under Review.
- Seidel, F., Müller, S., Haase, K., 2017. Choice-based revenue management with flexible substitution patterns. Working Paper. Hamburg University.
- Talluri, K., Van Ryzin, G., 2004. *The Theory and Practice of Revenue Management*. Springer Verlag.
- Vij, A., Walker, J. L., 2016. How, when and why integrated choice and latent variable models are latently useful. *Transportation Research Part B: Methodological* 90, 192–217.

2 Papers

2.1 Exposing Unobserved Spatial Similarity: Evidence from German School Choice Data

Exposing Unobserved Spatial Similarity: Evidence from German School Choice Data

Sven Müller, Knut Haase, Frauke Seidel

Institute of Transport Economics, University of Hamburg, Hamburg, Germany

In a spatial context, flexible substitution patterns play an important role when modeling individual choice behavior. Issues of correlation may arise if two or more alternatives of a selected choice set share characteristics that cannot be observed by a modeler. Multivariate extreme value (MEV) models provide the possibility to relax the property of constant substitution imposed by the multinomial logit (MNL) model through its independence of irrelevant alternatives (IIA) property. Existing approaches in school network planning often do not account for substitution patterns, nor do they take free school choice into consideration. In this article, we briefly operationalize a closed-form discrete choice model (generalized nested logit [GNL] model) from utility maximization to account for spatial correlation. Moreover, we show that very simple and restrictive models are usually not adequate in a spatial choice context. In contrast, the GNL is still computationally convenient and obtains a very flexible structure of substitution patterns among choice alternatives. Roughly speaking, this flexibility is achieved by allocating alternatives that are located close to each other into nests. A given alternative may belong to several nests. Therefore, we specify a more general discrete choice model. Furthermore, the data and the model specification for the school choice problem are presented. The analysis of free school choice in the city of Dresden, Germany, confirms the influence of most of the exogenous variables reported in the literature. The estimation results generally indicate the applicability of MEV models in a spatial context and the importance of spatial correlation in school choice modeling. Therefore, we suggest the use of more flexible and complex models than standard logit models in particular.

Introduction

Space plays an important role in evaluating individual choices for several goods and services. School choice decisions especially exhibit features of choice situations that are highly influenced by spatial factors. In this article, we first give a brief introduction and a short overview of literature concerned with spatial choice modeling. Later, we turn our attention to the main aspects of the German school system and school choice modeling in particular.

Spatial Choice Modeling

A frequently used statistical model to analyze discrete choices is the multinomial logit (MNL) model. Its popularity is owed to, among other reasons, its utility-maximizing behavior and closed-form choice

Correspondence: Sven Müller—Institute of Transport Economics, University of Hamburg, Von-Melle-Park 5, 20146 Hamburg, Germany
e-mail: sven.mueller@wiso.uni-hamburg.de

Submitted: August 17, 2009. Revised version accepted: January 20, 2011.

probabilities. This model exhibits the property of independence of irrelevant alternatives (IIA), which (according to some researchers) is seen as a major shortcoming. This property may lead to model misspecification or false prediction of market shares. Haynes, Good, and Dignan (1988) argue that spatial choice problems especially show characteristics (e.g., random taste variation) that are difficult to handle with the MNL. Hunt, Boots, and Kanaroglou (2004) further point out that some researchers emphasize that spatial choice models have to be seen as distinct from discrete choice modeling due to incapacities introduced by space. Meanwhile, developments in discrete choice analysis now allow existing models to account for a wide range of substitution patterns, including features of space (Bolduc, Fortin, and Fournier 1996; Train 1999; Walker and Li 2007). However, regarding the specifics of spatial choice (i.e., correlation in unobserved utility), little attention in the geographic literature is paid to the application of choice models other than the MNL. Hunt, Boots, and Kanaroglou (2004) state that discrete choice models should be increasingly applied in geographic contexts in order to evaluate their possible benefits. Attempts to account for spatial correlation involve the adjustment of the systematic component of utility or the implementation of a choice model that exhibits more flexible substitution patterns. The standard logit model enables constant substitution among alternatives. In contrast, the generalized nested logit (GNL) model allows for correlations in unobserved attributes by grouping alternatives that share unobserved (spatial) variability into common nests. Our focus here is on applying a GNL within the framework of random utility theory for school choice in the city of Dresden, Germany.

School Choice

Fluctuating student numbers over time and space force municipalities to adjust the number, the locations, and the capacities of schools. Within the framework of (long-term) school network planning, officials need to know factors that influence students to choose a certain school in order to derive expected utilization. The literature about school choice modeling usually focuses on racial mix, tuition fees, and travel-to-school distance (see section “School Choice Modeling”). Because free school choice is seldom found in many countries, most school location planning approaches do not account for spatial substitution (Müller 2008; Müller, Haase, and Kless 2009), but some references lead one to believe that spatial substitution patterns between school locations exist (Manski and Wise 1983; Borgers et al. 1999; Müller 2009).

The concept of utility entails a compensatory decision process. It presumes that students’ choices involve trade-offs among the attributes characterizing schools. For example, a student may choose a school located far away from her location if the profile offered by that school (e.g., math and languages) compensates for the increased travel distance. Based on such trade-offs, each student selects the school with the highest utility value. The focus on utility maximization in this article arises from its strong theoretical background.

The utility-maximization rule is robust; that is, it provides a good description of choice behavior even if students use different rules (Koppelman and Bhat 2006, pp. 12–13). German students are free to choose a secondary school in which to enroll. This means enrollment is not determined by location of the students, because school districts do not exist. In general, after 4 years in primary school, a student enrolls in a secondary school. Based on their academic ability, students are allowed to enroll in either *Mittelschule* or *Gymnasium*. The latter can be seen as a special type of secondary school that prepares graduates to attend the university. The degree that students have when graduating from *Gymnasium*, therefore, is equivalent to a high school degree in the United States. Figure 1 shows the structure of the German school system as well as the number of grades and the corresponding students’ ages.

Students rarely switch from one school to another on the same educational level (i.e., switch from one *Gymnasium* to another). The secondary school choice decision is strongly dependent on an educational recommendation a student receives after having finished 4 years of primary school. As a result, students showing good scholastic performance are allowed to enroll in *Gymnasium*, while less capable students have to attend *Mittelschule*. In general, enrolling in *Gymnasium* is prohibited if the educational recommendation is not for *Gymnasium*. Students rarely choose to enroll in *Mittelschule* when their educational recommendation qualifies them for *Gymnasium*.

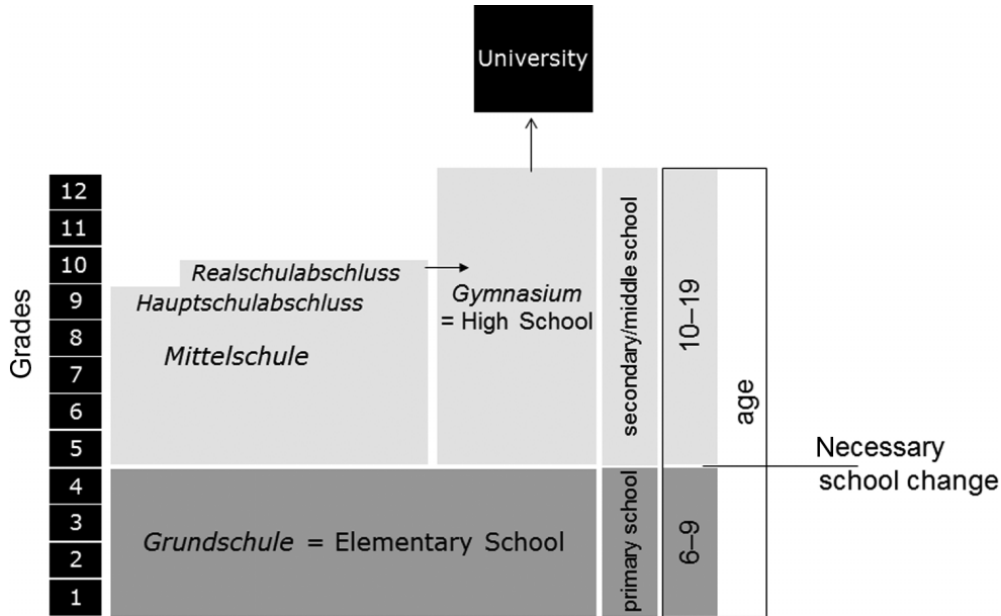


Figure 1. Main features of the educational system of Germany.

Unlike Gymnasium, Mittelschule schools generally are more homogeneous regarding their spatial distribution and offered profiles. Hence, in this article, we focus on students choosing a Gymnasium school in the city of Dresden, Germany. Gymnasium schools exhibit varying characteristics regarding the amount of education offered in subjects like sciences, languages, and music/arts. The objective here was to describe the development of a school choice model embedded in the framework of discrete choice analysis, considering spatial dependencies between the school locations under study. Although this is a specific application, the modeling framework for spatial choices presented can be easily applied to a wide range of spatial contexts, like demand modeling for recreational sites and other (non market) recreational goods and services. Valuable applications regarding the subject of spatial choice include recreational demand models (Train 1999) or housing location choice models (Guo and Bhat 2007).

Multivariate Extreme Value Models

The choice models we employ in this article are based on the assumptions of random utility theory. A decision maker n is assumed to choose from a set of available alternatives C_n alternative i such that utility $U_{ni} \geq U_{nj} \forall j \in C_n, j \neq i$. Note that $C_n \subseteq C$ with C : set of all alternatives under study. Because we do not observe all effects on utility-maximizing behavior, we decompose utility U_{ni} into a deterministic (or systematic) part V_{ni} and a stochastic part ϵ_{ni} :

$$U_{ni} = V_{ni} + \epsilon_{ni}. \quad (1)$$

Usually V_{ni} is linear in parameters:

$$V_{ni} = \sum_h \beta_{ih} x_{nih}. \quad (2)$$

The H independent variables x_{nih} describe alternative i and characteristics of decision maker n . The x_{nih} variables are weighted by coefficients β_{ih} . Because ϵ_{ni} is a random variable, we can only determine the probability that an individual n chooses i from her choice set of available alternatives C_n by

$$P_n(i | C_n) = \text{Prob}(V_{ni} + \epsilon_{ni} > V_{nj} + \epsilon_{nj}, \quad \forall j \in C_n, j \neq i) \quad (3)$$

$$= \text{Prob}(\epsilon_{nj} < V_{ni} - V_{nj} + \epsilon_{ni}, \quad \forall j \in C_n, j \neq i). \quad (4)$$

Now we have to make assumptions about the joint probability distributions for the random components of utility ϵ_{ni} in equation (1).

Multivariate extreme value (MEV) models constitute a large class of discrete choice models whose unifying attribute is that the stochastic part of utility ϵ_{ni} is distributed as a generalized extreme value for all alternatives (Train 2003, p. 80). Following McFadden (1978), different kinds of discrete choice models can be developed as special cases of the more general MEV model formulation. The generating function for different types of models (e.g., MNL and nested logit [NL]) is obtained by making specific assumptions about the cumulative distribution of the vector of unobserved utility $\epsilon_n = \langle \epsilon_{n1}, \dots, \epsilon_{nj} \rangle$ (Train 2003, pp. 83–100). Each instance of the MEV family is derived from a continuous and differentiable generating function,

$$G : \mathbb{R}_+^{C_n} \rightarrow \mathbb{R}_+, \quad (5)$$

which defines the cumulative distribution function (CDF) of the error terms and the choice model, respectively. The CDF of an MEV model takes the form

$$F_{\epsilon_n}(\xi_1, \dots, \xi_{C_n}) = e^{-G(e^{\xi_1}, \dots, e^{\xi_{C_n}})}, \quad (6)$$

whereas, in order for F to be a CDF, the μ -MEV-generating function G needs to exhibit the following properties:

- (1) $G(y)$ is a nonnegative function, $G(y) \geq 0 \quad \forall y_i \in \mathbb{R}_+^{C_n}$;
- (2) $G(y)$ is homogeneous of degree $\mu > 0$; that is, $G(\lambda y) = \lambda^\mu G(y)$, for $\lambda > 0$;
- (3) $G(y)$ asymptotically tends to infinity for each y_i tending to infinity: $\lim_{y_i \rightarrow \infty} G(y_1, \dots, y_i, \dots, y_{C_n}) = \infty$, for each $i = 1, \dots, C_n$; and,
- (4) the m th partial derivative of $G(y)$ with respect to m distinct y_i is nonnegative if m is odd and non positive if m is even, for any distinct indices $i_1, \dots, i_m \in \{1, \dots, C_n\}$.

The probability of choosing alternative i from a choice set C_n for an MEV model may be written as

$$P_n(i | C_n) = \frac{y_i G_i(y_1, \dots, y_{C_n})}{\mu G(y_1, \dots, y_{C_n})}, \quad (7)$$

where $y_i = e^{V_{ni}}$. By explicitly assuming that the generating function $G(y_1, y_2, \dots, y_{C_n})$ takes the form

$$G(y) = \sum_{i=1}^{C_n} y_i^\mu, \quad (8)$$

the MNL model is derived. Substituting equation (8) into (7) yields

$$P_n(i | C_n) = \frac{e^{\mu V_{ni}}}{\sum_{j \in C_n} e^{\mu V_{nj}}}, \quad (9)$$

where μ is a scale parameter that is not identified and has to be set to an arbitrary value (e.g., one) for model identification purposes. Equation 9 is the logit choice probability. In the case of the MNL, the random components of utility ϵ_{ni} in equation (1) are assumed to be independently and identically distributed extreme

value (iid EV), which is a special case of the assumption made for the error terms in MEV models. Note that equation (9) has a closed form and that the unknown coefficients β_{ik} of equation (2) can be provided relatively simply through maximum likelihood estimation.

Failure of IIA and Substitution Patterns

Although the MNL is applied in various situations, it has some severe shortcomings, particularly in a spatial choice context. The main issue concerning spatial choice (such as school choice) is the well-known IIA property (Luce 1959), a direct outcome of the assumption that the ϵ_{ni} are iid (Haynes, Good, and Dignan 1988). The IIA property ensures that the ratio of choice probabilities for any two alternatives is unaffected by the presence or change of any other alternative and its attributes. Therefore, a change in the probability of one alternative leads to identical changes in relative choice probabilities for all other alternatives. For example, let us assume, that a school network consists of five school locations available to a student and that the predicted choice probabilities from equation (9) equal 0.30, 0.12, 0.15, 0.18, and 0.25, respectively. Next, we assume that school location 5 is closed due to an expected overall decline in student numbers. Equation (9) predicts choice probabilities equal to 0.40, 0.16, 0.20, and 0.24 for the remaining four locations. The choice probability for every remaining alternative increases by one-third (i.e., a 33.33% relative change to choice probabilities). This rigid substitution pattern ignores the fact that some schools may be better substitutes for the closed site (e.g., because of spatial proximity to that school). Although whether IIA holds for given data is an empirical question and a matter of the specification of V_{ni} set, many geographers suggest that IIA is unlikely to hold in spatial choice applications. For example, Haynes and Fotheringham (1990) note that size, aggregation, dimensionality, spatial continuity, and variation and location characteristics of spatial choice data are likely to produce substitution patterns that violate IIA. In its strict form, IIA applies only to an individual student n and not to all students as a population. As Ben-Akiva and Lerman (1985, pp. 109–11) state, IIA often is misinterpreted as implying that the ratio of the shares of the population choosing any two alternatives is unaffected by the utilities of other alternatives (schools).

Many attempts in the past tried to overcome IIA weaknesses and to account for a richer pattern of substitution than that offered by the MNL (Hunt, Boots, and Kanaroglou [2004] for a more detailed overview). Unfortunately, most of these attempts were based on the logit model specified by McFadden (1975) (Timmermans and van der Waerden 1992). In general, the models used have not been consistent with random utility theory (Koppelman and Sethi 2000). As Hunt, Boots, and Kanaroglou (2004) point out, developments in discrete choice modeling are considerable, and today various models exist that are able to cope with spatial complexity. These models are classified as *closed-form* models, such as the MNL, and *open-form* models, such as the multinomial probit (MNP). The advantage of the closed-form models is their computational tractability, whereas the advantage of the open-form models is their flexibility. In the remainder of this article, we consider a closed-form model with a maximum of flexibility for considering (spatial) substitution patterns of choice alternatives (i.e., schools).

Two Closed-Form Discrete Choice Models with Flexible Substitution Patterns

The NL is a model that accounts for a wide range of substitution patterns that arise when alternatives share unobserved attributes. Its implementation is appropriate when alternatives faced by a decision maker can be grouped into subsets, or *nests*, in such a way that IIA holds between alternatives within each nest but not across nests. Due to the nesting of alternatives, the NL overcomes the proportional substitution across alternatives imposed by the MNL through IIA. Following Train (2003 p. 84), the NL is a more general formulation of the MNL that allows for correlation in unobserved utility. The generating function to derive the NL from equation (7) is

$$G(y) = \sum_{k=1}^K \left(\sum_{i=1}^{C_n} y_i^{\beta_{ik}} \right)^{\frac{\mu}{\beta_{ik}}}, \quad (10)$$

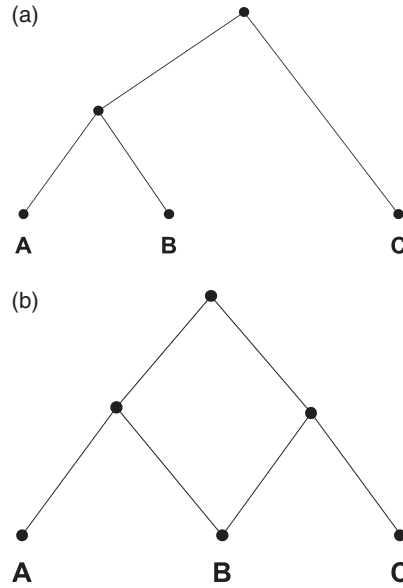


Figure 2. (a) Nesting structure of the nested logit. (b) Nesting structure of the generalized nested logit.

where K depicts the number of existing nests B_k . A separate scale parameter μ_k exists for each nest, so that only the ratios μ/μ_k are identified. Thus, a normalization of the scale parameter is required for model identification purposes.

Normalizing $\mu = 1$ is good practice, which is referred to as normalization from the top, although other normalization for the NL can be considered as well (Bierlaire 2006). Furthermore, $\mu/\mu_k = \sqrt{1 - \rho_{ij}}$, where ρ_{ij} denotes the correlation coefficient $\text{corr}(U_i, U_j)$. This is the correlation of the total utilities for any pair of alternatives in C_n that share the same nest (Ben-Akiva and Lerman 1985; Heiss 2002, p. 289). The scale parameter generally serves as an indicator for the independence among alternatives within a nest. Thus, a higher μ_k translates into a higher correlation between alternatives in that particular nest. Substituting equation (10) into (7) yields the following NL choice probability that individual n chooses alternative i :

$$P_n(i/C_n) = \frac{e^{\mu_k V_{ni}}}{\sum_{j \in C_{nk}} e^{\mu_k V_{nj}}} \frac{\left(\sum_{j \in C_{nk}} e^{\mu_k V_{nj}} \right)^{\mu/\mu_k}}{\sum_{m=1}^K \left(\sum_{j \in C_{nm}} e^{\mu_m V_{nj}} \right)^{\mu/\mu_m}}, \quad (11)$$

where $C_{nk} = B_k \cap C_n$, and k denotes the nest that contains alternative i . Figure 2a shows the nesting structure for an NL model with three alternatives, A , B , and C , available; that is, $C_n = \{A, B, C\}$. Alternatives that have similar unobserved attributes (here, alternatives A and B) are assigned to one nest.

For the NL model, every alternative belongs to only one nest. This aspect imposes an important restriction on the model insofar as this assumption might be inappropriate in some situations. Assume, for example, that alternative B shares some unobserved attributes not only with alternative A but also with alternative C . Such a nesting structure is presented in Fig. 2b and belongs to the GNL¹ model.

The proposed analytical formulation is derived from the MEV model in equation (7). An alternative may be a member of more than one nest to varying degrees. An allocation parameter α_{ik} reflects the extent to which alternative i is a member of nest k . The parameter α_{ik} is nonnegative, and $\sum_k \alpha_{ik} = 1 \quad \forall i$ for identification purposes. Further, α_{ik} may be interpreted as the portion of alternative i that is allocated to each nest k . If $\alpha_{ik} = 0$, alternative i does not belong to nest k , and if $\alpha_{ik} = 1$, the alternative belongs to nest k only. Values of α_{ik} between zero and one indicate a membership of an alternative i to multiple nests. A larger value

of α_{ik} means that alternative i shares a larger amount of common unobserved attributes with alternatives in nest k than with alternatives in other nests. The generating function to derive the choice probability for the GNL is

$$G(y) = \sum_{k=1}^K \left(\sum_{i=1}^{C_n} \alpha_{ik}^{\mu_k} y_i^{\mu_k} \right)^{\frac{\mu}{\mu_k}}. \quad (12)$$

Substituting equation (12) into (7) yields the probability function of the GNL

$$P_n(i | C_n) = \sum_{k=1}^K \frac{\left(\sum_{j \in C_{nk}} \alpha_{jk}^{\mu_k / \mu} e^{\mu_k V_{nj}} \right)^{\mu / \mu_k}}{\sum_{m=1}^K \left(\sum_{j \in C_{nm}} \alpha_{jm}^{\mu_m / \mu} e^{\mu_m V_{nj}} \right)^{\mu / \mu_m}} \frac{\alpha_{ik}^{\mu_k / \mu} e^{\mu_k V_{ni}}}{\sum_{j \in C_{nk}} \alpha_{jk}^{\mu_k / \mu} e^{\mu_k V_{nj}}}. \quad (13)$$

Due to the nest structure and the flexible allocation of alternatives to nests, the GNL does not exhibit the IIA of the MNL. Nevertheless, this advantage comes at the expense of an a priori assumption about the underlying correlation structure. If each alternative enters only one nest, with $\alpha_{ik} = 1 \forall i \in B_k$ and zero otherwise, the model becomes the NL of equation (11). If, in addition, $\mu_k = 1 \forall k$, the model becomes the MNL as in (9) (Train 2003, p. 95).

School Choice Modeling

The next section summarizes studies concerned with the modeling of school choice decisions and their influencing factors. In the past, several types of choice models have been employed. All studies identify distance to school as an important factor in individuals' choice decisions. Based on the findings in the literature and some data-related issues, we specify a spatial choice model for school choice in section "Data-related issues and model specification."

Literature Review

Manski and Wise (1983) initiated the growing now body of literature about the choice of educational facilities such as schools and universities. Borgers et al. (1999) employ an MNL based on stated choice data to identify the choice between Protestant, Catholic, and public schools in the Netherlands. They find evidence that school type (e.g., Montessori), religious affiliation, school size, and the distance between a student's location and a school are the most important decision-making factors. Moreover, they include substitution and availability effects to account for (spatial) competition between schools. Lankford, Lee, and Wyckoff (1995) model the choice across public, religious, and independent schools. Their MNP analysis reveals that school choice is affected by the racial composition of public schools, the crime rate, and the religious orientation of a school, as well as by the socioeconomic characteristics of a household, particularly the location of a household in a central city. Lankford and Wyckoff (2006) use a sequentially estimated NL to identify the effect of school choice on the racial segregation of students. They find that the racial composition of a school and the distance between a student's home and school influence school choice. They also find similarities between Catholic and private schools in unobserved factors. The mixed MNL (MMNL) model of Hastings, Kane, and Staiger (2006) furnishes evidence that distance traveled to school is the most important factor influencing the school decision. The combination of schools' mean test scores, household incomes, and parents' academic abilities results in a negative correlation between distance to school and mean test score. For German schools, Schneider (2004) shows that besides distance to school, household income has a strong influence on school choice. Finally, Jepsen and Montgomery (2009) use an NL to show that distance is the most important factor in deciding whether to enroll at a community college and about which school to choose. This finding is uncovered after controlling for tuition fees, school size, and socioeconomic variables.

This short review shows that distance seems to be by far the most important factor in the school choice process, indicating the possibility of spatial substitution between proximate schools. In our analysis, we apply the GNL model, which in contrast to MNP and MMNL is computationally easy to handle in identifying such substitution patterns. We control for most of the variables used in the studies mentioned here.²

Data-Related Issues and Model Specification

We aim to model the school choice of students living in the city of Dresden, Germany, and we use the survey by Müller, Tscharaktschiew, and Haase (2008). This study has been designed to model the travel-to-school mode choice. The data were collected at the schools under study, representing the endogenous variable in our study. The sample was stratified to l subsets of students with $l = 1, \dots, L$; $L \leq C$ contains all individuals who have chosen one particular alternative. Hence, this sample is choice based, which leads to problems in estimating GNL with standard maximum likelihood methods. Fortunately, we acquired data on the actual market share of each school. Therefore, we are able to employ the weighted exogenous sampling maximum likelihood (WESML) estimator³

$$\max_{\theta} \sum_{l=1}^L \sum_{n=1}^{N_l} \sum_{i \in C_n} y_{ni} \left(\frac{W_l}{H_l} \right) \ln [P_n(i | X_{ni}, \theta)], \quad (14)$$

where N_l denotes the set of students having chosen school l , y_{ni} is the choice by student n concerning school i (i.e., equals one, if student n chooses school i , zero otherwise), W_l denotes the known actual market shares, and H_l represent sample market shares, X_{ni} is the vector of exogenous variables, and θ is the vector of unknown coefficients β_{in} in equation (2). The fraction W_l/H_l is reported in the last column of Table 1. As stated by Ben-Akiva and Lerman (1985, pp. 238–9), this estimator yields a consistent estimate for θ . However, the WESML estimates are not necessarily asymptotically efficient.

From the survey sample, we select all students enrolled at Gymnasium ($N = 5,215$). Information about a chosen school, address (Fig. 3), and sex (about 44% of all students are male) is directly available from the survey. Table 1 reports the average travel distance to each school (based on street network) and its corresponding standard deviation. Moreover, from local authority statistics, we add the average income of the city district where a student is located. Average income is intended to account for differences among students' neighborhoods (Cullen, Jacob, and Levitt 2005).

Although we know that the median would be more appropriate, median income data are not available. Explanations in the remainder of this section are based on the following assumptions:

- (1) The likelihood of attending a private school is generally higher for a student originating from a wealthier city district than for a student living in a poorer district; and
- (2) As average income of a city district increases, the affinity of inhabitants toward education tends to increase.

City districts having a low average income are assumed to exhibit a large number of blue-collar workers, directly translating into a poorer social standing for the respective districts (Neu 2007). We also assume that the majority of households in a wealthier city district can afford tuition. Consequently, children of these households are more likely to enroll at private schools. If a student lives in a wealthy district, the student either stems from a wealthy household that enables him or her to enroll at private school or, if not, at least some in his or her peer group belong to a wealthy household. Hence, peer group pressure may influence the school choice decision of students from less wealthy families living in a well-off district. Finally, we have some attributes of the schools themselves, mainly profiles and size (Table 1 and Fig. 3). In the school choice context, one can imagine that one school is more similar to a second one than to other schools due to the same profile offered, authority, and spatial proximity.

While some of these similarities could be incorporated in V_{ni} , spatial similarity is particularly difficult to operationalize in V_{ni} . As Hunt, Boots, and Kanaroglou (2004) point out, spatial effects may be accounted

Table 1 School Attributes, Availability, Frequencies, and Weights

Short name	Profile			Size*	Distance to CBD (km)	Travel distance (m)		Avail. 2004	Number of observations	WESML weight
	a	b	c			Mean	SD			
STBE				3	0.8322	5,508.94	2,659.39	5,215	400	0.7806
MC	1			4	0.3966	5,560.22	2,641.01	5,215	199	1.5868
ANNE	1			3	0.8555	5,997.95	2,876.50	1,200		
BB		1		4	1.8995	5,455.92	2,811.18	5,215	566	0.5837
DKS			1	3	1.9790	6,140.89	2,564.16	5,215	120	1.6644
WALD				2	2.7259	6,518.75	2,541.77	5,215	56	2.3189
RORO			1	3	1.4277	5,837.08	2,576.96	5,215	373	0.6265
PEST	1			3	3.5188	7,429.34	2,776.76	5,215		
KLOT			1	3	7.5818	10,383.76	3,634.24	5,215	340	0.8869
DKS2				3	7.8815	9,176.73	3,645.66	5,215	32	2.8000
MAN		1		3	4.1813	6,124.14	3,333.74	5,215	261	0.9376
EVKZ				3	3.8280	5,904.65	3,268.67	5,215	364	0.8762
JOHA			1	4	2.4011	5,427.87	2,919.61	2,088	218	0.4689
HE			1	5	3.6305	5,811.50	3,343.25	5,215		
GZW	1			4	8.1933	8,874.82	4,597.43	3,538	168	1.0972
JAH			1	4	5.2381	6,732.89	3,945.26	5,215	619	0.5248
WUST			1	3	6.8566	7,451.84	4,222.19	2,055		
FL			1	3	1.9912	6,146.49	3,200.91	3,524	169	1.0697
VITZ			1	4	2.7463	6,023.21	3,276.62	5,215	266	0.9912
PLAU			1	5	3.7615	7,085.79	3,659.79	5,215	439	0.9171
COTT			1	5	4.2186	7,767.03	3,726.21	5,215	625	0.6322
JAS	1			5	4.2871	7,986.27	3,931.19	2,055		

*Measured in number of classes per grade.

a, Math profile; b, Math core; c, Languages core; d, Math and languages; e, Math and languages and music/arts (private school); g, no profile (private school); CBD, central business district; WESML, weighted exogenous sampling maximum likelihood; SD, standard deviation.

Geographical Analysis

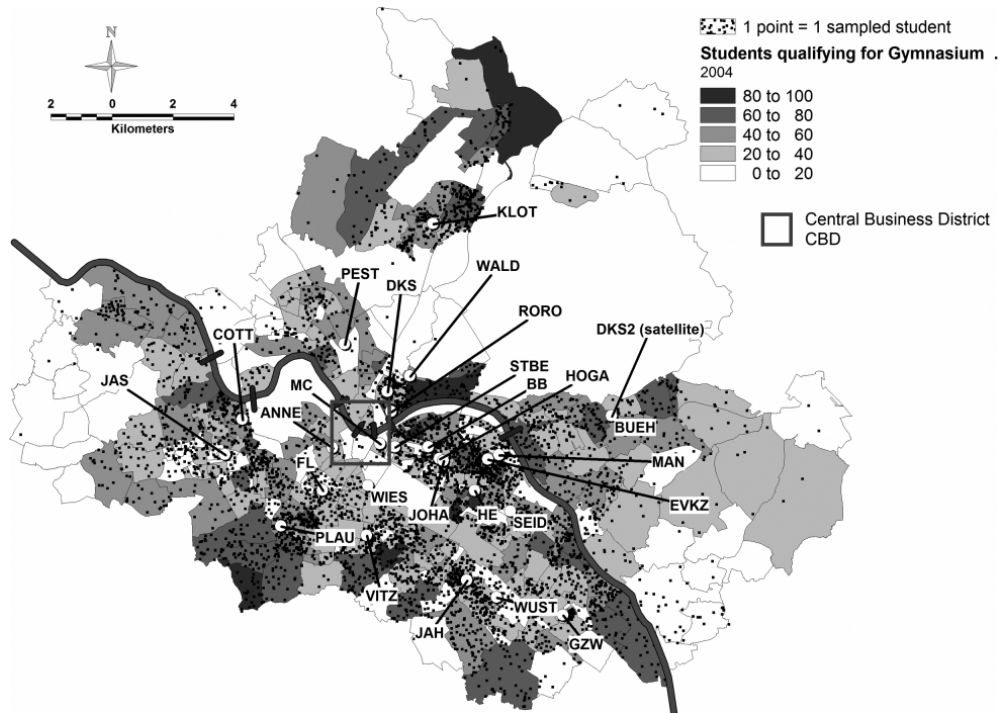


Figure 3. Locations of Gymnasien sampled students, and student numbers in Dresden, 2004.

for by adjusting the systematic utility of an alternative but at the expense of possibly affecting the behavioral underpinnings of the choice model. To operationalize spatial similarity in the deterministic component of utility, one would have to define explanatory variables that describe every kind of relation and spatial dependency that might exist between any pair of school locations. Such a model specification suffers from a remarkable increase in degrees of freedom (= number of coefficients to be estimated) and thus the tractability of the model. Furthermore, the corresponding model might not be consistent with the utility-maximizing theory of MEV models. In our model, spatial dependencies are at least partially incorporated in the stochastic part of utility, which causes correlation of certain alternatives and hence leads to the implementation of the models described in section “Two closed-form discrete choice models with flexible substitution patterns.” Spatial correlation could arise because of various reasons: for example, if two or more schools

- are located along the route of parents taking their children to school during their commute to work;
- are located near a transit stop served by many transit lines;
- use the same (sports) facilities; or
- are located in the same neighborhood or district.⁴

To account for spatial substitution, we group nearby schools into nests (spatial nests). These nests are imposed in order to capture spatial similarity or correlation only. Additional purposes are conceivable, but this would lead to even more complicated (multilevel) nesting structures (Daly and Bierlaire 2006; Müller 2008). For the basic specification, spatially proximate schools have been pooled into one of six nests, depending on the distance between pairs of school locations (Hartigan and Wong 1979). Within the process of specifying the model structure, we found that $K = 6$ results in a reasonable grouping of schools. A large number of nests would probably yield restrictive substitution patterns (i.e., only two or three schools are assumed to share unobserved common attributes). This, in turn, may result in a remarkable number of

insignificant nest parameters. A large number of feasible nesting structures can be found, and other values for K are possible. However, to have more or fewer nests results in fewer or more alternatives per nest, which complicates the finding of similarities between alternatives (a trade-off exists between number of nests and schools per nest). In our study, the ratio nest/alternatives is $6/26 = 0.23$, which is close to the nesting ratio of $3/15 = 0.20$ employed by Berkovec and Rust (1985), for example, who analyze the car choice of households. Similar ratios for nesting structures can be found in Gelhausen (2006) and Bhat (1998), with respective values of $6/21 = 0.28$ and $3/15 = 0.20$.

Finally, the decision about the overall nesting structure is subject to the discretion of a modeler. A researcher can impose an a priori structure. To determine the initial nesting structure shown in Fig. 4a, we employ R 2.10.0 and the *k-Means* function of the package *stats* version 2.10.1.⁵ An alternative approach is to search all possible nesting structures that might result in a large number of distinct structures for even moderate choice sets (Hensher and Button 2000, p. 216). For the estimation of the model coefficients and parameters, we use the public domain software package Biogeme 1.8 (Bierlaire 2003, 2008).

Results

Throughout the model-building process, we found several more or less equivalent specifications (for both V_{ni} and the nesting structure). Table 2 summarizes the estimation results for the MNL, NL, and GNL models. Maranzo and Papola (2008) show that for a given feasible substitution pattern, an infinite number of associated GNL specifications may exist. The nesting structures for the NL and GNL model specifications are as follows. For the NL model,

- Nest 1: ANNE, FL, VITZ, MC
- Nest 2: MAN, EVKZ, HE
- Nest 3: STBE, BB, JOHA
- Nest 4: PLAU, COTT, JAS
- Nest 5: DKS, DKS2, WALD, RORO, PEST, KLOT
- Nest 6: GZW, JAH, WUST,

and for the GNL model,

- Nest 1: ANNE, FL, VITZ, MC, **PLAU**
- Nest 2: MAN, EVKZ, **HE**
- Nest 3: **STBE**, BB, JOHA, **HE**
- Nest 4: **PLAU**, COTT, JAS
- Nest 5: DKS, DKS2, WALD, RORO, PEST, KLOT, **STBE**
- Nest 6: GZW, JAH, WUST.

As can be seen from these nesting structures, the GNL model allows three alternatives to enter two nests (the corresponding short names are in bold). For the NL structure, we found the following relationships: the variable distance between a student's home and the school location is considered to be semialternative-specific instead of generic.⁶ Hence, three different coefficients have been estimated for this variable: one for magnet schools ($\beta_{1,2} = -0.477$), which offer a unique profile, one for private schools ($\beta_{1,3} = -0.454$), and one for all others ($\beta_{1,1} = -0.573$). The corresponding coefficients (1.1–1.3) are significant, as measured by the value of the asymptotic *t*-test. This result indicates that for a two-tailed test, the respective coefficients differ from zero at the frequently used significance level of 0.05 (Ben-Akiva and Lerman 1985, pp. 161–2). Because we expected distance to be nonlinear, we tested selected modifications of the distance variable, such as the log of distance, a power series, and piecewise linearization. However, the linear specification presented in Table 2 yields the best model fit.

These results show that students enrolled in private or magnet schools are less sensitive to distance than others. This finding implies that a trade-off exists between distance traveled to school and the degree of

Geographical Analysis

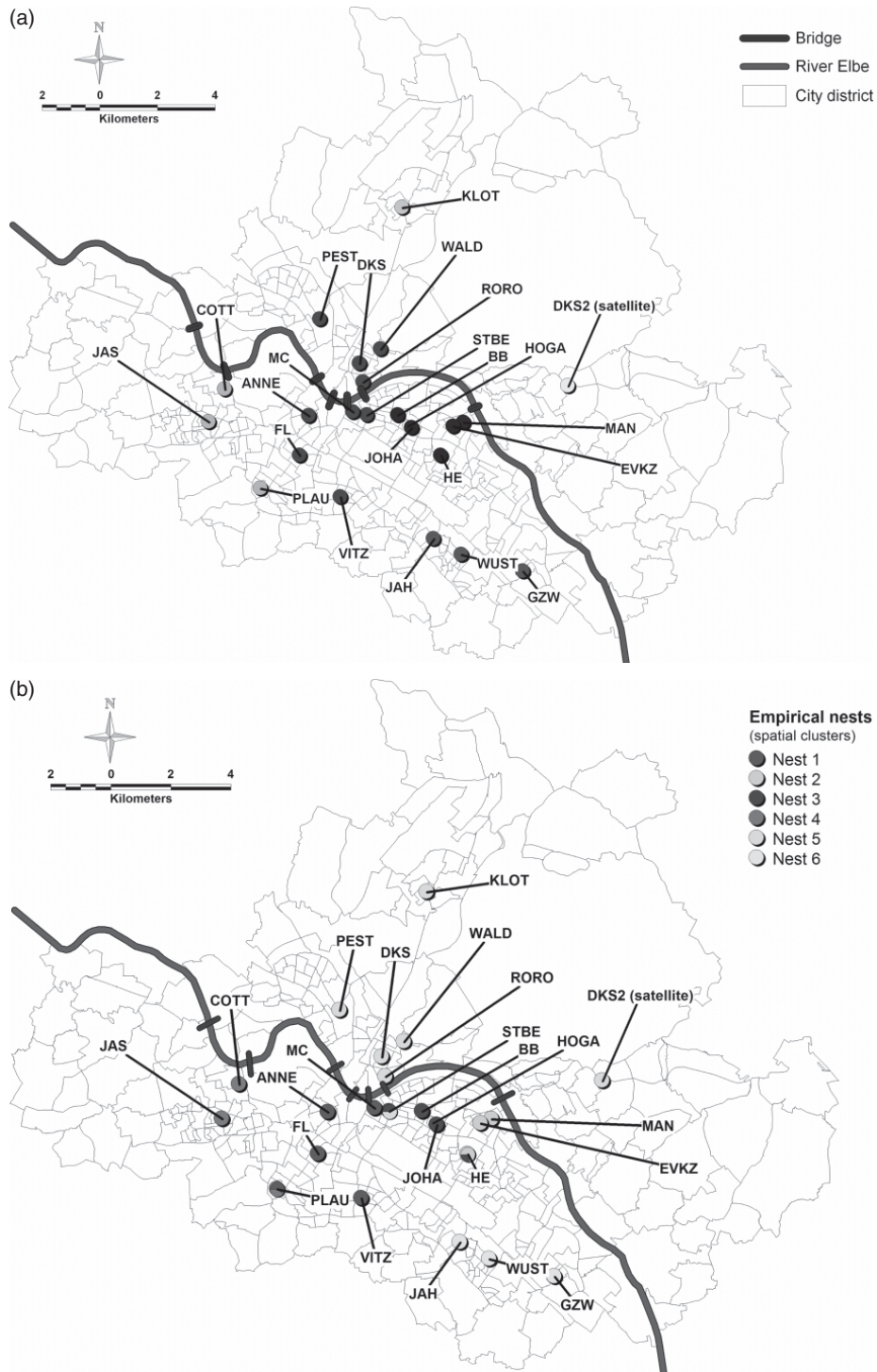


Figure 4. (a) Spatial nests of schools predetermined by cluster analysis. Schools that are allocated to the same nest show the same shading. (b) Empirically determined nests (by the generalized nested logit model). HE, PLAU, and STBE are proportionately allocated to different nests.

Table 2 Estimates for MNL, NL, and GNL Models

Variable number	Description	MNL		NL		GNL	
		Coeff. estimate	t-stat	Coeff. estimate	t-stat	Coeff. estimate	t-stat
Distance variable							
1.1	Distance student-school in km (other schools)	-0.660	-45.44	-0.573	-48.55	-0.571	-48.89
1.2	Distance student-school in km (magnet schools)	-0.544	-28.79	-0.477	-38.33	-0.488	-40.05
1.3	Distance student-school in km (private schools)	-0.465	-23.42	-0.454	-38.65	-0.470	-40.99
Other spatial variables							
2	Distance to school ≤ 1 km*	0.452	7.77	0.165	3.61	0.0703	1.93
3	Distance school-CBD in km	-0.142	-8.42	-0.0322	-2.71	-0.093	-7.28
4	School and student on same side of Elbe*	0.337	4.95	0.554	10.67	0.399	8.54
5	Site location of school's main campus*	-0.789	-2.97	-1.06	-8.54	-0.964	-8.34
6	School size (no. classes per grade)	0.230	10.23	0.273	13.77	0.234	13.80
7	Private school: no profile*	-2.14	-4.63	-1.22	-6.41	-1.03	-7.06
8	Private school: profile math & languages & music/arts*	-1.81	-4.45	-0.720	-4.17	-0.555	-4.50
9	Private school \times avg. neighborhood income in thousand euro	0.951	4.20	0.821	9.75	0.708	11.57
10	Profile math & languages*	0.111	1.58	0.0819	1.59	0.192	3.75
11	Profile math & music/arts*	0.359	5.03	0.386	7.34	0.466	9.54
12	Profile languages core*	0.560	4.52	0.867	9.54	0.775	8.87
13	Profile languages core \times gender male*	-0.983	-7.14	-0.839	-7.03	-0.799	-6.87
14	Profile math core*	0.221	-1.57	0.867	11.72	0.822	11.98
15	Profile math \times gender male*	0.566	3.95	0.112	3.15	0.087	3.23
Nest parameters\ddagger							
$\mu C1$	ANNE, FL, VITZ, MC (PLAU)			1.00	§	1.04	19.82 \ddagger
$\mu C2$	MAN, EVKZ, HE			4.73	10.00 \ddagger	6.29	11.71 \ddagger
$\mu C3$	STBE, BB, JOHA (HE)			1.12	19.58 \ddagger	3.82	7.37 \ddagger
$\mu C4$	PLAU, COTT, JAS			1.76	17.95 \ddagger	4.32	7.54 \ddagger
$\mu C5$	DKS, DKS2, WALD, RORO, PEST, KLOT (STBE)			1.29	26.05 \ddagger	1.29	23.55 \ddagger
$\mu C6$	GZW, JAH, WUST			1.18	13.71 \ddagger	1.16	12.79 \ddagger

Table 2 *Continued*

Variable number	Description	MNL		NL		GNL	
		Coeff. estimate	t-stat	Coeff. estimate	t-stat	Coeff. estimate	t-stat
α 's							
	$\alpha C1_PLAU$					0.599	10.60
	$\alpha C4_PLAU$					0.401	7.10
	$\alpha C3_STBE$					0.410	7.10
	$\alpha C5_STBE$					0.590	10.23
	$\alpha C2_HE$					0.710	30.15
	$\alpha C3_HE$					0.290	12.33
	Number of observations	5,215		5,215		5,215	
	$\mathcal{L}(0)$	-15,233.325		-15,233.325		-15,233.325	
	$\mathcal{L}(\hat{\beta})$	-9,883.091		-9,672.115		-9,572.170	
	\bar{p}^2	0.350		0.364		0.370	

*Dummy variables.

[†]t-test for nest parameters against 1.

[‡]Alternatives in brackets are considered in GNL only.

[§]Nest 1 is fixed due to modeling reasons.

MNL, multinomial logit; NL, nested logit; GNL, generalized nested logit.

specialization.⁷ The models with a generic distance variable and the semialternative-specific distance variable have been tested with a log-likelihood ratio test. The semialternative-specific distance variable outperforms the generic one at a 0.05 significance level. Due to space restrictions, we report the first specification only. Moreover, we consider a dummy variable that indicates whether a school is situated less than 1 km from a student's home. The positive sign of the corresponding coefficient indicates that schools within walking distance are favored. Variable 3 denotes the distance between a school and the central business district (CBD), which is a measure of location of that school. The negative sign of the corresponding coefficient indicates that schools located near the city center are more attractive than others. This measure is a proxy for the accessibility of schools. In Dresden, the transit system has a more or less radial network, and hence, schools near the central node are more accessible than others.⁸ In addition, parents who bring their children to school on their work commute are more likely to go to the CBD or at least pass the CBD. The trade-off between the distance school-CBD and the distance traveled to school supports this interpretation:

$$\frac{\partial U_{ni}/\partial \text{distance school - CBD}}{\partial U_{ni}/\partial \text{distance student - school}} = \frac{\hat{\beta}_3}{\hat{\beta}_1}, \quad (15)$$

which is 0.056 for the NL and 0.163 for the GNL model specification. In the GNL model, each 1-km increase in distance between a school and the CBD tends to be compensated for by a decrease of 0.163 km in the distance traveled by a student to school without affecting the utility of the student. For constant utility, more peripherally located schools have smaller catchment areas than central ones. According to equation (15), 1 km traveled to a central school has higher utility compared with 1 km traveled to a school located in the outskirts. Variables 4 and 5 have not been considered in the first place in the estimation. An outlier analysis points to observations with poorly predicted choice probabilities. First, there was an overprediction for DKS2, which is a site location of DKS's main campus. Without variable 7, the predicted market share for DKS2 was far too high. The negative coefficient of this variable corrects for this misprediction. Furthermore, for some observations, the predicted choice probabilities for nonchosen schools are remarkably high, an issue especially related to schools located on the opposite side of the river Elbe, based on a student's place of residence. Introducing variable 4, corrects the predicted choice probabilities of the nonchosen schools.

The remaining school attributes suggest a number of implications. Larger schools tend to be more attractive than smaller ones (variable 6) because large schools are less likely to deny enrollment based on capacity constraints. The dummy variables 7 and 8 indicate that private schools are less preferred than public schools, because most private schools often are associated with school fees, and some of them additionally have religious affiliation restrictions. For private schools that offer a wide range of profiles, the disutility is less remarkable (variable 8). Nevertheless, the higher the average neighborhood income, the more attractive private schools become (variable 9). The most requested profile is the math core (variable 14), particularly by male students (variables 14 and 15). This preference is followed by the languages core profile if students are female (variable 12). Less attractive are standard profiles (variables 10 and 11) and the languages core profile if students are male (variables 12 and 13).⁹

As expected, the nest parameters μ_{c2} to μ_{c6} are significantly different from one, indicating similarities between nearby schools and the failure of IIA in the standard logit case. Significant nest parameters with values consistent with utility maximization (i.e., if $0 < 1/\mu_k < 1$) are a sufficient indicator that IIA does not hold for an MNL in the school choice context (Train 2003, p. 54). The nesting coefficients presented in Table 2 for both the NL and the GNL models indicate that IIA holds for alternatives that are in the same nest but not for those across nests (Train 2003, p. 82). Nest 1 was fixed for the estimation of the NL model. Although alternative feasible nesting structures other than the one presented for the NL model in Table 2 could have been found, this option was abandoned here in favor of comparability between the NL and GNL models. Both models (NL and GNL) are normalized from the top (i.e., $\mu = 1$). The presumed spatial substitution pattern (Fig. 4a) is empirically confirmed to some extent (Fig. 4b). The strongest spatial

Table 3 Cross-Elasticities with Respect to the *Distance to CBD*-Variable for Nest 3 (GNL) and Two Alternatives from Other Nests*

	HE	BB	STBE	JOHA	ANNE	COTT
HE	-4.738	2.284	1.437	2.691	0.045	0.0133
BB	0.105	-6.285	2.499	6.664	0.045	0.0133
STBE	0.024	0.899	-0.266	1.059	0.045	0.0133
JOHA	0.105	5.657	2.499	-5.217	0.045	0.0133
ANNE	0.0135	0.292	0.810	0.344	-2.780	0.0133
COTT	0.0135	0.292	0.810	0.344	0.045	-17.9581

*School names printed in bold indicate the nest membership.

CBD, central business district; GNL, generalized nested logit.

Table 4 Cross-Elasticities with Respect to the *Distance to CBD*-Variable for Nest 3 (NL) and Three Alternatives from Other Nests*

	HE	BB	STBE	JOHA	ANNE	COTT
HE	-11.667	0.175	0.298	0.179	0.073	0.024
BB	0.066	-3.230	0.443	0.277	0.073	0.024
STBE	0.066	0.271	-2.358	0.277	0.073	0.024
JOHA	0.066	0.271	0.443	-3.206	0.073	0.024
ANNE	0.066	0.175	0.298	0.179	-2.649	0.024
COTT	0.066	0.175	0.298	0.179	0.073	-6.955

*School names printed in bold indicate the nest membership.

CBD, central business district; NL, nested logit.

similarity exists between schools in nest C2 (μ_{C2}). However, particularly near the CBD, the substitution patterns and thus correlation between schools are somewhat more complicated than had been assumed. Instead of having two large nests, as presumed, we empirically find six smaller nests. Moreover, all schools north of the river Elbe are allocated in one nest. Although we account for the separation effect of the river with variable 6, schools north of the river share some unobserved spatial factors. This indicates that spatial similarity exhibits substitution effects that are difficult to account for in the deterministic part of utility (V_{ni}). Comparing the results between the NL and the GNL models, the coefficients of the deterministic utility functions are mostly similar under different error structures. This outcome signifies the reliability of the model specification. However, concerning the small difference in coefficients between the NL and the GNL models, most of the GNL model coefficients are smaller in magnitude than the NL model coefficients.

The relationship among pairs of alternatives in the NL and the GNL models can be examined further by comparing the respective cross-elasticities, or the proportional change in the choice probability of an alternative with respect to a proportional change in an explanatory variable of another alternative (Koppelman and Bhat 2006, p. 50). The elasticity increases between pairs of alternatives as the corresponding value of $1/\mu_k$ decreases from one. The magnitude of this effect is further related to the choice probability of the respective nest and the conditional probability of the alternatives in that nest. This effect can also be seen in Tables 3 and 4, which include the cross-elasticities for alternatives in nest 3 and alternatives outside the nest associated with a change in the distance to the CBD variable.

The elasticity measure for distance to CBD, for example, can be used to evaluate a relocation. A change in the distance to CBD for a given school occurs if that school is relocated for a certain period

of time due to extensive renovation of the original school building. In the GNL model, a 1% change in the respective attribute of alternative BB, for instance, causes a 2.5% change in the choice probability of STBE, which is in the same nest (Table 3). For an alternative outside the nest, like ANNE, the respective change in choice probability is only 0.045%, which is disproportionately small as these alternatives do not share a common nest. Thus, alternatives that share a common nest are much better substitutes for each other than alternatives from different nests. The given values of the elasticities quantify this distinction concerning substitutability. Furthermore, in case of the GNL model, the fraction of each alternative included in one or more common nests determines the implied correlation and substitution between alternatives (Hensher and Button 2000, p. 218). For our analysis, we chose schools allocated to more than one nest that are located nearby schools of a different nest. Hence, STBE, HE, and PLAU are assigned to a second nest, as displayed in Fig. 4b. Several GNL model specifications lead to the feasible, reasonable, and easy-to-interpret model presented in Table 2. STBE, located south of the river, exhibits spatial similarity with schools north of the river. Allocation parameters α_{C3_STBE} and α_{C5_STBE} indicate that STBE is allocated to nest C5 by nearly 60% and to nest C3 by 40%. Thus, STBE shares stronger common unobservable attributes with schools of nest C5 than with schools of nest C3. This is a reasonable finding that may be explained by the bridges across the river Elbe surrounding the area around STBE. We further derive correlation matrices from the NL and the GNL models, which are displayed in Tables A1 and A2, respectively, in the Appendix. The eighth row of Table A2 (STBE) documents the advantage of the GNL model. Because STBE is allocated to nests C3 and C5, STBE is correlated with many more schools than indicated by a simple NL model. Besides this flexibility in substitution patterns, the GNL model yields a higher log-likelihood ($\mathcal{L}[\hat{\beta}]$ in Table 2). We can reject the null hypothesis that the NL and the GNL models are equivalent at the 0.05 level of significance using a nonnested hypothesis test (Ben-Akiva and Lerman 1985, p. 171ff.).

Conclusion

Due to the possibility of free school choice and fluctuating student numbers, schools in Germany face increasing competition, which can be seen particularly in the expanding number of profiles (e.g., math, languages, sciences, arts) or extracurricular activities offered by schools to stimulate enrollment and, thus, avoid school closings. To analyze the effects of changes in the school network, mid- and long-term forecasts of demographic trends as well as of students' decisions about school choice needs to be taken into account to derive possible future scenarios. Therefore, we feel that choice models reproducing the decision-making processes of individuals that are as realistic as possible (i.e., choice models accounting for spatial substitution) are a valuable instrument in school planning and school assignment. Until now, literature about school network planning seems to have ignored spatial substitution between competing school locations. Moreover, school choice literature with a focus on spatial substitution is scarce. The model presented here explicitly accounts for spatial substitution. Fortunately, the model still takes a computationally convenient closed-form. We can verify most of the findings in the literature concerning the variables that enter the systematic part of utility, like school size and travel-to-school distance. Moreover, we find new evidence about spatial effects. First, we see that the catchment area of a school (based on constant utility values) decreases in relation to increased distance from the CBD. Second, our analysis shows that a significant and remarkable correlation exists between schools within proximity to one another. Furthermore, correlation patterns are allowed to vary due to a flexible allocation of schools to nests. From a methodological perspective, more sophisticated approaches are worth using (i.e., discrete choice models based on utility-maximizing behavior) in order to attain more insights into the spatial patterns of locational choice. Through relaxation of the distinct membership of a school to one nest, we incorporate spatially overlapping substitution effects. This analysis strongly suggests that spatial substitution should be focused on more when designing a school network. Accordingly, empirically determined substitution between locations should be accounted for in location-allocation problems and urban models in general.

Acknowledgements

We thank four anonymous referees for their comments on an earlier version of this article. We are also very grateful to the editor for his editing of this article. Of course, we retain responsibility for all remaining errors, omissions, and opinions.

Notes

- 1 This model is similar to the cross-nested logit (CNL) model by Vovsha 1997.
- 2 We do not, however, incorporate the variables of mean test score and religious affiliation in our study due to lack of information. Race does not play an important role within the school choice process in most eastern German cities (except Berlin) because the percentage of students of color is very low ($\leq 10\%$).
- 3 If actual market shares are not known, one can use (under certain conditions) the weighted conditional maximum likelihood (WCML) estimator by Bierlaire, Bolduc, and McFadden (2008).
- 4 If two schools are located in the same neighborhood, as perceived by a student (Guo and Bhat 2007), we expect that they are more correlated to each other than to other schools.
- 5 **R** is a programming language and software environment for statistical computing and graphics.
- 6 A full alternative specific specification yields $I - 1$ coefficients.
- 7 We consider private and magnet schools as specialized schools.
- 8 Nearly 60% of students enrolled at Gymnasium schools in Dresden choose public transport for their commute to school (Müller, Tscharaktschiew, and Haase 2008).
- 9 Profile math is the reference category.

Appendix

Table A1 Correlation Matrix of the NL Model

	MC	ANNE	FL	VITZ	MAN	EVKZ	HE	STBE	BB	JOHA	PLAU	COTT	JAS	DKS	DKS2	WALD	RORO	PEST	KLOT	GZW	JAH	WUST		
MC	1																							
ANNE		1																						
FL			1																					
VITZ				1																				
MAN					1	0.96	0.96																	
EVKZ					0.96	1	0.96																	
HE					0.96	0.96	1																	
STBE								1	0.20	0.20														
BB								0.20	1	0.20														
JOHA								0.20	0.20	1														
PLAU											1	0.68	0.68											
COTT											0.68	1	0.68											
JAS											0.68	0.68	1											
DKS														1	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40
DKS2														0.40	1	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40
WALD														0.40	0.40	1	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40
RORO														0.40	0.40	0.40	1	0.40	0.40	0.40	0.40	0.40	0.40	0.40
PEST														0.40	0.40	0.40	0.40	1	0.40	0.40	0.40	0.40	0.40	0.40
KLOT														0.40	0.40	0.40	0.40	0.40	1	0.40	0.40	0.40	0.40	0.40
GZW														0.40	0.40	0.40	0.40	0.40	0.40	1	0.28	0.28	0.28	0.28
JAH																				0.28	1	0.28	0.28	0.28
WUST																				0.28	0.28	1	0.28	0.28

Table A2 Correlation Matrix of the GNL Model

	MC	ANNE	FL	VITZ	MAN	EVKZ	HE	STBE	BB	JOHA	PLAU	COTT	JAS	DKS	DKS2	WALD	RORO	PEST	KLOT	GZW	JAH	WUST	
MC	1																						
ANNE	0.754	1																					
FL	0.754	0.754	1																				
VITZ	0.754	0.754	0.754	1																			
MAN					1	0.8912	0.794																
EVKZ					0.8912	1	0.794																
HE					0.794	0.794	1	0.2875	0.426	0.426													
STBE							0.2875	1	0.5356	0.5356				0.2997	0.2997	0.2997	0.2997	0.2997	0.2997				
BB							0.426	0.5356	1	0.923													
JOHA							0.426	0.5356	0.923	1													
PLAU											1	0.5349	0.5349										
COTT											0.5349	1	0.9283										
JAS																							
DKS														1	0.3991	0.3991	0.3991	0.3991	0.3991	0.3991			
DKS2														0.3991	1	0.3991	0.3991	0.3991	0.3991	0.3991			
WALD														0.3991	0.3991	1	0.3991	0.3991	0.3991	0.3991			
RORO														0.3991	0.3991	0.3991	1	0.3991	0.3991	0.3991			
PEST														0.3991	0.3991	0.3991	0.3991	1	0.3991	0.3991			
KLOT														0.3991	0.3991	0.3991	0.3991	0.3991	1	0.3991			
GZW														0.3991	0.3991	0.3991	0.3991	0.3991	0.3991	1	0.2568	0.2568	
JAH																					0.2568	1	0.2568
WUST																					0.2568	0.2568	1

References

- Ben-Akiva, M., and S. Lerman. (1985). *Discrete Choice Analysis, Theory and Application to Travel Demand*. Cambridge, MA: MIT Press.
- Berkovec, J., and J. Rust. (1985). "A Nested Logit Model of Automobile Holdings for One Vehicle Households." *Transportation Research Part B: Methodological* 19(4), 275–85.
- Bhat, C. (1998). "Analysis of Travel Mode and Departure Time Choice for Urban Shopping Trips." *Transportation Research Part B: Methodological* 32(6), 361–71.
- Bierlaire, M. (2003). "BIOGEME: A Free Package for the Estimation of Discrete Choice Models." Proceedings of the Swiss Transport Research Conference, Ascona, Switzerland.
- Bierlaire, M. (2006). "A Theoretical Analysis of the Cross-Nested Logit Model." *Annals of Operations Research* 144(1), 287–300.
- Bierlaire, M. (2008). Estimation of Discrete Choice Models with BIOGEME 1.8, biogeme.epfl.ch.
- Bierlaire, M., D. Bolduc, and D. McFadden. (2008). "The Estimation of Generalized Extreme Value Models from Choice-Based Samples." *Transportation Research B: Methodological* 42(4), 381–94.
- Bolduc, D., B. Fortin, and M. Fournier. (1996). "The Effect of Incentive Policies on the Practice Location of Doctors: A Multinomial Probit Analysis." *Journal of Labor Economics* 14(4), 703–32.
- Borgers, A., H. Oppewal, M. Ponjé, and H. Timmermans. (1999). "Assessing the Impact of School Marketing: Conjoint Choice Experiments Incorporating Availability and Substitution Effects." *Environment and Planning A* 31, 1949–64.
- Cullen, J., B. Jacob, and S. Levitt. (2005). "The Impact of School Choice on Student Outcomes: An Analysis of the Chicago Public Schools." *Journal of Public Economics* 89(5–6), 729–60.
- Daly, A., and M. Bierlaire. (2006). "A General and Operational Representation of Generalised Extreme Value Models." *Transportation Research Part B: Methodological* 40(4), 285–305.
- Gelhausen, M. (2006). "Flughafen- und Zugangsverkehrsmittelwahl in Deutschland—Ein verallgemeinerter Nested Logit-Ansatz." Proceedings of the Deutscher Luft- und Raumfahrtkongress, Braunschweig.
- Guo, J. Y., and C. R. Bhat. (2007). "Operationalizing the Concept of Neighborhood: Application to Residential Location Choice Analysis." *Journal of Transport Geography* 15, 31–45.
- Hartigan, J., and M. Wong. (1979). "Algorithm AS 136: A K-Means Clustering Algorithm." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28(1), 100–8.
- Hastings, J., T. Kane, and D. Staiger. (2006). "Parental Preferences and School Competition, Evidence from a Public School Choice Program." National Bureau of Economic Research Working Paper No. 11805.
- Haynes, K., and A. Fotheringham. (1990). "The Impact of Space on the Application of Discrete Choice Models." *The Review of Regional Studies* 20, 39–49.
- Haynes, K., D. Good, and T. Dignan. (1988). "Discrete Spatial Choice and the Axiom of Independence from Irrelevant Alternatives." *Socio-Economic Planning Science* 22(6), 241–51.
- Heiss, F. (2002). "Structural Choice Analysis with Nested Logit Models." *The Stata Journal* 2(3), 227–52.
- Hensher, D., and K. E. Button. (2000). *Handbook of Transport Modelling*. Oxford, U.K: Pergamon.
- Hunt, L., B. Boots, and P. Kanaroglou. (2004). "Spatial Choice Modelling: New Opportunities to Incorporate Space into Substitution Patterns." *Progress in Human Geography* 28, 746–66.
- Jepsen, C., and M. Montgomery. (2009). "Miles to Go before I Learn: The Effect of Travel Distance on the Mature Person's Choice of Community College." *Journal of Urban Economics* 65, 64–73.
- Koppelman, F., and C. Bhat. (2006). A Self-Instructing Course in Mode Choice Modelling: Multinomial and Nested Logit Models. Tech. rep. U.S. Department of Transportation Federal Transit Administration.
- Koppelman, F. S., and V. Sethi. (2000). "Closed-Form Discrete-Choice Models." In *Elsevier's Handbooks in Transport 1*, 211–27, edited by D. A. Hensher and K. J. Button. Vol. Handbook of Transport Modelling. Amsterdam, Netherlands: Pergamon, Chap. 13.
- Lankford, H., E. Lee, and J. Wyckoff. (1995). "An Analysis of Elementary and Secondary School Choice." *Journal of Urban Economics* 38, 236–51.
- Lankford, H., and J. Wyckoff. (2006). "The Effect of School Choice and Residential Location on the Racial Segregation of Students." In *Improving School Accountability—Check-Ups or Choice, Advances in Applied Microeconomics, Volume 14*. 185–239, edited by T. Gronberg and D. Jansen. Albany, NY: Elsevier.
- Luce, R. (1959). *Individual Choice Behaviour: A Theoretical Analysis*. New York: Wiley.
- Manski, C., and D. Wise. (1983). *College Choice in America*. Cambridge, MA: Harvard University Press.
- Maranzo, V., and A. Papola. (2008). "On the Covariance Structure of the Cross-Nested Logit Model." *Transportation Research B: Methodological* 42(2), 83–98.
- McFadden, D. (1975). On Independence Structure and Simultaneity in Transportation Demand Analysis. Working Paper 7511. Berkeley: Department of Economics, University of California.
- McFadden, D. (1978). "Modeling the Choice of Residential Location." *Spatial Interaction Theory and Planning Models* 25(477), 75–96.

Geographical Analysis

- Müller, S. (2008). *Dynamic School Network Planning in Urban Areas*, 5. Stadt- und Raumplanung Stadt- und Raumplanung. Berlin: LIT Verlag.
- Müller, S. (2009). "A Spatial Choice Model Based on Random Utility." *Journal of the Institute of Transport and Economics* 11(2), 2–25.
- Müller, S., K. Haase, and S. Kless. (2009). "A Multi-Period School Location Planning Approach with Free School Choice." *Environment and Planning A* 41(12), 2929–45.
- Müller, S., S. Tscharaktschiew, and K. Haase. (2008). "Travel-to-School Mode Choice Modelling and Patterns of School Choice in Urban Areas." *Journal of Transport Geography* 16(5), 342–57.
- Neu, M. (2007). "Sozialraumstrukturen im Wandel: Eine Längsschnittanalyse des Essener Stadtgebietes 1970-1987-2006." Diskussionspapiere aus der Fakultät für Sozialwissenschaft, Ruhr-Universität Bochum 2007-1.
- Schneider, T. (2004). Der Einfluss des Einkommens der Eltern auf die Schulwahl. Vol. 446. Discussion Papers. Berlin: DIW.
- Timmermans, A., and P. van der Waerden. (1992). "Mother Logit Analysis of Substitution Effects in Consumer Shopping Destination Choice." *Journal of Business Research* 24(2), 177–89.
- Train, K. (1999). "Mixed Logit Models for Recreation Demand." In *Valuing Recreation and the Environment*, 121–40, edited by J. Herriges and C. Kling. Northampton, MA: Edward Elgar.
- Train, K. (2003). *Discrete Choice Methods with Simulation*. Cambridge, MA: Cambridge University Press.
- Vovsha, P. (1997). "Application of Cross-Nested Logit Model to Mode Choice in Tel Aviv, Israel Metropolitan Area." *Transportation Research Record: Journal of the Transportation Research Board* 1607, 6–15.
- Walker, J., and J. Li. (2007). "Latent Lifestyle Preferences and Household Location Decisions." *Journal of Geographical Systems* 9(1), 77–101.

2.2 Students' perceptions, academic departments' image, and major choice in business administration studies - The example of Hamburg Business School.

Students' perceptions, academic departments' image,
and major choice in business administration studies -
The example of Hamburg Business School.

Frauke Seidel*

*Hamburg University, Institute for Transport Economics, Moorweidenstraße 18, D-20148
Hamburg, Germany*

Sven Müller

*Karlsruhe University of Applied Sciences, Department of Transport Systems
Management, Moltkestraße 30, D-76133 Karlsruhe, Germany*

Knut Haase

*Hamburg University, Institute for Transport Economics, Moorweidenstraße 18, D-20148
Hamburg, Germany*

Abstract

In Germany enrollment in majors is of large interest to academic departments, because their budget depends crucially on the number of enrolled students. Hence, understanding students major choice decision is important to academic departments. Besides observed factors such as job opportunities and aptitude for a certain subject, we presume that unobserved latent variables significantly influence the major choice decision of students. Using data that we collected in 2013 among students pursuing a bachelors degree in business administration at Hamburg University, we employ an integrated choice and latent variable model. Thereby, we model the influence of nine major-specific latent variables that we label *image* on the major choice decision of the sampled students. To identify the latent variables, we utilize

*Corresponding author

Email addresses: frauke.seidel@uni-hamburg.de (Frauke Seidel),
sven.mueller@hs-karlsruhe.de (Sven Müller), knut.haase@uni-hamburg.de (Knut Haase)

ordered categorical indicators that we obtained from the assessment of various major-specific psychometric factors such as the students perceptions of *supervision quality* and *achievements in research*. Our findings reveal that latent variables *image* have a significant influence on the major choice decision. We further show that the consideration of psychometric factors alone, detached from an ICLV modeling approach, may result in erroneous decisions. By investigating the distributions of the image variables, we identify their differing impacts on the utilities of the given majors. Based on the results of our analysis, we examine factors that lead to the identification of the latent variables and provide insights on how academic departments at Hamburg Business School can raise their attractiveness from the students perspective.

Keywords: behavioral economics, discrete choice modeling, major choice, hybrid choice models

1. Introduction

Choosing a major at university is, after deciding on a course of study, a decision that importantly influences the career path of an undergraduate student. Within the course of study, the choice of an undergraduate major lays the foundation for the profession students may pursue in the future. When making this decision, students take into account various factors including expected earnings (Berger, 1988), interest in the subject (Malgwi et al., 2005) as well as the likelihood of graduation (Montmarquette et al., 2002).

Majors are areas of specialization in which the academic departments are grouped to provide a variety of teaching contents. At Hamburg Business School, students that are enrolled in either one of the degree programs *Business Administration*, *Business Engineering* or *Business Information Systems* choose one out of nine majors offered by the respective academic department for their third year of undergraduate studies: Finance & Insurance, Healthcare Management, Marketing & Media, Operations & Supply Chain Management, Business Law, Statistics, Strategic Management, Information Management, or Auditing & Taxes. These majors compete for constantly high student enrollment while, at the same time, aiming to attract the most smart and dedicated students, because a permanently high number of students results in more full professor positions or positions being renewed after retirement. Another aspect is, that high student numbers increase the proba-

bility of discovering promising academic offspring. We suppose that the most important reason that academic departments strive for constantly high student numbers lies in the allocation of departmental budgets by the business school. At Hamburg Business School, these budgets depend to a large part on the number of students enrolled in the respective major. With regard to the allocation of departmental budgets, we presume that the knowledge of unobserved and observed influences on students major choice decisions as well as their direction of impact can be a very powerful tool for academic departments. We suppose that this knowledge can be of particular value since it might enable departments to steer students towards their education offers actively. Thus, understanding the dynamics of major choice facilitates the adjustment of certain factors that influence the attractiveness of the respective majors as perceived by the students. So far, the influences on the major choice decision of students have been studied in many different settings. [Leppel et al. \(2001\)](#), for example, investigate how socioeconomic status and parental occupation affects the major choice decision of students by applying a multinomial logit (MNL) model to data collected from beginning postsecondary students in 1990. The authors find that particularly business students are influenced in their major choice decision by their parents' occupations and socioeconomic status while effects differ by gender. For example, both female and male students were more likely to choose male dominated careers provided by a major in science or engineering when their fathers were in professional or executive positions. With mothers in professional or executive positions daughters were more likely to choose non-traditional majors while sons had a higher probability of choosing female-dominated majors like education, health or social sciences. Surprisingly, the study also found that a father in a professional or executive occupation has a larger effect on females major choice than a mother in the same occupation. The opposite was found to be true for male students. From a survey among undergraduate business students at 40 US universities [Kim et al. \(2002\)](#) find that the choice of a major is mainly influenced by interest in a certain subject, job opportunities, opportunities for self employment, abilities and the level of potential earnings after graduation. The importance of each factor on the major choice decision thereby varies depending on the particular major. Furthermore, the authors found factors such as reputation of the major at school, perceived quality of instruction, amount and type of promotional information, and parents and friends influence to be less important to the students major choice decision. By examining students' test scores in a quantitative outcomes assessment

Pritchard et al. (2004) show that students who possess better quantitative skills favor the choice of more quantitative majors like accounting/finance while students with weaker computational and algebra skills tend to choose management or marketing majors. Furthermore, Malgwi et al. (2005) reveal differences in the ranking of influential factors on the major choice decision between women and men. They find that interest in the subject is the most influential factor regardless of gender. For women, the second most important factor is aptitude for the subject while for men it is career advancement and job opportunities. Zafar (2013) shows that gender differences in the major choice decision are complex and can be attributed to differing preferences and tastes rather than women not being confident about their academic abilities. By applying a choice model to subjective expectations data the author demonstrates how to infer decision rules under uncertainty and, in particular, when expectations differ across groups in unknown ways. A more recent study investigates how students beliefs about future earnings and lifestyle influence the major choice decision. Wiswall and Zafar (2015) examine data from undergraduate college students of New York University on their self beliefs about certain major-specific aspects. The authors utilize a structural lifecycle utility model on students beliefs about future earnings as well as ability perceptions and discover that both play an important role in the choice of a major.

Literature provides evidence that expected observable factors have a considerable effect on the major choice decision of students while effects of unobserved latent factors are disregarded so far. However, according to research in the social sciences attitudes, perceptions, norms, and beliefs play an important role in individual decision making and can even override the influence of observable variables on individual choice behavior (Vij and Walker, 2016). The theoretical framework for modeling disaggregate choice behavior is provided by random utility theory. An overview on the development of analyzing economic choices is given by McFadden (2001) in his Nobel lecture. As shown by Marschak (1960) a choice model that is derived under the assumption that an individual maximizes its personal utility is called a random utility model (RUM). Discrete choice models belong to this category of models (McFadden, 1974). They directly link observed attributes of alternatives and characteristics of individuals to observed choices. Utility, which is a random variable from the researchers' perspective, is thereby used to explain the individual choice decisions. Thus, all influences on the choice decision are then supposed to be captured by the model. This assumption is, among

behavioral scientists in particular, seen as quite restrictive since individual decision making behavior is inadequately represented by traditional RUMs. According to [Zeid \(2009, p. 29\)](#) the decision-making process in models that are based on random utility theory is more likely to refer to a 'black box' where the role of psychological factors such as perceptions, attitudes and beliefs is disregarded. Furthermore, the accuracy of the rationality assumption that these models are based upon has been doubted by recent studies from social scientists and behavioral economists ([McFadden et al., 1999](#); [Loomes and Pogrebna, 2016](#); [Ariely, 2008](#), pp. 240-244). To overcome these restrictive assumptions, approaches have been developed that allow for the incorporation of psychometric factors into discrete choice models. A first prototype method for the integration of psychometric data on perceptions and tastes and discrete responses was proposed by [McFadden \(1986\)](#). These so-called integrated choice and latent variable (ICLV) models reveal the effects of attitudes and perceptions on individual utility through latent variables while taking advantage of the simultaneous estimation of the ICLV's structural and measurement models to obtain efficient and consistent estimates. Although latent variables cannot be measured directly, their effects on certain measurable variables (i.e., indicators) are observable ([Ben-Akiva et al., 2002a,b](#); [Walker and Ben-Akiva, 2002](#); [Glerum et al., 2014](#); [Voleti et al., 2016](#)).

We contribute to existing research by investigating the impact of unobserved influences on the major choice decision of students at Hamburg Business School. Therefore, we employ an ICLV model with nine alternative-specific latent variables to study the influence of latent factors on students major choice decision. Our work is based on the assumption that these latent variables reflect the perceived *image* students have of each major. Besides one relatively short meeting where majors are introduced to the undergraduates who are about to make their choice decision, students at Hamburg Business School receive only little official information about the available majors. Thus, we assume that the perceived *image* of a certain major is shaped to a large part by grapevine that circulates among students. To identify these latent *image* variables, we utilize ordinal responses from students with regard to various major-specific perceptual attributes. These responses serve as indicators for the unobserved latent variables and represent the students perceptions towards each major. Examples for such indicators are the perceived requirements to pass a course of a certain major, the perceived quality of supervision and the perceived practical relevance of a course. We further presume, that the major-specific *image* forms as a response to information

about and experience with each particular major (Schweitzer and Cachon, 2000). Therefore, the aim of this study is to examine how unobserved latent factors, besides observed ones, affect the students major choice decision. Our investigation is particularly based on how students' differing perceptions towards available majors influence the outcome of the decision process. Based on the modeling results, we further examine the possibilities of how majors can raise the attractiveness of their teaching offers by adjusting certain factors that are important to students.

We organize this paper as follows: the general modeling framework for the ICLV model that we apply in the remainder of this article is presented in Section 2. Section 3 gives an overview of the outline of the survey and the data. Section 4 provides the specification of the major choice ICLV model that is used for estimation while results and implications for academic departments are discussed in Section 5 followed by a conclusion in Section 6.

2. Integrated Choice and Latent Variable Modeling Framework

An ICLV generally consists of two sub-models: a discrete choice model and a latent variable (LV) model. For each sub-model we specify structural and measurement equations that reflect the formal relationship between choice and measurement indicators on the one hand and observed attributes, observed characteristics, and unobserved latent variables on the other hand. The structural equations for the discrete choice sub-model, are defined according to random utility theory (McFadden, 1974). The utility an individual n perceives from choosing an alternative m (major) is given by:

$$u_{nm} = v_{nm} + \varepsilon_{nm}, \quad \varepsilon_{nm} \sim D^\varepsilon(0, \Sigma_\varepsilon). \quad (1)$$

Utility u_{nm} has been decomposed into a deterministic part v_{nm} and a stochastic part ε_{nm} . In the ICLV model the deterministic part v_{nm} of the choice model is a function $d(\cdot)$ of observed attributes x_{nm} describing the alternatives, individual characteristics s_n , alternative-specific latent variables z_{nm}^* and parameters β and λ :

$$v_{nm} = d(x_{nm}, s_n, \beta, z_{nm}^*, \lambda). \quad (2)$$

In general, parameters β capture the influence of observed attributes x_{nm} and characteristics s_n on utility u_{nm} while parameters λ capture the influence of unobserved latent variables z_{nm}^* on utility. The stochastic part of utility ε_{nm}

captures all influences on utility that are not accounted for within the specification of v_{nm} . For the disturbance term ε_{nm} , the researcher specifies the distribution function D^ε with covariances Σ_ε according to her assumptions about the underlying choice behavior.

The discrete choice model gives the distribution f_1 of utility u_{nm} conditional on the values of the observed attributes x_{nm} and characteristics s_n , and the latent variables z_{nm}^* :

$$f_1(u_{nm}|x_{nm}, s_n, \beta, z_{nm}^*, \lambda, \Sigma_\varepsilon). \quad (3)$$

For more detailed information on random utility models we refer to [Ben-Akiva and Lerman \(1985\)](#) and [Train \(2009\)](#).

In a RUM, the highest utility value u_{nm} of an alternative m within individual's n choice set C_n determines the choice decision of individual n . Since values of utility u_{nm} are not directly observable it is considered a latent variable. However, following the theoretical framework of latent variable models, we assume that utility affects the values of response variables: the observed choice indicators. Thus, we define the measurement equations of the discrete choice model part of the ICLV model as:

$$y_{nm} = \begin{cases} 1, & u_{nm} = \max_{j \in C_n} u_{nj}, \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

Thereby, y_{nm} denotes the choice indicator. According to these specifications the probability of the observed choice is

$$P(y_{nm}|x_{nm}, s_n, \beta, z_{nm}^*, \lambda, \Sigma_\varepsilon). \quad (5)$$

The general formulation of the structural equations for the latent variable (LV) sub-model, is given by:

$$z_{nm}^* = g(x_{nm}, s_n, \gamma) + \omega_{nm}, \quad \omega_{nm} \sim D^\omega(0, \Sigma_\omega). \quad (6)$$

The latent variable z_{nm}^* is a function $g(\cdot)$ of alternative-specific attributes x_{nm} , individual-specific characteristics s_n , and parameters γ . The distribution D^ω of the disturbance term ω_{nm} has to be specified by the researcher. From the structural equations of the LV sub-model we obtain the distribution f_2 of the latent variables z_{nm}^* :

$$f_2(z_{nm}^*|x_{nm}, s_n, \gamma, \Sigma_\omega). \quad (7)$$

The measurement equations of the LV sub-model provide the links between observed indicators and latent constructs (Kenny et al., 1998). Again, we cannot directly observe the values of the latent variables z_{nm}^* but we assume that they affect the values of observed indicators. Values of the observed indicators are captured at individual level by $q = 1, \dots, Q$ survey questions that gather the assessments of psychometric factors on a Likert scale (Likert, 1932). These psychometric factors are thereby defined according to hypotheses that the researcher has about the links between latent variables and the according indicators.

Eventually, the values of observed survey responses to question q with regard to a certain psychometric factor, provide discrete and ordered indicator values i_{nmq} . Within the ICLV modeling framework they are modeled as dependent variables:

$$i_{nmq} = h(z_{nm}^*, \alpha) + \nu_{nmq}, \quad \nu_{nmq} \sim D^\nu(0, \Sigma_\nu). \quad (8)$$

Observed indicators i_{nmq} are a function $h(\cdot)$ of latent explanatory variables z_{nm}^* and a parameter vector α . An error component ν_{nmq} is added to $h(\cdot)$. It follows distribution function D^ν that has to be further specified by the researcher. The measurement equations of the LV sub-model give the distribution f_3 of the indicators conditional on the values of the latent variables z_{nm}^* :

$$f_3(i_{nmq} | z_{nm}^*, \alpha, \Sigma_\nu) \quad (9)$$

Figure 1 provides an overview of the described model components of the ICLV model as well as the previously defined relationships between observed explanatory and unobserved latent variables. Measurement equations represent the links

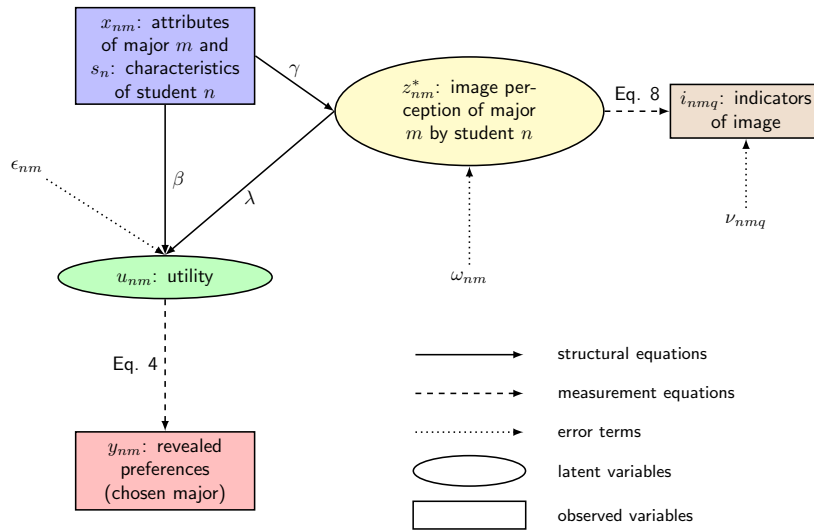


Figure 1: General framework of the integrated choice and latent variable model (illustration based on Walker (2001, p. 90)).

between the latent variables u_{nm} and z_{nm}^* and the observed choice indicators y_{nm} and psychometric factors i_{nmq} . Indicator values are manifestations of latent constructs that help to identify the latent variables but do not affect behavior (Walker, 2001, p. 89). Structural equations provide the links between the observed explanatory variables and the latent variables. These links are also denoted as cause-and-effect relationships that govern the decision making process (Ben-Akiva et al., 2002b). For estimating the ICLV model the joint likelihood function of all model components has to be determined and solved by simulated integration (i.e., Maximum Simulated Likelihood (MSL) estimation). This is necessary as numerical integration methods for models with more than one latent variable quickly become infeasible (Walker, 2001, p. 12).

The joint likelihood function is, if more than one latent variable is included in the model formulation, a multi-dimensional integral of the choice model over the distribution of the latent constructs. Assuming that the disturbances

ε_{nm} , ω_{nm} and ν_{nm} are independent, the joint probability of choosing alternative m while observing indicator i_{nmq} , conditional on the exogenous variables x_{nm} and s_n is given as:

$$\mathcal{L}_n(y_{nm}, i_{nmq} | x_{nm}, s_n, \alpha, \beta, \gamma, \lambda, \Sigma_\varepsilon, \Sigma_\omega, \Sigma_\nu) = \int_{z_{nm}^*} P(y_{nm} | x_{nm}, s_n, \beta, z_{nm}^*, \lambda, \Sigma_\varepsilon) f_3(i_{nmq} | z_{nm}^*, y_{nm}, \alpha, \Sigma_\nu) f_2(z_{nm}^* | s_n, \beta, \Sigma_\omega) dz_{nm}^*. \quad (10)$$

The first term of equation 10 corresponds to the likelihood of the discrete choice model (5). The second term corresponds to the measurement equations of the latent variable model (9) and the third term denotes the distribution of the latent variables as defined by the formulation of structural equations (7).

By jointly estimating the structural and measurement equations of the described ICLV framework, we obtain a fully efficient estimator (Kamakura et al., 1994; Frischknecht et al., 2014). Therefore, Equation 10, that represents the according joint likelihood function, is maximized by applying simulation methods (Bierlaire and Fetiariison, 2009). In contrast to sequential estimation, this simultaneous estimation approach leads to efficient and consistent parameter estimates (Walker, 2001, p. 95).

3. Data

We apply the ICLV modeling framework to data collected at Hamburg Business School to reveal the influence of unobserved latent factors on the major choice decision of students.

We carried out a survey among students enrolled in the undergraduate degree programs Business Administration, Business Engineering, or Information Systems during summer term 2013. The population of interest for our study consists of students in their 4th semester since at this time, students actually decide on a major for their third year (= semesters 5 and 6) of studies. From our survey, we have eventually obtained responses from 377 students. The number of respondents is made up of 301 students enrolled in the Business Administration degree program, followed by 56 Business Engineering students and one student enrolled in Information Systems. Another 19 students chose not to provide information about their degree program. The questionnaire consisted of three parts that were relevant for our study.

First, we obtained student-specific information such as gender, age, degree program, whether the questioned students friends enrolled in the same major, previous apprenticeships, study related internships, and A-level grades. Of course, more detailed information about the students (and their parents) would be interesting. Such information is very difficult to obtain in Germany, though. Our sample consists 55.1% male and 44.9% female students. The largest group of students with 49.2% is between 22 and 25 years old. Another 33.5% of students lie within the age group of 19-21 and only 13.8% are older than 25. Table 1 gives an overview of the relative sample frequencies regarding the respondents' degree program, gender and age. We do not report

Respondents	Sample
<i>Degree Program</i>	
Business Administration	80.0 %
Business Engineering	14.9 %
Information Systems	0.3 %
Unknown	5.0 %
<i>Gender</i>	
Female	44.9 %
Male	55.1 %
<i>Age</i>	
19 - 21	33.5 %
22 - 25	49.2 %
> 25	13.8 %
Unknown	3.5 %

Table 1: Summary statistics of the sampled students. The sample size is 377 out of a total of 383 enrolled students in 2013.

the remaining individual-specific characteristics here, since they are not included in the final model specification that we present in Section 4. The choice set that was utilized in the survey includes the following alternatives:

1. Finance & Insurance (FI)
2. Healthcare Management (HM)
3. Marketing & Media (MM)
4. Operations & Supply Chain Management (OM)
5. Business Law (BL)
6. Statistics & Econometrics (ST)

7. Strategic Management (SM)
8. Information Management (IM)
9. Auditing & Taxes (AT)

In the second part of the survey, we obtained stated choices from the students by conducting a choice experiment. For each student, the choice experiment consisted of multiple choice situations (i.e., repetitions). Per repetition, we displayed two majors with relevant attributes, and the no-choice option. We chose the following attributes to describe the major: *expected income after examination*, *average grades given in all courses of a major*, and *number of courses available per term*. We assumed these attributes within the choice experiment as they reflect the most important influencing factors known to the students at Hamburg Business School. Income and career opportunities are found to play a significant role in major choice decisions (Kim et al., 2002; Malgwi et al., 2005). We suppose the majority of students aims on achieving good grades during their studies to increase their chances for certain job opportunities. Average grades provide students a (rough) estimate for their own grades. By the number of courses available per term, we account for a factor that academic departments can adjust to increase their own attractiveness among students. We assume that more courses increase the attractiveness because the number of courses offered by a major determines the menu of lectures, classes, and seminars students can choose from. The attributes and their levels are shown in Table 2.

Attributes	Attribute Levels		
Expected income after examination in € (x_m^{income})	45,000	55,000	65,000
Average grade received per major (x_m^{grade})	1.3	2.3	3.3
Number of courses offered per term (x_m^{course})	4	6	-

Table 2: Attributes and attribute levels for choice experiment.

With regard to the *grades* variable, it should be noted that smaller values reflect better achievements at German universities. A 1.0 is given for a very good performance while a 4.0 means a course was just passed. Overall the following gradations are possible: 1.0, 1.3, 1.7, 2.0, 2.3, 2.7, . . . 4.0. Beyond 4.0 no further gradations exist since a worse performance automatically results in a failed course or test. From our experience, we know that 3.3 is

a rather bad and seldom average grade. Hence, we decided to consider the interval [1.3, 3.3]. To the best of our knowledge, this interval covers the most usual grades.

The values for the expected income after graduation were chosen according to an annual report on the salaries for graduates from various academic areas that is provided by the German job board StepStone (StepStone, 2016). To limit the number of repetitions per student in the choice experiment we choose a D-efficient design. That way, we ensure to obtain treatment combinations such that the variance of the obtained parameter estimates is minimized (Kuhfeld et al., 1994). We further assume that choice situations and attribute levels are independent and generate the according combinations of alternatives and attribute levels by applying the R package 'AlgDesign' (Wheeler, 2014). As a result we obtained a D-efficient design with 18 different treatment combinations for the experiment. Since this is a large number of repetitions, we decided to divide our sample population into two groups. Hence, within each group respondents provide stated choice decisions for nine repeated choice situations. This results in 3392 observations in our dataset. For the third part of our survey, we consider alternative-specific latent variables that capture the *image* of each major. In the context of our major choice ICLV model, the required response variables are alternative-specific indicators. We utilize a set of psychometric factors q that represent the students perceptions towards each major. We gather the response variables that are needed for the measurement part of the LV sub-model by asking students to assess $q = 1, \dots, 9$ psychometric factors that are listed in the following. Possible answers for each assessment were given on a five point Likert scale.

q=1 Career opportunities after examination: Besides personal interest and skills we assume that choice of a certain major is influenced by the perceived career opportunities a student expects from graduating in a certain major. Thus, we asked students to assess the perceived career opportunities for each major. Answers range from: *very high, high, moderate, low* and *very low*.

q=2 Potential earnings after examination: In line with career opportunities we assume that the expected potential earnings after graduating in a certain major influence the image of a certain major. Thus, we asked students about their expected annual gross income in occupational fields that are typical for the chosen major. The given answers were: $< 35,000\text{€}$, $35,001\text{€}-45,000\text{€}$, $45,001\text{€}-55,000\text{€}$, $55,001\text{€}-65,000\text{€}$ and $> 65,000\text{€}$.

q=3 Internationality of courses: We assume the perceived internationality of courses and lectures is important to students. Internationality of a major is thereby defined by the availability of lectures in English and the possibility to receive credits for courses taken during semesters abroad. Answer possibilities are: *very high, high, moderate, low* and *very low*.

q=4 Practical relevance of courses: This question aims at assessing the students perceptions regarding the skills they expect to learn when choosing a certain major and whether the taught knowledge has any practical relevance in a later job. Thus, we asked students to assess the practicality of courses offered by each major with answers ranging from: *very high, high, moderate, low* and *very low*.

q=5 Ability to cope with course requirements: During the first two years of their studies students attend lectures and courses from academic departments that belong to different majors. We suppose that students have an idea on how well they will be coping when enrolling in a certain major. Furthermore, we assume that an ability to better cope with course requirements leads to higher image values of the respective major. Again, this perception was assessed by asking the students to select one of the following answer possibilities per major: *very good, good, moderate, bad* and *very bad*.

q=6 Achievements in research: We assume that students have a certain perception regarding research activities within the majors (i.e., the respective academic department). These activities include for example the number of publications and the acquisition of external funding. By this question we assessed the perceived achievements in research for each major by asking students to select one of the following answer possibilities per major: *very high, high, moderate, low* and *very low*.

q=7 Quality of supervision: We asked students to evaluate the perceived quality of supervision for each major on a Likert scale with possible assessments: *very high, high, moderate, low* and *very low*. In particular, we aimed to assess the perceived supervision quality when the student is writing a Bachelor thesis or requires consultation regarding lecture related topics. Thereby, we assume that a high perceived quality of supervision increases the image and hence the utility of choosing the respective major.

q=8: Variability in lectures: Within each major the number of choices regarding modules and lectures vary. Thus, the possibilities to combine certain lectures for gathering the required credit points are more or less restrictive. We assume that a high variability increases the perceived attractiveness (image) of a major. We asked students to state the perceived variability in lectures for each major by choosing one of the following answers: *very high*, *high*, *moderate*, *low* and *very low*.

q=9 Amount of work required per course: Besides the students' perception regarding their ability to cope with course requirements, we also want them to evaluate the amount of work that they expect to be required to pass a course in a certain major. Again, we assume experiences from lectures in the early stages of the students studies favors the development of perceptions. We hypothesize that the amount of course work is negatively related to the image of a certain major. The answer possibilities we provided for the assessment of the amount of course work per major were: *very high*, *high*, *moderate*, *low* and *very low*.

We then translated the Likert scale values that students selected for their assessments into the according numerical values $l_k = \{1, \dots, 5\}$. Thereby, one refers to the most negative and five to the most positive assessment of an item. These numerical values represent the categorical and ordered levels of indicators i_{nmq} that are manifestations of the unobserved latent variables *image* z_{nm}^* . We provide a complete overview of the distribution of these assessments per question q and major m in Figure 2.

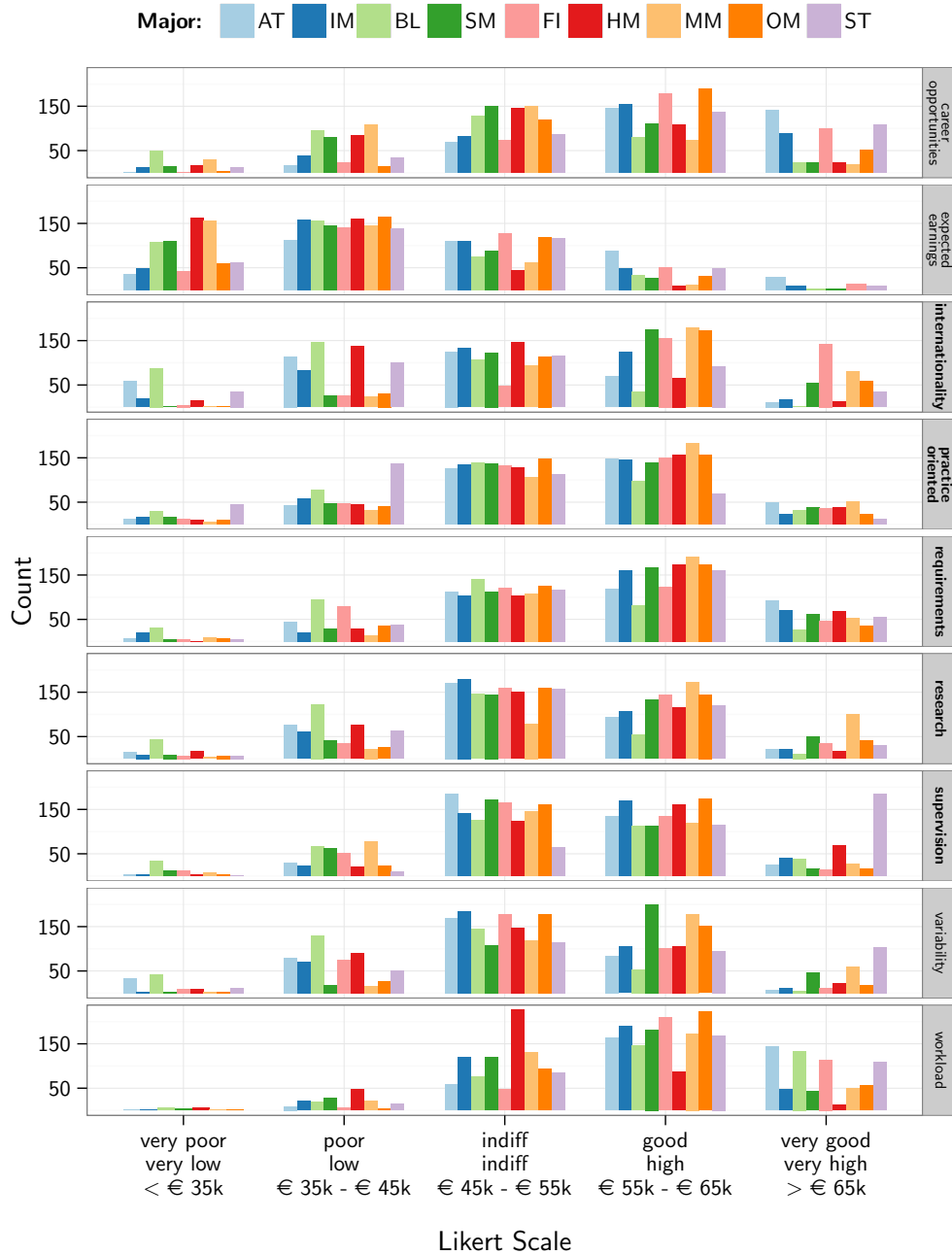


Figure 2: Distribution of the responses to the psychometric factor questions per major. Psychometric factors that are utilized within the measurement equations of the ICLV model are printed in bold.

The graphic shows how the students' answers regarding the psychometric factors q are distributed among the indicator levels l_k . It is apparent that the majority of students evaluates most of the factors with moderate/medium and good/high which translates into indicator values three and four respectively. An exception is the perception for expected earnings after graduation that most students assess to be less than € 45k. We think that the outcome of this assessment results from the fact that we asked for the expected earnings just after graduation. Obviously, the majority of students expects rather low earnings initially after graduation while our study does not provide insights about students expectations on earnings growth rates. Particularly for the HM and MM major the majority of students seems to have the perception of very low or low earnings when graduating in these fields. However, according to Berger (1988), students do not base their major choice decision on starting salaries but on the present value of the predicted future earnings stream. Thus, the probability of choosing one particular major increases with an increasing present value of the future earnings relative to other available majors.

In the long run, the evaluation of perceived career opportunities per major provides a more differentiated view. Here, we assume that students used the perceived salary development as one important influencing factor for their assessment. There are certain majors that are perceived by the majority of students as offering good or very good career opportunities. These include especially the AT, ST, FI, BI, and OM majors. This seems reasonable, since majors with an approach that focuses on mathematical, analytical or IT skills prepare students for higher paid jobs in IT, finance and consulting such as software engineer, IT consultant, Data Analyst, Data Scientist, investment or financial advisor.

There are some majors that clearly stand out for some of the evaluated factors. In particular, the majority of students has the perception that the FI major provides very international courses and lectures. Furthermore, the majority of the students perceives the ST major to have the best quality of supervision, a very high variability in lectures and a very high workload. The ST major also stands out for being the major with the lowest perceived practice orientation. Surprisingly, ST is perceived to offer students quite good career opportunities. From Figure 2 we can also draw conclusions about the reasons for the termination of the BL major. Small student numbers may be attributed to perceived high workloads at moderate career opportunities, poor research activities and internationality.

4. Specification of the major choice ICLV model

For the specification of the structural equations of the discrete choice sub-model we utilize the general formulation of Equation 1 as

$$u_{nm} = \beta_m^{asc} + \beta^{income} \cdot \frac{x_m^{income}}{10,000} + \beta^{grade} \cdot x_m^{grade} + \lambda^{image} \cdot z_{nm}^* + \varepsilon_{nm}. \quad (11)$$

We assume that the utility u_{nm} a student n receives when choosing major m is affected by the observed explanatory variables *income* and *grade* and by unobserved alternative-specific latent variables *image* z_{nm}^* . We do not consider the variable for *variability in courses* here, since estimation results neither showed significant influence on the utility of the given majors nor improvement model fit. The *income* variable is divided by 10,000 to avoid scaling issues in the model estimation. The impact of these variables on the alternative-specific utility value is captured by the unknown parameters β^{income} , β^{grade} and λ^{image} . We include an alternative-specific constant β_m^{asc} . Each has generic influence on utility u_{nm} . We further assume that ε_{nm} is independently and identically (iid) extreme value (EV) distributed. Hence, we obtain from Equation 5 the multinomial logit model (MNL) as the choice model with the logit choice probability (McFadden, 1974)

$$P_{nm} = \frac{e^{v_{nm}}}{\sum_j e^{v_{nj}}}. \quad (12)$$

We specify the structural equations of the latent variable sub-model (6) as follows:

$$z_{nm}^* = \gamma_m^{mean} + \gamma_m^{gender} \cdot s_n^{gender} + \omega_{nm}. \quad (13)$$

We presume that each latent variable z_{nm}^* is a function of the individual-specific explanatory variable *gender*=1, if male, parameters γ_m^{mean} and γ_m^{gender} , and a normally distributed disturbance term ω_{nm} with mean 0 and standard deviation σ_ω . The measurement equations of the LV sub-model (8) that link the latent variables to observed values of indicators are formally defined as:

$$i_{nmq} = \alpha_{mq}^{image} \cdot z_{nm}^* + \alpha_{mq}^{choice} \cdot y_{nm} + \nu_{nmq}. \quad (14)$$

Parameter α_{mq}^{image} captures the influence of the unobserved latent variables z_{nm}^* on the distribution of the observed indicator values. We further include a parameter α_{mq}^{choice} to capture a systematic response bias (Walker, 2001, p. 90). Thereby, we control for exaggerated responses to the psychometric factors survey questions that result in the major-specific indicator values i_{nmq} . Hence, we correct for the response bias when student n prefers major m by linking parameter α_{mq}^{choice} to the according choice indicator y_{nm} . Following Hoyos et al. (2015) we assume the error component ν_{nm} to be logistically distributed with mean 0 and scale 1.

For indicators that are collected in form of a Likert scale, each observed categorical response i_{nmq} is related to the latent variable z_{nm}^* through a threshold model. For a discrete and ordered indicator with $k = 1, \dots, K$ levels $l_1 < l_2 < \dots < l_{k-1} < l_K$, the q -th measurement equation for individual n and alternative m is described by an ordered logit (OL) model (Train, 2009, pp. 163-166). In the ICLV major choice model, the formal relation between the values of latent variables *image* z_{nm}^* and the five indicator values is given as:

$$i_{nmq} = \begin{cases} 1 & \text{if} & -\infty < z_{nm}^* \leq \tau_{m1} \\ 2 & \text{if} & \tau_{m1} < z_{nm}^* \leq \tau_{m2} \\ 3 & \text{if} & \tau_{m2} < z_{nm}^* \leq \tau_{m3} \\ 4 & \text{if} & \tau_{m3} < z_{nm}^* \leq \tau_{m4} \\ 5 & \text{if} & \tau_{m4} < z_{nm}^* \leq +\infty \end{cases} \quad (15)$$

According to the formulation of an OL model in equation 15, we calculate the probability that individual n chooses indicator category k when answering survey question q as follows (Bierlaire, 2010):

$$\begin{aligned}
P(i_{nmq} = 1) &= Pr(\alpha_{mq}^{\text{image}} \cdot z_{nm}^* + \alpha_{mq}^{\text{choice}} \cdot y_{nm} \leq \tau_{m1}) \\
&= \frac{1}{1 + e^{\alpha_{mq}^{\text{image}} \cdot z_{nm}^* + \alpha_{mq}^{\text{choice}} \cdot y_{nm} - \tau_{m1}}} \tag{16}
\end{aligned}$$

$$\begin{aligned}
P(i_{nmq} = 2) &= Pr(\alpha_{mq}^{\text{image}} \cdot z_{nm}^* + \alpha_{mq}^{\text{choice}} \cdot y_{nm} \leq \tau_{m2}) - Pr(\alpha_{mq}^{\text{image}} \cdot z_{nm}^* + \alpha_{mq}^{\text{choice}} \cdot y_{nm} \leq \tau_{m1}) \\
&= \frac{1}{1 + e^{\alpha_{mq}^{\text{image}} \cdot z_{nm}^* + \alpha_{mq}^{\text{choice}} \cdot y_{nm} - \tau_{m2}}} - \frac{1}{1 + e^{\alpha_{mq}^{\text{image}} \cdot z_{nm}^* + \alpha_{mq}^{\text{choice}} \cdot y_{nm} - \tau_{m1}}} \tag{17}
\end{aligned}$$

$$\begin{aligned}
P(i_{nmq} = 3) &= Pr(\alpha_{mq}^{\text{image}} \cdot z_{nm}^* + \alpha_{mq}^{\text{choice}} \cdot y_{nm} \leq \tau_{m3}) - Pr(\alpha_{mq}^{\text{image}} \cdot z_{nm}^* + \alpha_{mq}^{\text{choice}} \cdot y_{nm} \leq \tau_{m2}) \\
&= \frac{1}{1 + e^{\alpha_{mq}^{\text{image}} \cdot z_{nm}^* + \alpha_{mq}^{\text{choice}} \cdot y_{nm} - \tau_{m3}}} - \frac{1}{1 + e^{\alpha_{mq}^{\text{image}} \cdot z_{nm}^* + \alpha_{mq}^{\text{choice}} \cdot y_{nm} - \tau_{m2}}} \tag{18}
\end{aligned}$$

$$\begin{aligned}
P(i_{nmq} = 4) &= Pr(\alpha_{mq}^{\text{image}} \cdot z_{nm}^* + \alpha_{mq}^{\text{choice}} \cdot y_{nm} \leq \tau_{m4}) - Pr(\alpha_{mq}^{\text{image}} \cdot z_{nm}^* + \alpha_{mq}^{\text{choice}} \cdot y_{nm} \leq \tau_{m3}) \\
&= \frac{1}{1 + e^{\alpha_{mq}^{\text{image}} \cdot z_{nm}^* + \alpha_{mq}^{\text{choice}} \cdot y_{nm} - \tau_{m4}}} - \frac{1}{1 + e^{\alpha_{mq}^{\text{image}} \cdot z_{nm}^* + \alpha_{mq}^{\text{choice}} \cdot y_{nm} - \tau_{m3}}} \tag{19}
\end{aligned}$$

$$\begin{aligned}
P(i_{nmq} = 5) &= 1 - Pr(\alpha_{mq}^{\text{image}} \cdot z_{nm}^* + \alpha_{mq}^{\text{choice}} \cdot y_{nm} \leq \tau_{m4}) \\
&= 1 - \frac{1}{1 + e^{\alpha_{mq}^{\text{image}} \cdot z_{nm}^* + \alpha_{mq}^{\text{choice}} \cdot y_{nm} - \tau_{m4}}} \tag{20}
\end{aligned}$$

Parameters τ_{mk} in equations 15 to 20 are thresholds of the unobserved latent variable z_{nm}^* and are to be estimated. Thereby, the indicator value i_{nmq} chosen by individual n as a response to one of the questions regarding the psychometric factors q depends on whether or not a certain threshold is crossed (Winship and Mare, 1984). For example, in assessment $q = 6$ (see Section 3) we asked students about their perception on how well they expect themselves to cope with the requirements to pass a certain major m . According to Equation 15 a given student responds to the question with "very good", which relates to indicator level $l_{\text{very good}} = 5$ in our study, if the value of the according latent variable z_{nm}^* is greater than the cutoff value τ_{m4} (Train, 2009, p. 164). From Equation 15 we see that higher values of

the latent variable z_{nm}^* correspond to higher indicator values i_{nmq} which is the result of a more positive assessment of the according psychometric factor q . Finally, equations 16 - 20 give the probability for an indicator value that students choose for their assessment of a given psychometric factor q . We studied $|Q| = 9$ psychometric indicators using confirmatory factor analysis to identify those that explain the relation to their respective latent variable in the best possible way. We follow the recommendations of [Kenny et al. \(1998\)](#) and [Kline \(2011, pp. 137-138\)](#) to choose three indicators per latent variable including: *quality of supervision*, *achievements in research*, *ability to cope with course requirements*, *internationality* and *practical relevance*. Figure 5 in the appendix provides a complete overview of the model specification and shows in detail which indicators are used to identify a given latent variable. We estimate our major choice ICLV model following an approach proposed by [Walker \(2001, p. 95\)](#). Thereby, we utilize the extended version of BIOGEME ([Bierlaire and Fetiariison, 2009](#)) to estimate all model parts and parameters simultaneously.

We simplify our model as we ignore the panel structure of the collected data (stated preference/choice experiment) by treating it as cross section. According to [Honoré \(2002\)](#) this leads to consistent and asymptotically normal estimates. For the estimation of the model parameters as specified in Equations 11 - 20 the joint likelihood function (10) is maximized. Since, our model specification yields a total of nine latent variables z_{nm}^* (i.e., one latent variable per major m) the joint likelihood function is the nine-dimensional integral of the MNL choice model over the distribution of the latent constructs. For estimation we apply 5000 random draws to each of the 3392 observations in our dataset.

5. Results

Choosing an extended approach like the ICLV modeling framework to investigate the major choice decision enormously increases the amount of parameters to be estimated and, due to simulated integration, the run time for model estimation. Therefore, we keep the overall model specification as simple as possible while ensuring that the assumed effects regarding the latent variables are still observable from the results.

Parameter	Estimate	t-Statistic	Estimate	t-Statistic	Estimate	t-Statistic
	MNL 1		MNL 2		MNL sub-model (ICLV)	
β^{asc}	-0.602	-2.45	-0.852	-3.25	-0.569	-2.24
β^{asc}_{FI}	-0.893	-4.22	-0.460	-1.99	-0.951	-4.36
β^{asc}_{HM}	-0.254	-1.38	0.108	0.53	-1.14	-2.99
β^{asc}_{MM}	-0.359	-1.46	-0.422	-1.59	-0.436	-1.69
β^{asc}_{DM}	-1.63	-7.23	-1.36	-5.37	-2.03	-7.87
β^{asc}_{DL}	-0.853	-3.00	-1.00	-3.33	-1.53	-4.45
β^{asc}_{SI}	-0.271	-1.19	-0.0697	-0.29	-2.43	-2.18
β^{asc}_{SM}	-0.836	-3.91	-0.973	-4.18	-1.53	-5.11
β^{asc}_{IM}	-0.938	-4.29	-0.907	-3.83	-1.56	-5.74
β^{asc}_{AT}	-0.176	-3.17	-0.180	-3.22	-0.183	-3.21
β^{income}	0.310	6.72	0.304	6.53	0.319	6.70
β^{gender}_{FI}	.	.	0.552	3.31	.	.
β^{gender}_{HM}	.	.	-0.747	-4.42	.	.
β^{gender}_{MM}	.	.	-0.599	-3.62	.	.
β^{gender}_{DM}	.	.	0.215	1.31	.	.
β^{gender}_{DL}	.	.	-0.451	-2.08	.	.
β^{gender}_{SI}	.	.	0.357	2.04	.	.
β^{gender}_{SM}	.	.	-0.292	-1.87	.	.
β^{gender}_{IM}	.	.	0.320	1.91	.	.
β^{gender}_{AT}	.	.	-0.0230	0.13	.	.
λ^{image}	0.311	4.94
Observations	3392		3392		3392	
Parameters	11		20		109	
$\mathcal{L}(\beta_0)$	-3726.493		-3726.493		-558723.756	
$\mathcal{L}(\hat{\beta})$	-3433.775		-3396.581		-120196.834	
ρ^2	0.079		0.089		0.784	
$\hat{\rho}^2$	0.076		0.083		0.784	

Table 3: Comparison of the estimation results of two simple MNL models and the MNL sub-model of the major choice ICLV model.

In total, we estimate 109 model parameters from 45 model equations: the impact of observed attributes and latent variables on utility as defined by Equations 11, the impact of socio-economic characteristics on latent variables as defined by Equations 13 and the impact of the latent variables on the observed indicator values as defined by Equations 14.

We display the estimated parameters for the structural equations of the MNL sub-model as well as the summary statistics of the overall ICLV model in Table 3. Furthermore, in Table 4 in the appendix we include the estimation results for the parameters of the MNL sub-model as well as the overall summary statistics of the ICLV that we obtained by applying different numbers of random draws to the MSL procedure (Equation 10). The results show that estimates are consistent in their values while the number of draws increases. In Table 3 we also display the results obtained from estimating two simple MNL models without latent variables from the same data set. Differences in the initial log-likelihood values $\mathcal{L}(\beta_0)$ result from differences in the log-likelihood functions for the simple MNL models and the ICLV model (Equation 10). We choose a specification for the MNL 1 model that is similar to the MNL sub-model of Equation 11. Due to a more fair comparison the specification of the MNL 2 model includes the individual characteristic *gender*, since *image* z_{nm}^* is a function of gender (Equation 13).

With regard to the $\bar{\rho}^2$ values, we see large differences between the MNL 1 and MNL 2 model specifications compared to the MNL sub-model of the ICLV. The model fit for the MNL 1 and MNL 2 models is 0.076 and 0.083 respectively. The slightly better fit of the MNL 2 model can be addressed to the introduction of heterogeneity to the model by adding the *gender* variable. However, with $\bar{\rho}^2 = 0.784$, we see a large difference in model fit between the simple MNL models and the ICLV model. This finding supports our assumption that the major choice decision of students at Hamburg Business School is influenced by unobserved latent factors.

Within the model specifications of both the two MNL models and the MNL sub-model, the explanatory attributes are included as linear terms in the utility functions of the alternatives. Thereby, the utility functions of the MNL sub-model additionally include the alternative-specific latent variables. To ensure model identification, we choose the no-choice option as the reference alternative for both models.¹

¹However, the decision about which attributes to normalize is arbitrary since either

In terms of the MNL sub-model, we can reject the null hypothesis $\beta = 0$ on a 95% significance level for all coefficients except for the alternative-specific constant (ASC) of the OM major β_{OM}^{asc} . Furthermore, we find that all estimated parameters show the expected signs. As expected, parameter β^{grade} shows a negative sign since we assume that students receive less utility from a major where grades are poor on average. The *income* parameter β^{income} as well as the parameter that estimates the impact of the latent variable *image* λ^{image} both show positive signs. This is also expected since we assume that a higher income and a higher image favor the selection of a certain major. From these results we conclude that latent variables can be identified within the major choice ICLV model that do in fact have an influence on the utility of the major alternatives. Please note, that we did not include the explanatory variable for the number of courses into the MNL sub-model, since we discovered during model specification that it does not have a statistically significant impact on the major-specific utilities.

The estimated parameters of the MNL 1 model are statistically significant at least on the 95% confidence level except for the ASC's β_{MM}^{asc} , β_{OM}^{asc} and β_{SM}^{asc} . For the MNL 2 model, we find that estimates for the ASC's β_{HM}^{asc} , β_{MM}^{asc} and β_{OM}^{asc} are not statistically significant. For all three models, the *income* and *grade* coefficients are similar in value and have the expected signs. With regard to the MNL 2 model, we find that being a male student at Hamburg Business School negatively influences the utility of choosing majors HM, MM, BL and SM. In turn, this means that the utility a female student gains from choosing one of these majors is positively affected. This result is in line with studies that examine gender-specific issues in major choice. Research finds that for different reasons that are not related to abilities women prefer to choose non-quantitative majors more often than men. For example [Zafar \(2013\)](#) shows, that women are less likely to major in a mathematically demanding field like engineering because they believe they won't enjoy coursework. [Correll \(2001\)](#) finds that men assess their mathematical competence higher than females. In the context of our study, this might result in women choosing less-quantitative majors to increase their probability of graduation. [Sutter and Glätzle-Rützler \(2014\)](#) study the gender gap in several experiments with 3 to 18 year olds. They find that females, although they performed equally well or better in the experiments than males, are less

normalization results in the same model ([Ben-Akiva and Lerman, 1985](#), p. 287).

willing to compete. In addition to this, [Buser et al. \(2014\)](#) show that the willingness to compete is positively related to choosing a demanding education track.

In the LV sub-model nine alternative-specific latent variables *image* are used to explain the values of 27 perceptual indicators. Estimation results are displayed in Table 5 in the appendix. We do not report fixed parameters. According to the specification in Equations 13, the influence of the individual characteristic on the latent variables is weighted by alternative-specific parameters γ_m^{gender} and γ_m^{mean} while the latter has the function of an intercept. Besides the explanatory characteristic *gender* we also tested the inclusion of other individual-specific variables like *age* and *friends within the same major*. In both cases, we did not obtain meaningful results from adding these dummy variables. Thus, we omit them from the final model specification.

Rules for model identification for the structural equations of the LV sub-model apply in the same way as for the MNL model ([Walker, 2001](#), p. 94). Hence, from testing various specifications we normalize parameters $\gamma_{HM}^{\text{mean}}$ and $\gamma_{HM}^{\text{gender}}$. This results in the HM major being the reference alternative. Furthermore, the values of γ_m^{mean} serve as intercepts in a similar way as ASC's in the MNL choice model.

For the LV sub-model, we can reject the null hypotheses $\gamma_m^{\text{mean}} = 0$ for all majors m on at least the 95% significance level and $\gamma_m^{\text{gender}} = 0$ for all majors except for OM, BL and AT on a 95% significance level. Thus, for these majors we can not assume that *gender* influences the value of the latent variable. However, from the estimates we can conclude that for male students the *image* value is higher for the FI major compared to the reference major HM. Correspondingly, *image* values are lower, compared to the HM major, for the remaining majors.

From the estimates of the normal error components ω_{nm} we are able to calculate the standard deviations for each of the latent variables. Thus, we determine the distribution of the latent variables z_{nm}^* across majors for male students from Equations 13. Figure 3 displays the distribution of the latent variables z_{nm}^* for male students ($s^{\text{gender}} = 1$) across majors. We find that the SM major has the highest *image* values within the population. In line with the findings of [Pritchard et al. \(2004\)](#), we assume that this particular major is popular among students since its teaching contents do not require quantitative or analytical skills. Thus, it may be perceived by some students as a major that is easier to pass than more quantitative majors. Choosing a major that requires less quantitative abilities might increase a given student's

chances of graduation which is in line with the findings of [Montmarquette et al. \(2002\)](#). Another aspect might be the perception that a less demanding major provides a higher probability of receiving good grades. In contrast, we find that the OM major and the FI major have the lowest mean values of the latent variables. Both majors have a reputation for providing lectures that require very good quantitative and analytic skills combined with an affinity for working with IT systems and tailored software. Students who are less confident about their quantitative skills may perceive these requirements as a factor that decreases their probability to graduate since failed exams can only be repeated twice. Another aspect of OM and FI having a low perceived image might be that both the major's lectures for undergraduate students are currently scheduled to the 4th semester. Thus students gain very little or no experience with specific OM and FI teaching contents prior to making their major choice decision. However, besides the mean values for the latent variables their standard deviations suggest that the perceived *image* values per major vary to different degrees across students. We find that the standard deviation ranges from 0.51 for the HM major to 1.62 for the AT major. For the SM major we have a mean value of $\mu^{SM} = 6.73$ and standard deviation of $\sigma^{SM} = 0.77$. Thus, for 95% of the questioned students the perceived image values for the SM major are in the range of 5.23 to 8.24 which is simply the interval $\mu^{SM} \pm 1.96\sigma^{SM}$. The means of the image variable for the IM, ST and AT major are quite similar at values of $\mu^{IM} = 2.04$, $\mu^{ST} = 1.98$ and $\mu^{AT} = 1.88$ respectively. However, standard deviations for these majors vary with values of $\sigma^{IM} = 0.97$, $\sigma^{ST} = 0.70$ and $\sigma^{AT} = 1.62$. Thus, for the ST major 95% of the *image* values lie between 0.61 and 3.35, for the IM major between 0.15 and 3.93 and for the AT major between -1.30 and 5.06.

The distributions of *image* variables across majors shows, that perceptions for the same major can vary greatly across students. However, from our model we are not able to identify specific reasons for this variation. At this point, further analysis with latent class models that account for unobserved heterogeneity within the population of students might reveal more differentiated results ([Walker, 2001](#), p. 128).

The measurement equations of the LV sub-model provide the links between the values of the observed response indicators and the latent variables. This relationship is

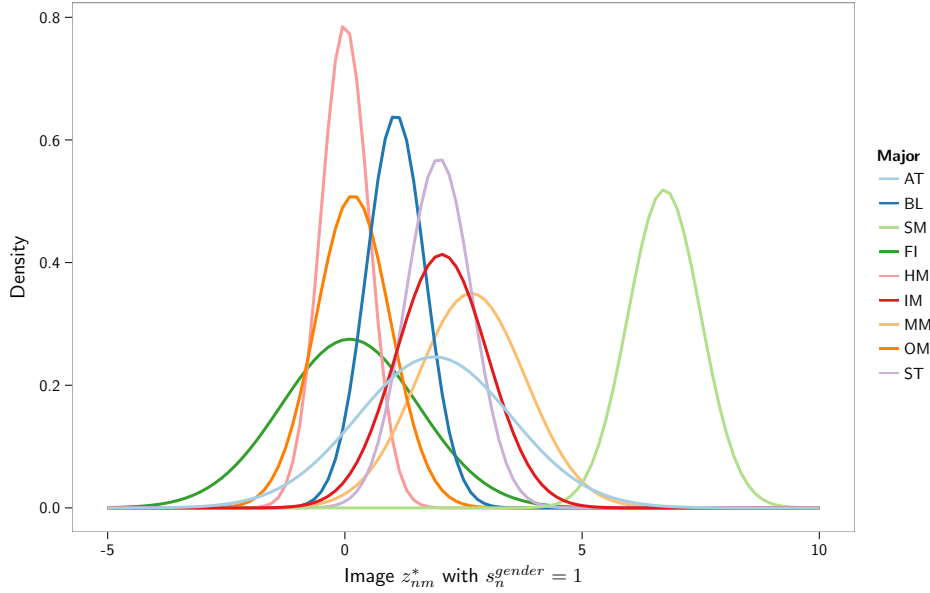


Figure 3: Distribution of the latent variable *image* for male students z_{nm}^* across majors.

defined by equations 14. From the estimation results we are able to investigate whether the latent variables *image* do influence the selection of the response indicator values. As described in Section 3 we obtain values of these indicators through psychometric factors survey questions. Since the values of the latent variables are unmeasured, their scale has to be set for estimation. Thus, for each latent variable we fix one of the interaction terms α_{mq}^{image} in Equations 15 to one for normalization. Daly et al. (2012) refer to this as Ben-Akiva normalization following a normalization strategy set out by Ben-Akiva et al. (2002b). The authors also provide a comparison with another normalization strategy by Bolduc et al. (2005) and find both strategies to be equivalent.

By normalizing the interaction terms in our specification, we put constraints on the parameters in one of the measurement equations per latent variable. The choice of the interaction term is generally arbitrary. However, we found that the decision about which interaction terms to fix might influence the overall model fit and should be examined during the process of model building. Therefore, in case of multiple latent variables we suggest to first specify single ICLV models for each latent variable to identify the best specification.

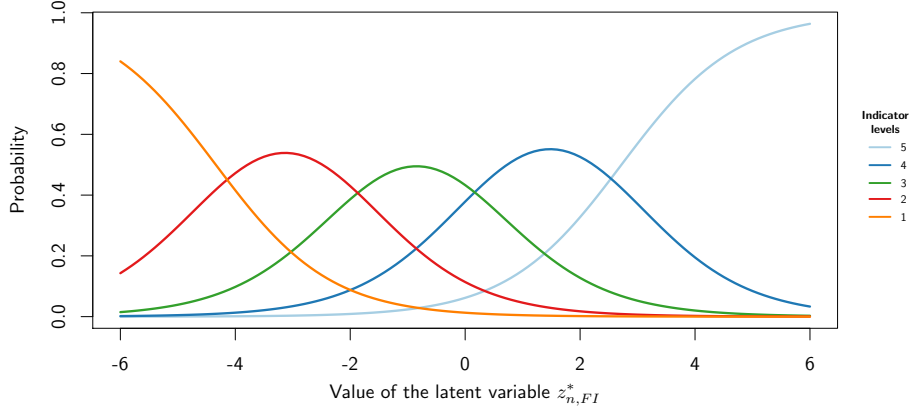


Figure 4: Probabilities $P(i_{nmq} = l_k)$ according to Equations 16 - 20 for the psychometric factor 'achievements in research' for the FI major.

Within the LV sub-model each measurement equation contains two parameters that are estimated: (1) parameter $\alpha_{mq}^{\text{image}}$ weighting the influence of the latent variable z_{nm}^* image on the according indicator value; (2) parameter $\alpha_{mq}^{\text{choice}}$ captures exaggerated responses in the assessments of the perceptual survey questions for the chosen alternative that might be caused by justification bias. Table 6 displays the estimated parameters for the measurement equations of the LV sub-model whereas fixed parameters are not reported. Thus, we ensure that the estimated results for the remaining coefficients are unbiased. From the estimation results, we conclude that except for parameters $\alpha_{FI,\text{research}}^{\text{choice}}$, $\alpha_{FI,\text{research}}^{\text{choice}}$, $\alpha_{OM,\text{research}}^{\text{choice}}$, $\alpha_{BL,\text{research}}^{\text{choice}}$ and $\alpha_{ST,\text{research}}^{\text{choice}}$ the chosen major, that is represented by indicator y_{nm} in Equation 14, has a statistically significant influence on the value that students select for the assessment of the psychometric factors of that same major.

From the t-statistics of the $\alpha_{mq}^{\text{image}}$ parameters we conclude that the indicator values that we obtain from the students responses to the psychometric factors survey questions are actually influenced by the unobserved latent variables *image* z_{nm}^* . Furthermore, from $\alpha_{HM,\text{research}}^{\text{image}} = -1.47$ we conclude that the HM major increases its *image* value through poor research activities. According to Equation 14 higher values of $z_{n,HM}^*$ result in a decreasing assessment of the psychometric factor for achievements in research. An additional decrease in the indicator value $i_{n,HM,6}$ by -1.68 ($\alpha_{HM,\text{research}}^{\text{choice}}$) results from HM being the chosen major of a given student with $y_{n,HM} = 1$, correcting for exagger-

ated assessments of the respective psychometric factor. For the remaining majors the contribution of either the *image* variable and the choice indicator to the value of the indicators i_{nmq} are positive. Both, higher *image* values and the major being the chosen alternative result in higher assessments of the respective majors psychometric factors.

The parameters τ_{mk} of the threshold model as defined by Equation 15 are listed in Table 7 in the appendix. We choose a generic specification for the threshold parameters as defined in Equations 15 - 20. We estimate parameters τ_{mk} for each major and psychometric factor. Nine latent variables yield 36 additional parameters compared to 72 parameters if we decide to let the thresholds vary across indicators. However, although we are not able to capture variation in the thresholds for the indicators of the same major, we limit computational complexity, which is important for a model of this size.

Except for $\tau_{SM,1}$ all estimated threshold parameters are statistically significant on the 99% confidence interval. Knowing the values of the τ_{mk} we can also provide a graphical illustration of the indicator probabilities in Equations 16 -20. Exemplary for the FI major and psychometric factor 'achievements in research' we present these probabilities in Figure 4. Thereby, each curve represents the probability function for the selection of an indicator level l_k . Hence, for increasing values of the latent variable $z_{n,FI}^*$ the probability of selecting indicator level $l_k = 5$ for the assessment of the psychometric factors of the FI major approaches 100%. In general, larger values of the latent variable favor the selection of higher indicator levels by the respondents. Figure 4 reveals the intervals of the latent variable values in which the according indicator levels have the highest probabilities of being chosen. The same is valid for very low values of the latent variable. In particular, when the image value a student n perceives for the FI major, $z_{n,FI}^*$, is less than -4.34 which corresponds to $\tau_{FI,1}$, indicator level $l_k = 1$ (very low) is the option that is selected with the highest probability for assessing the psychometric factor 'achievements in research'.

6. Conclusion

Our study shows the application of an ICLV model with ordered perceptual indicators that we apply for modeling the major choice decision of students at Hamburg Business School. Estimation of the model components is performed simultaneously using a full information maximum likelihood (FIML) estimator. We investigate the influence of major-specific latent vari-

ables *image* on the decision process of the students. From the results of our study, we draw the following important conclusions:

1. We find the latent variables *image* influence the assessment of selected psychometric factors that we employ within the measurement equations of the LV sub-model. The results from the MNL sub-model confirm that the *image* of a major indeed has significant influence on the students choice decision.
2. By plotting the distributions of the nine latent variables, we identify Strategic Management (SM) as the major with the highest *image* value. We are not able to draw this exact conclusion solely based on the assessments of the psychometric factors as shown in Figure 2. This is an important finding which shows that indicator values cannot be directly translated into latent *image* variables. It reveals the advantage of the ICLV model that enables us to identify variables that directly influence the students choice decision and are crucial from the academic departments' perspective. Without the ICLV model, decisions would be based upon the findings of Figure 2 leading to erroneous orientation of the academic departments.
3. The remaining majors show quite similar mean values of the latent variable while their standard deviations vary to a large degree. We find the *image* variable of the AT and FI majors varies the most. On the contrary, the variance of the distribution of the latent variable is lowest for the HM major that serves as the reference alternative. With regard to the MNL sub-model, we further find that observed explanatory variables *income* and *grade* significantly influence the students choice decision.

From the obtained results, we conclude that study deans and academic departments can further investigate how to obtain and preserve the students' interest throughout their academic career. However, from the proposed specification of the MNL sub-model it is clear, that observable explanatory variables are less suitable to be adjusted in order to increase enrollment numbers for the different majors. Neither students grades nor expected income after graduation lie within the sphere of influence of the academic departments. However, to increase attractiveness we conclude that academic departments could focus on improving the external representation of majors. This can be done by improving students' perceptions with regard to the areas covered by the psychometric factors that we applied in our model. For example,

supervision quality could be improved by implementing best practices for addressing different students' needs. In particular, majors could provide more time slots for consultation. While the majority of students usually requires little assistance, some students prefer to attend consultation regularly. To date, consultation hours that are offered by academic staff are not accounted for within the individual teaching loads although they can easily sum up to several hours per week. Besides research, preparations for lectures and teaching itself, this leaves little incentive to increase the amount of consultation hours.

Another aspect to improve the major's image might be achieved by displaying current research projects proactively. For example, selected topics could enter lectures and seminars. Thus, students gain first hand experience in working on real world problems. This would also help with the improvement of the practical relevance perception. Of course, not all of the above indicators can be easily adjusted from the perspective of the different majors. Especially, if students think they are lacking interest or the ability to pass a certain major. Although instructors cannot influence students abilities they do have influence on students attitudes towards classes and the learning environment which are related to students' performance and success (Depaolo and McLaren, 2006). Another possibility is to address students very early in their studies, when they attend the first basic lectures and get in touch with the different majors for the first time. This could be achieved by rescheduling certain lectures to the earlier stages of the studies. For example, the modules "Introduction to Operations Research" and "Logistics", offered by the OM major were both scheduled in the 4th semester (2nd year) in 2013. Students choose their majors with little or no experience in OM since these are the only modules that are offered by OM during the students first two years of studies. According to Haselhuhn et al. (2012) personal experience changes behavior. Thus, lectures that are commonly perceived as difficult to pass, such as operations research (OR) or quantitative lectures in general, could help such majors to gain popularity. That way, students would be able to familiarize with challenging but nonetheless interesting topics early on. This might raise the probability of choosing majors like IM, AT and OM over SM and MM. We assume that the popularity of the latter is attributed to a perception among students of less quantitative lectures and thus a higher probability of graduation (Montmarquette et al., 2002). Another interesting approach for quantitative majors is proposed by Wilder and Ozgur (2015). The authors suggest the revision of current quantitative courses and to shift

contents to business analytics topics like data visualization, data mining and prescriptive analytics. In the context of increasing job opportunities in the data science field, the authors justify their recommendation by explaining that future managers are required to be more and more data-savvy to be successful in their jobs. Since data analytics are already part of some lectures in the OM major, adding such contents to undergraduate lectures might also help to improve the *practical relevance* perceptions of students towards that particular major.

At this point, an additional analysis that identifies different types of students and their respective preferences in academia could be helpful. We therefore suggest to further investigate the presented major choice problem based on a latent class (LC) model. Additionally, more flexible choice models than MNL should be applied within the discrete choice model component of the ICLV to improve forecasts. Thereby, we suggest the application of NL or CNL models that exhibit non-constant substitution patterns between alternatives while providing a closed-form choice model.

7. Appendix

7.1. Specification of the major choice ICLV model

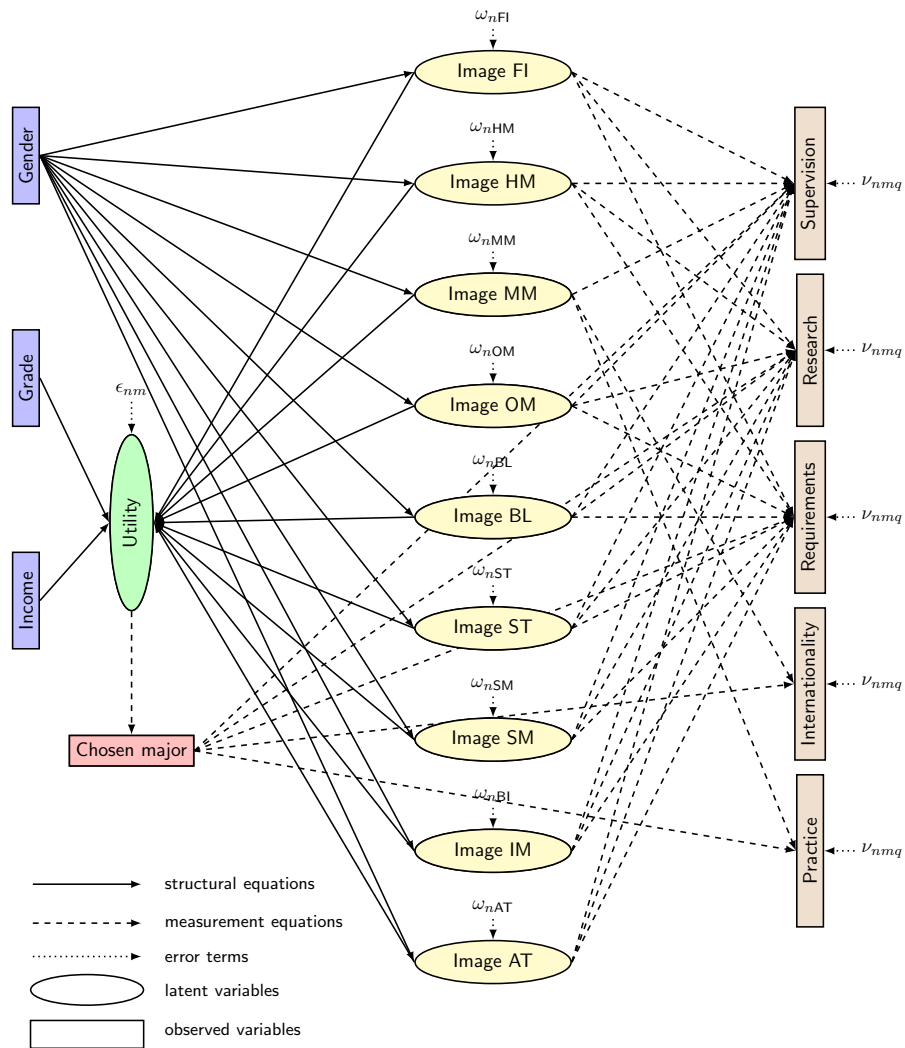


Figure 5: Structural and measurement equations of the major choice ICLV model.

7.2. Estimation results of the discrete choice sub-model for varying number of draws

Parameter	Estimate t-Statistic		Estimate t-Statistic		Estimate t-Statistic	
	5000 Draws		2000 Draws		1000 Draws	
<i>Structural Equations - MNL</i>						
β_{TF}^{ASC}	-0.569	-2.24	-0.565	-2.23	-0.558	-2.20
β_{TM}^{ASC}	-0.951	-4.36	-0.939	-4.32	-0.936	-4.30
β_{MT}^{ASC}	-1.14	-2.99	-1.08	-2.95	-1.09	-2.88
β_{OM}^{ASC}	-0.436	-1.69	-0.419	-1.64	-0.422	-1.64
β_{OL}^{ASC}	-2.03	-7.87	-2.01	-7.83	-1.99	-7.78
β_{TL}^{ASC}	-1.53	-4.45	-1.48	-4.38	-1.51	-4.35
β_{SM}^{ASC}	-2.43	-2.18	-2.30	-1.74	-2.33	-1.91
β_{TM}^{ASC}	-1.53	-5.11	-1.50	-4.99	-1.55	-4.92
β_{AT}^{ASC}	-1.56	-5.74	-1.53	-5.67	-1.51	-5.62
β^{grade}	-0.183	-3.21	-0.181	-3.19	-0.182	-3.20
β^{income}	0.319	6.70	0.317	6.68	0.317	6.67
λ^{image}	0.311	4.94	0.294	4.71	0.298	4.52
Number of observations	3392		3392		3392	
Number of estimated parameters	109		109		109	
$\mathcal{L}(\hat{\beta}_0)$	-558723.756		-558723.756		-558723.756	
$\mathcal{L}(\hat{\beta})$	-120196.834		-120209.385		-120262.534	
$\hat{\rho}^2$	0.784		0.784		0.784	
$\hat{\rho}^2$	0.784		0.784		0.784	

Table 4: Estimation results for the structural equations of the discrete choice sub-model for different numbers of random draws.

7.3. Estimation results of the latent variable sub-model of the major choice ICLV model

Parameter	Estimate	Std. Error	t-Statistic
<i>Structural Equations - LVM</i>			
γ_{FI}^{mean}	-0.397	0.126	-3.16
γ_{MM}^{mean}	3.01	0.774	3.89
γ_{OM}^{mean}	0.274	0.136	2.02
γ_{BL}^{mean}	1.13	0.190	5.97
γ_{ST}^{mean}	2.19	0.267	8.19
γ_{SM}^{mean}	7.20	3.47	2.08
γ_{IM}^{mean}	2.30	0.405	5.68
γ_{AT}^{mean}	1.77	0.192	9.23
γ_{FI}^{gender}	0.499	0.109	4.59
γ_{MM}^{gender}	-0.344	0.0579	-5.94
γ_{OM}^{gender}	-0.102	0.0587	-1.74
γ_{BL}^{gender}	-0.0583	0.0425	-1.37
γ_{ST}^{gender}	-0.208	0.0399	-5.22
γ_{SM}^{gender}	-0.466	0.0442	-10.55
γ_{IM}^{gender}	-0.258	0.0526	-4.91
γ_{AT}^{gender}	0.113	0.0932	1.22
$\sigma_{\omega_n, FVI}$	1.45	0.0574	25.24
$\sigma_{\omega_n, HM}$	0.506	0.0774	6.54
$\sigma_{\omega_n, MM}$	1.14	0.0516	22.02
$\sigma_{\omega_n, OM}$	0.783	0.0957	8.18
$\sigma_{\omega_n, BL}$	0.622	0.0401	15.51
$\sigma_{\omega_n, ST}$	0.700	0.0426	16.41
$\sigma_{\omega_n, SM}$	0.769	0.0617	12.47
$\sigma_{\omega_n, IM}$	0.965	0.0370	26.06
$\sigma_{\omega_n, AT}$	1.62	0.0915	17.72

Table 5: Estimation results for the structural equations of the LV sub-model as defined by Equations 13 (fixed coefficients are not reported).

Parameter	Estimate	Std. Error	t-Statistic
<i>Measurement Equations - LVM</i>			
<i>image</i>			
$\alpha_{FI, \text{research}}$	0.421	0.0585	7.20
<i>choice</i>			
$\alpha_{FI, \text{research}}$	0.402	0.119	3.39
<i>image</i>			
$\alpha_{FI, \text{supervision}}$	0.447	0.0772	5.79
<i>choice</i>			
$\alpha_{FI, \text{supervision}}$	-0.198	0.116	-1.71
<i>image</i>			
$\alpha_{HM, \text{research}}$	-1.47	0.241	-6.10
<i>choice</i>			
$\alpha_{HM, \text{research}}$	-1.68	0.287	-5.84
<i>image</i>			
$\alpha_{HM, \text{supervision}}$	0.915	0.172	5.32
<i>choice</i>			
$\alpha_{HM, \text{supervision}}$	0.766	0.123	6.22
<i>image</i>			
$\alpha_{MM, \text{internationality}}$	0.927	0.0260	35.67
<i>choice</i>			
$\alpha_{MM, \text{internationality}}$	0.366	0.120	3.04
<i>image</i>			
$\alpha_{MM, \text{practice}}$	0.772	0.0607	12.72
<i>choice</i>			
$\alpha_{MM, \text{practice}}$	0.381	0.116	3.28
<i>image</i>			
$\alpha_{OM, \text{research}}$	1.06	0.232	4.58
<i>choice</i>			
$\alpha_{OM, \text{research}}$	0.202	0.121	1.67
<i>image</i>			
$\alpha_{OM, \text{supervision}}$	0.720	0.104	6.95
<i>choice</i>			
$\alpha_{OM, \text{supervision}}$	0.185	0.0994	1.86
<i>image</i>			
$\alpha_{BL, \text{research}}$	0.606	0.0484	12.53
<i>choice</i>			
$\alpha_{BL, \text{research}}$	0.190	0.170	1.12
<i>image</i>			
$\alpha_{BL, \text{supervision}}$	1.45	0.103	14.15
<i>choice</i>			
$\alpha_{BL, \text{supervision}}$	0.419	0.211	1.99
<i>image</i>			
$\alpha_{ST, \text{research}}$	0.718	0.0355	20.25
<i>choice</i>			
$\alpha_{ST, \text{research}}$	0.0589	0.114	0.52
<i>image</i>			
$\alpha_{ST, \text{supervision}}$	1.86	0.118	15.71
<i>choice</i>			
$\alpha_{ST, \text{supervision}}$	0.337	0.158	2.13
<i>image</i>			
$\alpha_{SM, \text{research}}$	1.12	0.0597	18.71
<i>choice</i>			
$\alpha_{SM, \text{research}}$	0.335	0.108	3.09
<i>image</i>			
$\alpha_{SM, \text{requirements}}$	1.18	0.0866	13.60
<i>choice</i>			
$\alpha_{SM, \text{requirements}}$	0.953	0.105	9.11
<i>image</i>			
$\alpha_{IM, \text{research}}$	0.605	0.0619	9.78
<i>choice</i>			
$\alpha_{IM, \text{research}}$	0.261	0.132	1.97
<i>image</i>			
$\alpha_{IM, \text{requirements}}$	1.16	0.0480	24.22
<i>choice</i>			
$\alpha_{IM, \text{requirements}}$	1.19	0.0529	22.41
<i>image</i>			
$\alpha_{AT, \text{research}}$	0.279	0.0473	5.89
<i>choice</i>			
$\alpha_{AT, \text{research}}$	0.400	0.149	2.69
<i>image</i>			
$\alpha_{AT, \text{supervision}}$	0.513	0.0530	9.69
<i>choice</i>			
$\alpha_{AT, \text{supervision}}$	0.455	0.131	3.48

Table 6: Estimation results for the parameters α_{mq}^{image} and α_{mq}^{choice} of the measurement equations of the LV sub-model as defined by Equation 14.

Parameter	Estimate	Std. Error	t-Statistic
<i>Thresholds τ_{mk}</i>			
$\tau_{FI,1}$	-4.34	0.128	-33.89
$\tau_{FI,2}$	-1.93	0.0810	29.70
$\tau_{FI,3}$	0.24	0.0371	58.51
$\tau_{FI,4}$	2.72	0.0456	54.42
$\tau_{HM,1}$	-4.12	0.0798	-51.59
$\tau_{HM,2}$	2.01	0.0720	29.30
$\tau_{HM,3}$	-0.10	0.0331	57.55
$\tau_{HM,4}$	2.03	0.0342	62.34
$\tau_{MM,1}$	-2.58	0.782	-3.30
$\tau_{MM,2}$	-0.34	0.103	21.68
$\tau_{MM,3}$	1.69	0.0402	50.49
$\tau_{MM,4}$	4.27	0.0397	64.90
$\tau_{OM,1}$	-4.14	0.152	-27.22
$\tau_{OM,2}$	-2.30	0.0807	22.79
$\tau_{OM,3}$	0.17	0.0423	58.40
$\tau_{OM,4}$	2.89	0.0450	60.57
$\tau_{BL,1}$	-1.31	0.164	-7.94
$\tau_{BL,2}$	0.42	0.0346	50.00
$\tau_{BL,3}$	2.16	0.0297	58.57
$\tau_{BL,4}$	4.06	0.0454	41.84
$\tau_{ST,1}$	-2.48	0.266	-9.31
$\tau_{ST,2}$	-0.09	0.0942	25.36
$\tau_{ST,3}$	1.93	0.0358	56.45
$\tau_{ST,4}$	4.00	0.0384	53.82
$\tau_{SM,1}$	3.54	3.45	1.03
$\tau_{SM,2}$	5.58	0.0650	31.39
$\tau_{SM,3}$	7.81	0.0388	57.35
$\tau_{SM,4}$	10.16	0.0424	55.53
$\tau_{IM,1}$	-1.95	0.369	-5.29
$\tau_{IM,2}$	-0.33	0.0614	26.44
$\tau_{IM,3}$	2.01	0.0362	64.63
$\tau_{IM,4}$	4.51	0.0436	57.28
$\tau_{AT,1}$	-3.19	0.174	-18.36
$\tau_{AT,2}$	-0.95	0.0708	31.66
$\tau_{AT,3}$	1.40	0.0368	63.83
$\tau_{AT,4}$	3.60	0.0441	49.90

Table 7: Estimation results for the threshold parameters of the measurement equations of the LV sub-model as defined by Equations 15 - 20.

References

- Ariely, D., 2008. Predictably Irrational. HarperCollins New York.
- Ben-Akiva, M., Mc Fadden, D., Train, K., Walker, J., Bhat, C., Bierlaire, M., Bolduc, D., Boersch-Supan, A., Brownstone, D., Bunch, D., Daly, A., De Palma, A., Gopinath, D., Karlstrom, A., Munizaga, M., 2002a. Hybrid choice models: Progress and challenges. *Marketing Letters* 13 (3), 163–175.
- Ben-Akiva, M., Walker, J., Bernardino, A. T., Gopinath, D. A., Morikawa, T., Polydoropoulou, A., 2002b. Integration of choice and latent variable models. In: Mahmassani, H. S. (Ed.), *In perpetual motion: Travel behaviour research opportunities and application challenges*. Elsevier Science, pp. 431–470.
- Ben-Akiva, M. E., Lerman, S. R., 1985. *Discrete Choice Analysis, Theory and Application to Travel Demand*. MIT Press, Cambridge, MA.
- Berger, M. C., 1988. Predicted future earnings and choice of college major. *Industrial & Labor Relations Review* 41 (3), 418–429.
- Bierlaire, M., 2010. Estimating hybrid choice models with the new version of biogeme. In: Seminar series. No. EPFL-TALK-152415.
- Bierlaire, M., Fetiariison, M., 2009. Estimation of discrete choice models: extending biogeme. In: *Swiss Transport Research Conference (STRC)*.
- Bolduc, D., Ben-Akiva, M., Walker, J., Michaud, A., 2005. Hybrid choice models with logit kernel: Applicability to large scale models. In: Martin E.H. Lee-Gosselin, S. T. D. (Ed.), *Integrated Land-Use and Transportation Models*. Emerald Group Publishing, pp. 275–302.
- Buser, T., Niederle, M., Oosterbeek, H., 2014. Gender, competitiveness, and career choices. *The Quarterly Journal of Economics* 129 (3), 1409–1447.
URL <http://qje.oxfordjournals.org/content/129/3/1409.abstract>
- Correll, S. J., 2001. Gender and the career choice process: The role of biased self-assessments¹. *American Journal of Sociology* 106 (6), 1691–1730.

- Daly, A., Hess, S., Patrui, B., Potoglou, D., Rohr, C., 2012. Using ordered attitudinal indicators in a latent variable choice model: a study of the impact of security on rail travel behaviour. *Transportation* 39 (2), 267–297.
- Depaolo, C., McLaren, C. H., 2006. The relationship between attitudes and performance in business calculus. *INFORMS Transactions on Education* 6 (2), 8–22.
- Frischknecht, B. D., Eckert, C., Geweke, J., Louviere, J. J., 2014. A simple method for estimating preference parameters for individuals. *International Journal of Research in Marketing* 31 (1), 35–48.
- Glerum, A., Stankovikj, L., Thémans, M., Bierlaire, M., 2014. Forecasting the demand for electric vehicles: accounting for attitudes and perceptions. *Transportation Science* 48 (4), 483–499.
- Haselhuhn, M. P., Pope, D. G., Schweitzer, M. E., Fishman, P., 2012. The impact of personal experience on behavior: Evidence from video-rental fines. *Management Science* 58 (1), 52–61.
- Honoré, B. E., 2002. Nonlinear models with panel data. *Portuguese Economic Journal* 1 (2), 163–179.
- Hoyos, D., Mariel, P., Hess, S., 2015. Incorporating environmental attitudes in discrete choice models: An exploration of the utility of the awareness of consequences scale. *Science of the Total Environment* 505, 1100–1111.
- Kamakura, W. A., Wedel, M., Agrawal, J., 1994. Concomitant variable latent class models for conjoint analysis. *International Journal of Research in Marketing* 11 (5), 451–464.
- Kenny, D., Kashy, D., Bolger, N., 1998. Data analysis in social psychology. In: *The Handbook of Social Psychology*. McGraw-Hill, pp. 233–265.
- Kim, D., Markham, F. S., Cangelosi, J. D., 2002. Why students pursue the business degree: A comparison of business majors across universities. *Journal of Education for Business* 78 (1), 28–32.
- Kline, R. B., 2011. *Principles and practice of structural equation modeling*. Guilford publications.

- Kuhfeld, W. F., Tobias, R. D., Garratt, M., 1994. Efficient experimental design with marketing research applications. *Journal of Marketing Research* 31 (4), 545–557.
- Leppel, K., Williams, M. L., Waldauer, C., 2001. The impact of parental occupation and socioeconomic status on choice of college major. *Journal of Family and Economic Issues* 22 (4), 373–394.
- Likert, R., 1932. A technique for the measurement of attitudes. *Archives of Psychology*.
- Loomes, G., Pogrebna, G., 2016. Do preference reversals disappear when we allow for probabilistic choice? *Management Science, Papers in Advance*.
- Malgwi, C. A., Howe, M. A., Burnaby, P. A., 2005. Influences on students' choice of college major. *Journal of Education for Business* 80 (5), 275–282.
- Marschak, J., 1960. Binary-choice constraints and random utility indicators. In: *Mathematical Methods in the Social Sciences*. Stanford University Press, pp. 312–329.
- McFadden, D., 1974. Conditional logit analysis of qualitative choice behaviour. In: Zarembka, P. (Ed.), *Frontiers of Econometrics*. Academic Press, New York, pp. 105–142.
- McFadden, D., 1986. The choice theory approach to market research. *Marketing Science* 5, 275–297.
- McFadden, D., 2001. Economic choices. *American Economic Review* 91 (3), 351–378.
- McFadden, D., Machina, M. J., Baron, J., 1999. Rationality for economists? *Journal of Risk and Uncertainty* 19 (1-3), 73–105.
- Montmarquette, C., Cannings, K., Mahseredjian, S., 2002. How do young people choose college majors? *Economics of Education Review* 21 (6), 543–556.
- Pritchard, R. E., Potter, G. C., Saccucci, M. S., 2004. The selection of a business major: Elements influencing student choice and implications for outcomes assessment. *Journal of Education for Business* 79 (3), 152–156.

- Schweitzer, M. E., Cachon, G. P., 2000. Decision bias in the newsvendor problem with a known demand distribution: Experimental evidence. *Management Science* 46 (3), 404–420.
- StepStone, 2016. Stepstone Gehaltsreport 2016 für Fach- und Führungskräfte.
- Sutter, M., Glätzle-Rützler, D., 2014. Gender differences in the willingness to compete emerge early in life and persist. *Management Science* 61 (10), 2339–23354.
- Train, Kenneth, E., 2009. *Discrete choice methods with simulation*. Cambridge University Press.
- Vij, A., Walker, J. L., 2016. How, when and why integrated choice and latent variable models are latently useful. *Transportation Research Part B: Methodological* 90, 192–217.
- Voleti, S., Srinivasan, V., Ghosh, P., 2016. An approach to improve the predictive power of choice-based conjoint analysis. *International Journal of Research in Marketing* <http://dx.doi.org/10.1016/j.ijresmar.2016.08.007>.
- Walker, J., 2001. *Extended discrete choice models: Integrated framework, flexible error structures and latent variables*. Ph.D. thesis, Massachusetts Institute of Technology.
- Walker, J., Ben-Akiva, M., 2002. Generalized random utility model. *Mathematical Social Sciences* 43 (3), 303–343.
- Wheeler, B., 2014. *AlgDesign: Algorithmic Experimental Design*. R package version 1.1-7.3.
URL <http://CRAN.R-project.org/package=AlgDesign>
- Wilder, C. R., Ozgur, C. O., 2015. Business analytics curriculum for undergraduate majors. *INFORMS Transactions on Education* 15 (2), 180–187.
- Winship, C., Mare, R. D., 1984. Regression models with ordinal variables. *American Sociological Review* 49 (4), 512–525.
- Wiswall, M., Zafar, B., 2015. Determinants of college major choice: Identification using an information experiment. *The Review of Economic Studies* 82 (2), 791–824.

Zafar, B., 2013. College major choice and the gender gap. *Journal of Human Resources* 48 (3), 545–595.

Zeid, M. A., 2009. Measuring and modeling activity and travel well-being. Ph.D. thesis, Massachusetts Institute of Technology.

2.3 Synthetic Data Sets with Non-Constant Substitution Patterns for Fare Class Choice

Synthetic Data Sets with Non-Constant Substitution Patterns for Fare Class Choice

BY FRAUKE SEIDEL, HAMBURG

1. Introduction

Synthetic data sets that are based on discrete choice models are applied in various research areas. A major field of study utilizing generated data focuses on the properties of newly developed discrete choice models and their predictive performance (Chiou and Walker, 2007). A prominent example is the Mixed Logit model whose development has led to an increase in studies applying synthetic data (Garrow et al., 2010). In addition to testing the performance of discrete choice models by applying synthetic data sets as done by Walker (2001, pp. 57), generated data is also applied to verify estimation results obtained by new estimators. In this context, Bierlaire et al. (2008) provide a study based on the comparison of two estimators for choice based samples. Synthetic data also provides the basis for evaluating the process of data generation itself. For instance, Garrow et al. (2010) compare three methodologies for generating such data and offer recommendations based on their empirical findings. Another option for applying generated data occurs when real data is not available. In mathematical optimization models disaggregate choice decisions from synthetic data can be utilized to represent demand for a certain product or service. For example, in revenue management in the airline industry the seat inventory control provides a solution to whether a seat in a fare class is offered to a passenger for a certain price (Andersson, 1998). Therefore, demand data is needed. In a discrete choice context this requires the generation of utility functions at the level of the decision maker, i.e. the individual. According to its respective definition, the deterministic part of utility may contain variables like travel cost, in-vehicle travel time, out- of-vehicle travel time and distance as well as income, gender and trip purpose (Williams and Ortuzar, 1982). When applying synthetic data these values are generated using probability distributions that can be verified by real data. The stochastic part of utility is then generated according to the assumptions the modeler makes about the underlying choice behavior of the generated population. Specifically in airline revenue management accurate data on fare class choice decisions is not available or is lacking important information (Hess et al., 2010).

Anschrift der Verfasserin:

Dipl.-Verk.wirtsch. Frauke Seidel
Universität Hamburg
Fakultät für Betriebswirtschaft
Institut für Verkehrswirtschaft
Von-Melle-Park 5
20146 Hamburg
email: frauke.seidel@uni-hamburg.de

To fill this gap, this paper provides a methodology to generate synthetic data for this purpose. Thereby it is assumed that the provided approach is valid for the case of a single flight with fixed capacity between a certain city pair. Furthermore, the airline as a monopolist is able to differentiate the fares offered to customers on that particular flight. This construct is known as price discrimination (Talluri and Van Ryzin, 2005, pp. 352-363). For data generation we assume in the following, that the utility a passenger receives from choosing a particular fare is dependent on product attributes, individual characteristics as well as factors that are not observed by the modeler. These unobserved factors are assumed to be correlated over alternatives and to have an important impact on demand by causing non-constant substitution between these fare classes (Berry, 1994). Insights on the methodology used to generate individual discrete choice utility values with constant and non-constant substitution patterns are given in section 2. Assumptions regarding the attributes of the considered alternatives as well as the characteristics of the generated population are provided in section 3. Furthermore, estimation results and elasticities are presented proving that the assumed substitution patterns are well recovered by the generated data.

2. Modeling Framework

Our considerations regarding the demand model for fare class choice are based on the theoretical framework of random utility theory. According to Marschak (1960) a choice model derived under the assumption that a decision maker maximizes its personal utility is called a random utility model (RUM). Thereby, utility is a random variable from the researcher's perspective as some influences affecting the choice decision usually remain unknown. Discrete choice models belong to this category of models (McFadden, 1974).

In this context, we consider fare class choice in airline revenue management. Airline passengers indexed $n = 1, \dots, N$ are assumed to choose exactly one fare class f out of an individual choice set C_n . The number of choice alternatives in C_n is required to be finite and exhaustive. Furthermore, alternatives within the choice set are mutually exclusive (Train, 2009, p. 15).

2.1. Utility and Decision Rule

Passengers are assumed to evaluate each fare class according to fare class attributes. Choices are further influenced by passenger characteristics. Each passenger n receives a certain utility u_{nf} of choosing fare class f . Utility u_{nf} is decomposed into a deterministic part v_{nf} and a stochastic part ε_{nf} and is formally defined as

$$u_{nf} = v_{nf} + \varepsilon_{nf} \quad (2.1)$$

with

$$v_{nf} = \sum_k \beta_{fk} z_{nfk}. \quad (2.2)$$

Pursuant to random utility theory a decision maker chooses the alternative with highest utility. Hence, the decision rule for the fare class choice problem can be stated as follows: A passenger n chooses fare class f only if $u_{nf} > u_{nf'} \forall f' \neq f$ (McFadden, 2001; Ben-Akiva and Lerman, 1985, p. 101). The probability of choosing fare class $f \in C_n$ over fare class f' is then defined by

$$\begin{aligned} P_{nf} &= \text{Prob}(u_{nf} \geq u_{nf'}, \forall f' \in C_n, f' \neq f) \\ &= \text{Prob}(v_{nf} + \varepsilon_{nf} \geq v_{nf'} + \varepsilon_{nf'}, \forall f' \in C_n, f' \neq f) \\ &= \text{Prob}(\varepsilon_{nf'} \leq v_{nf} - v_{nf'} + \varepsilon_{nf}, \forall f' \in C_n, f' \neq f) \end{aligned} \quad (2.3)$$

Given a specific assumption about the joint distribution of the stochastic utility component any choice model can be derived from equation 2.3 (Ben-Akiva and Lerman, 1985, p. 101). Thus, the specification of the joint distribution of ε_{nf} differs according to the choice model the researcher believes best represents the underlying choice situation.

2.2. Generation of Stochastic Utility

In a synthetic data set the stochastic utility component is generated such that the assumed behavioral process can be represented by the choice model that is considered for the data generation process. In simulation this approach is known as input modeling. Hence, to generate data that complies with our assumptions regarding the behavioral process for fare class choice we consider probability distributions according to a Multinomial Logit (MNL) and Nested Logit (NL) model for the generation of ε_{nf} . The disturbances of the discrete choice utilities are simulated by applying a pseudo random number generator (Garrow et al., 2010; Rosenthal, 2004).

An alternative approach to generate synthetic discrete choice data sets includes the utilization of NL choice probabilities to determine chosen alternatives. According to Garrow et al. (2010) the approximation of Gumbel distributed random variables with normals should be avoided as it yields biased data sets that do not reflect the desired behavioral model.

2.2.1 MNL Errors

The MNL model is derived from equation 2.3 if we assume that ε_{nf} is independently and identically (iid) type I¹ extreme value (EV) distributed with location parameter η and scale parameter μ (Train, 2009, p. 38). Thus, MNL error terms are distributed with density

$$f(\varepsilon_{nf}) = \mu e^{-\mu(\varepsilon_{nf}-\eta)} e^{-e^{-\mu(\varepsilon_{nf}-\eta)}} \quad (2.4)$$

¹ The type I extreme value distribution is also referred to as Gumbel distribution (Coles et al., 2001, pp. 46-48).

and cumulative distribution

$$F(\varepsilon_{nf}) = e^{-e^{-\mu(\varepsilon_{nf}-\eta)}}. \quad (2.5)$$

Let $F(\varepsilon_{nf}) = d$ be the probability of retrieving a draw that is equal or below ε_{nf} with d being a number between zero and one. Then, we can define $F(\varepsilon_{nf}) = \delta$ with δ being a draw of the standard uniform distribution. By solving for ε_{nf} we obtain a draw from distribution 2.5 as $\varepsilon_{nf} = F^{-1}(\delta) = \frac{1}{\mu} (-\ln(-\ln(\delta))) + \eta$ for decision maker n and a fare class f (Train, 2009, pp. 209-210). The inverse cumulative distribution of equation 2.5 is denoted by $F^{-1}(\delta)$ and is also called the quantile function $Q(\delta)$ (Gilchrist, 2000, pp. 12-14). The cumulative distribution function (CDF) $F(\cdot)$ is always invertible in a unique way if the argument is univariate and the corresponding probability density is nonzero.

The value of μ in the MNL model is arbitrary as it only sets the scale of the utilities. Thus, for convenience μ is usually chosen to equal one (Ben-Akiva and Lerman, 1985, p. 71). Without loss of generality, the location parameter is assumed to be $\eta = 0$ if a full set of alternative-specific constants (i.e., $|C| - 1$ constants in the considered fare class choice problem with $C = \bigcup_n C_n$) is included in the choice model (Hunt, 2000).

2.2.2 Substitution Patterns

By definition, the MNL model is not able to capture correlations between alternatives as the unobserved utility components for different alternatives are unrelated (Train, 2009, p. 39). Fare classes on a single flight are defined as differing products exhibiting several combinations of travel restrictions as well as differing prices. They are distinguished by compartment (first, business and economy) and are further characterized by additional benefits customers gain beyond the actual flight between an origin and destination. Within each of the mentioned compartments fare classes exhibit similar characteristics like advance purchase requirements, length-of-stay requirements, rebooking and cancellation penalties, the possibility to upgrade, the possibility to collect frequent flyer miles and many more. More complex combinations of restrictions imposed on a fare class result in lower prices (Talluri and Van Ryzin, 2005, pp. 521). Restrictions, thus, provide a necessary fencing between low and high fare products to prevent certain customers (i.e., business travelers) from buying down to a cheaper fare class (Zhang and Bell, 2010).

A buy down occurs when a customer who is willing to purchase a high fare product in the first place actually chooses a discount fare when both products are available. Thus, fencing serves as a justification for the disregard of up sell and down sell between fare classes. In particular, high fare customers are discouraged from purchasing low fare tickets as fare restrictions reduce the attractiveness of cheaper fares (Fiig et al., 2010).

As various fares on a single flight provide customers with similar restrictions we suppose

that dependencies in demand between alternatives with common characteristics exist. Hence, MNL substitution patterns that are constant between alternatives represent an inappropriate assumption for the considered fare class choice problem.

However, in airline revenue management research it is common practice to assume that demand for alternatives offered at the same time is independent. This is also known as the independent demand assumption. Thereby, demand for each fare class is supposed to be an independent stochastic process that is not influenced by the availability of other alternatives. An endogenization of customer behavior is not considered in the independent demand model (Talluri and Van Ryzin, 2005, p. 301).

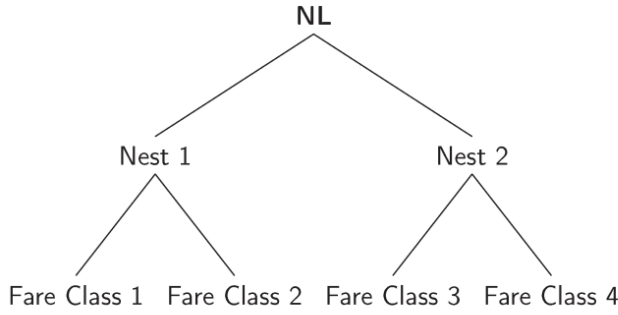
We presume, in the following, that correlation between fare classes is caused by the above mentioned fare attributes that are not included in the deterministic utility component v_{nf} . According to Berry (1994) these unobserved factors have important influence on demand as they lead to non-constant substitution patterns.

To account for the supposed demand dependencies between available alternatives our approach focuses on the generation of synthetic data based on a discrete choice model with more flexible substitution patterns. The NL model, for instance, is able to account for correlation in unobserved factors of alternatives and provides a more realistic representation of choice behavior.

2.2.3 NL Errors

Individual choices that comply with an NL model can be derived by assuming that the stochastic utility components of equation 2.3 follow a generalized extreme value (GEV) distribution (Train, 2009, pp. 80-81). For further details on the derivation of the NL choice probabilities see McFadden et al. (1978).

Figure 1 – NL nesting structure for a two-level NL model with four alternatives grouped into two nests.



Source: Own illustration

The assumption stated above, allows the grouping of alternatives that share common unobserved attributes into $m = 1, \dots, M$ non-overlapping nests. Thus, the stochastic utility component ε_{nf} of equation 2.1 can be decomposed into a nest specific term ε_{nm} that is the same for all alternatives in nest m and an alternative specific term ε_{nfm} that is independent across all alternatives (Bhat, 1996). The sum of both disturbances again has the same variance as the disturbance of the MNL model (Ben-Akiva and Lerman, 1985, p. 287). A general example on the nesting of four alternatives in a two-level NL model is given in figure 1. Although, the choice decision according to an NL model is not a hierarchical process we distinguish between the upper (nest) level and lower (alternative) level choice decision to derive the required terms for the data generation procedure.

Individual utility for choosing an alternative according to the NL model is obtained as

$$u_{nf} = v_{nf} + \varepsilon_{nm} + \varepsilon_{nfm} \quad \forall n, f \in C_{nm}|m \quad (2.6)$$

where C_{nm} denotes the choice set of individual n for a given nest m . The error terms ε_{nfm} are iid Gumbel distributed with scale parameter μ_m whereas the distribution of ε_{nm} is not known (Garrow et al., 2010). The ε_{nfm} are generated according to the procedure for the MNL error terms as stated in section 2.2.1. The scale parameter μ_m hereby describes the variances of the unobserved effects of utility u_{nf} on the lower level of the nesting structure. Thus, for all alternatives in the same nest m the scale parameter μ_m is identical. Alongside

the decomposition of the total error term ε_{nf} in equation 2.6 we also consider a compound error term for the generation of stochastic utility in the NL model as

$$\bar{\varepsilon}_{nf} = \varepsilon_{nm} + \tilde{\varepsilon}_{nfm} \quad \forall n, f \in C_{nm} | m \quad (2.7)$$

Thereby, ε_{nm} is the disturbance associated with the choice decision of an individual n on the upper level of the choice problem while $\tilde{\varepsilon}_{nfm}$ is the disturbance of the maxima of the individual utilities associated with the lower level choice decision. For each individual, the choice decision on the lower level is determined by the maximum of the utility values associated with the available alternatives. In the following, this maximum is denoted by \tilde{u}_{nf} .

The compound error $\bar{\varepsilon}_{nf}$ from equation 2.7 is non-independently and identically Gumbel distributed with scale parameter μ (Hunt, 2000; Silberhorn et al., 2008).

In an NL model formulation only the ratio of the two scale parameters μ/μ_m can be identified from the data. Therefore, the scale of utility is set by normalizing one of the scale parameters to one. The decision as to which parameter is to normalize is arbitrary as either possibility results in the same model (Ben-Akiva and Lerman, 1985, p. 287; Hensher and Greene, 2002). For the sake of generalization we will in the following illustrate the generation of the NL error terms by explicitly considering both scale parameters within the formal representations. Thus, the data generation process can be easily reproduced regardless of the normalization applied by the modeler.

The difficulty in generating a data set with the desired NL correlation structure lies in the disturbances ε_{nm} that are distributed such that the maxima of the individual utility values, \tilde{u}_{nf} , are iid Gumbel distributed with scale parameter μ . This is an indirect conclusion as the distribution of ε_{nm} is unknown. However, it can be obtained from the information about the mean value and variance of the compound error $\bar{\varepsilon}_{nf}$ and the independent errors of each individuals maximum utility $\tilde{\varepsilon}_{nfm}$ (Garrow et al., 2010). Therefore, in the following we utilize the relation of these error terms as given by equation 2.7.

In general, the mean value and variance of an iid type I EV random variable X with location parameter η and scale μ are formally given by

$$E(X) = \eta + \frac{\gamma}{\mu} \quad (2.8)$$

and

$$Var(X) = \frac{\pi^2}{6\mu^2} \quad (2.9)$$

with γ being Euler's constant.

According to Ben-Akiva and Lerman (1985, pp. 104-105) the maximum of $|C|$ iid Gumbel distributed random variables (i.e.; $\tilde{\varepsilon}_{nfm}$) with location and scale parameters $(\eta_1, \mu_m), (\eta_2, \mu_m), \dots, (\eta_F, \mu_m)$ is also iid Gumbel distributed with parameters

$$\left(\frac{1}{\mu_m} \ln \sum_{f \in C_{nm}} e^{\mu_m \eta_f}, \mu_m \right). \quad (2.10)$$

Furthermore, the variance of the independent error term is

$$\text{Var}(\varepsilon_{nfm}) = \frac{\pi^2}{6\mu_m^2} \quad (2.11)$$

and since $\eta_f = 0$, together with 2.8 and 2.11, its mean value is

$$E(\varepsilon_{nfm}) = \frac{\gamma}{\mu_m}. \quad (2.12)$$

Following Hunt (2000) the location parameter of an iid Gumbel distributed random variable can be set to zero, as any nonzero location parameter is eliminated by an alternative-specific constant. Assuming a full set of constants in our choice model we can set $\eta_f = 0$. Hence, the location parameter of the distribution of the maximum values of the independent disturbances $\tilde{\varepsilon}_{nfm}$, becomes

$$\tilde{\eta} = \frac{1}{\mu_m} \ln |C_{nm}|. \quad (2.13)$$

Substituting 2.13 in 2.8, we derive the mean value of the disturbance maxima as

$$E(\tilde{\varepsilon}_{nfm}) = \frac{1}{\mu_m} \ln |C_{nm}| + \frac{\gamma}{\mu_m}. \quad (2.14)$$

Furthermore, the variance of $\tilde{\varepsilon}_{nfm}$ equals the variance of ε_{nfm} :

$$\text{Var}(\tilde{\varepsilon}_{nfm}) = \frac{\pi^2}{6\mu_m^2}. \quad (2.15)$$

To derive the unknown distribution of ε_{nm} we need the mean value and the variance of the compound error from equation 2.7. Both values are defined as

$$E(\bar{\varepsilon}_{nfm}) = E(\varepsilon_{nm} + \tilde{\varepsilon}_{nfm}) = \frac{\gamma}{\mu} \quad (2.16)$$

and

$$\text{Var}(\bar{\varepsilon}_{nfm}) = \text{Var}(\varepsilon_{nm} + \tilde{\varepsilon}_{nfm}) = \frac{\pi^2}{6\mu^2}. \quad (2.17)$$

Having made the above definitions the derivation of the mean value and variance of ε_{nm} is now straightforward. From equations 2.14, 2.16, 2.15 and 2.17 we obtain the mean value and variance of ε_{nm} as

$$\begin{aligned} E(\varepsilon_{nm}) &= E(\bar{\varepsilon}_{nfm}) - E(\tilde{\varepsilon}_{nfm}) \\ &= \left(\frac{1}{\mu_m} \ln |C_{nm}| \right) + \left(\frac{\gamma}{\frac{\mu_m \mu}{\mu_m - \mu}} \right) \end{aligned} \quad (2.18)$$

and

$$\begin{aligned} Var(\varepsilon_{nm}) &= Var(\bar{\varepsilon}_{nfm}) - Var(\tilde{\varepsilon}_{nfm}) - 2Cov(\varepsilon_{nm} \tilde{\varepsilon}_{nfm}) \\ &= \frac{\pi^2}{6\mu^2} - \frac{\pi^2}{6\mu_m^2} - 0 \\ &= \frac{\pi^2}{6 \left(\frac{\mu_m^2 \mu^2}{\mu_m^2 - \mu^2} \right)}. \end{aligned} \quad (2.19)$$

As mentioned before, the independent error term ε_{nfm} from the NL utility function 2.6 is Gumbel distributed with location parameter 0 and scale parameters μ_m . As the scale parameters μ_m are predetermined by the modeler the disturbance is obtained according to the generation of the MNL errors as described in section 2.2.1. From equations 2.18 and 2.19 we get the location and scale parameter of the nest specific error terms ε_{nm} :

$$\left[\underbrace{- \left(\frac{1}{\mu_m} \ln |C_{nm}| \right)}_{\text{location parameter}}; \underbrace{\sqrt{\left(\frac{\mu_m^2 \mu^2}{\mu_m^2 - \mu^2} \right)}}_{\text{scale parameter}} \right]. \quad (2.20)$$

The NL disturbances can now be generated by combining the inverse of the Gumbel cumulative distribution function with the parameters from 2.20:

$$\begin{aligned} \varepsilon_{nm} &= F^{-1}(\delta) = \frac{1}{\mu} (-\ln(-\ln(\delta))) + \eta \\ &= \frac{1}{\left(\frac{\mu_m^2 \mu^2}{\mu_m^2 - \mu^2} \right)} (-\ln(-\ln(\delta))) - \left(\frac{1}{\mu_m} \ln |C_{nm}| \right) \end{aligned} \quad (2.21)$$

with δ being a uniform [0, 1] distributed random variable.

2.2. Generation of Systematic Utility

For the generation of the deterministic part of utility v_{nf} , we consider $k = 1, \dots, 3$ observable attributes of the alternatives as well passenger characteristics. According to equation 2.2 coefficients β_{fk} provide a weighting regarding the influence of each observed attribute z_{nfk} on deterministic utility v_{nf} . Let

$$v_{nf} = \beta_{f0} + \beta_1 \cdot z_{nf1} + \beta_{f2} \cdot z_{n2} + \beta_{f3} \cdot z_{n3} \quad \forall n, f \quad (2.22)$$

be the function of the deterministic part of individual utility for the fare class choice problem.

2.3.1 Attributes and Characteristics z_{nfk}

According to equation 2.22 the value of v_{nf} depends on

- the price paid by passenger n for a ticket in fare class f - z_{nf1} ,
- passengers gender - z_{n2} and
- the trip purpose of passenger n - z_{n3} .

As for the error terms, values for these variables are again generated by making assumptions about their distribution across the synthetic population. Characteristics gender and trip purpose, are dummy variables and can easily be generated by drawing from a uniform $[0,1]$ distribution. Prices for each passenger and fare class are assumed to be truncated normally distributed random variables.

2.3.2 Coefficients β_{fk}

Each of the considered attributes z_{nfk} is weighted by a coefficient β_{fk} . For existing real data on a specific choice problem these coefficients are determined by estimation. Hence, in a data generation context these coefficients are chosen to ensure that the data set reflects the assumptions regarding the influence of each attribute and characteristic on individual utility. Furthermore, it is important that neither the deterministic part nor the stochastic part of utility dominates the overall utility value u_{nf} (Munizaga et al., 2002). Therefore, the values of the two utility components of equation 2.1 and their respective influence on the overall value of utility u_{nf} have to be verified and adjusted prior to data generation.

Further details on the specification of the values of the variables z_{nfk} and the coefficients β_{fk} are provided in section 3.

3. Fare Class Choice

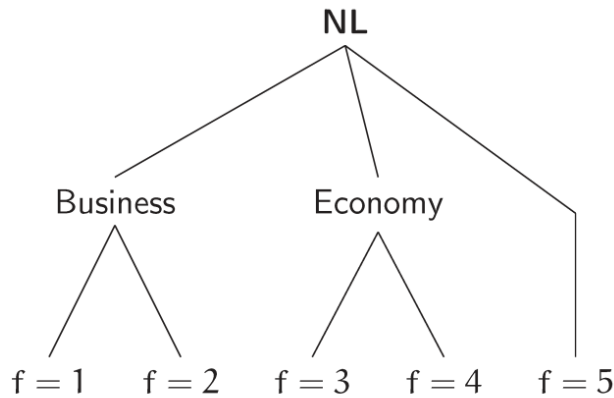
Since the liberalization of the airline market in the early 1970s airlines have started to utilize price discrimination for differentiated products as an instrument to maximize revenues (McGill and Van Ryzin, 1999; Anderson et al., 1992, pp. 1-5). Based on this, we suppose that the considered fare class choice problem is described by the choice decision between four ticket fares and a no choice option. We act on the assumption that the following alternatives exist:

- a regular ($f = 1$) and a discount fare ($f = 2$) in business class,
- a regular ($f = 3$) and a discount fare ($f = 4$) in economy class, and
- a no choice option ($f = 5$).

The no choice option expresses the decision of a potential customer to not buy a ticket in a certain fare class at all. It further ensures that the choice set of each generated individual is realistic and exhaustive. In the presented approach the alternative $f = 5$ serves as the reference alternative of the choice model. Furthermore, it is not assigned an attribute value for fare class price as the decision of not choosing an alternative is assumed to not impose any cost on a particular passenger. The functions of deterministic utility for a passenger n are defined as follows:

$$\begin{aligned}
 V_{n1} &= \beta_{10} + \beta_1 \cdot z_{n11} + \beta_{12} \cdot z_{n2} + \beta_{13} \cdot z_{n3} \\
 V_{n2} &= \beta_{20} + \beta_1 \cdot z_{n21} + \beta_{22} \cdot z_{n2} + \beta_{23} \cdot z_{n3} \\
 V_{n3} &= \beta_{30} + \beta_1 \cdot z_{n31} + \beta_{32} \cdot z_{n2} + \beta_{33} \cdot z_{n3} \\
 V_{n4} &= \beta_{40} + \beta_1 \cdot z_{n41} + \beta_{42} \cdot z_{n2} + \beta_{43} \cdot z_{n3} \\
 V_{n5} &= 0
 \end{aligned}$$

Figure 1 - Assumed nest structure for utility generation in a fare class choice demand model.



Source: Own illustration

As already stated in the previous section fare class attributes for the remaining alternatives as well as passenger characteristics have to be defined.

Thereby, prices z_{nf1} in € for each passenger n and fare class f are assumed to be normally distributed $N(\mu, \sigma)$ random variables with mean μ and standard deviation σ :

- $\mu = 1000, \sigma = 200$ for $f=1$,
- $\mu = 800, \sigma = 150$ for $f=2$,
- $\mu = 400, \sigma = 100$ for $f=3$,
- $\mu = 200, \sigma = 50$ for $f=4$.

The data set is furthermore generated such that 47% of the passengers of the synthetic population are female and 53% male, respectively. Additionally, the population can be segmented in passengers traveling for leisure or business purposes. Thereby, passengers with leisure trips represent 72% of the population and passengers with business trips the remaining 28% (Brey and Walker, 2011).

As for the coefficients β_{fk} , we have to make sure that the assumptions made about the influence of the attributes and characteristics on the deterministic utility of each alternative are reflected by the synthetic data. The corresponding coefficient values that are applied in the data generation process are displayed in column 'true value' in table 1.

Besides true coefficient values the estimates for both an NL model and an MNL model are compared. Estimates are obtained from the synthetic data set with NL errors.

By setting the true coefficient values, we first assume that the alternative specific constants (ASC) β_{fk} provide the desired market shares of the fare classes considered in our choice model. Second, price sensitivity is included in the behavioral model by assuming that a higher price decreases the utility value an individual receives from choosing a certain fare class. This is achieved by assuming a negative price coefficient β_1 that is the same for all alternatives. This is known as generic specification (Garrow et al., 2007).

Table 1 - True and estimated coefficients of the synthetic data set with confidence levels ***= 99%, ** = 95% and * = 90%.

Coefficient		NL			MNL		
β_{fk}	True value	Estimate	t-statistic against zero	t-statistic against true value	Estimate	t-statistic against zero	t-statistic against true value
ASC							
β_{10}	0.50	0.065	0.27	-1.81*	0.348	1.33	-0.58
β_{20}	1.50	1.560	8.34***	0.32	2.310	16.14***	5.66***
β_{30}	1.60	1.420	20.08***	-2.24***	1.160	13.77***	-5.24***
β_{40}	2.00	1.820	31.66***	-3.14***	1.980	38.54***	-0.39
β_{50}	0.00	0.00	-	-	-	-	-
Price							
β_1	-0.0040	-0.00427	-17.55***	-1.11	-0.00535	-30.78***	-7.76***
Gender							
β_{12}	0.80	0.899	5.47***	0.60	0.930	4.58***	0.64
β_{22}	0.50	0.561	6.51***	0.71	0.537	6.47***	0.42
β_{32}	0.20	0.184	3.01***	-0.26	0.299	4.27***	1.41
β_{42}	-0.10	-0.0934	-1.83*	0.13	-0.124	-2.40***	-0.46
β_{52}	0.00	0.00	-	-	-	-	-
Trip Purpose							
β_{13}	2.00	2.290	12.53***	1.58	2.560	12.18***	2.67***
β_{23}	1.50	1.380	14.84***	-1.29	1.360	14.26***	-1.47
β_{33}	1.00	1.070	14.15***	0.93	1.280	16.08***	3.53***
β_{43}	0.50	0.557	8.60***	0.88	0.496	7.56***	-0.06
β_{53}	0.00	0.00	-	-	-	-	-
Nest Coefficients² μ_m							
μ_1	1.80	1.40	8.75***	-2.50***			
μ_2	1.60	1.570	13.76***	-0.26			
μ_3	1.00	1.00	-	-			

² The value of the t-statistic as displayed for the nest coefficients tests the null hypothesis that $\mu_m = 1$ for all m .

Source: Own calculations

Finally, coefficients β_{f2} and β_{f3} provide a weighting of the socio-economic characteristics gender and trip purpose that are both included in the choice model as dummy variables. As both characteristics do not vary across alternatives the corresponding coefficients are defined alternative-specific. We further assume male business ($z_{nf2} = 1$, $z_{nf3} = 1$) travelers to receive higher utility from choosing a business fare over economy. Buying a ticket in business class is, due to better seat comfort and less restrictive regulations regarding re-

booking and cancellation, supposed to have a positive effect on utility for male passengers as well as business travelers.

The latter are generally considered relatively price-insensitive (Talluri and Van Ryzin, 2005, pp. 516-517) while leisure travelers, in particular, are found to have a higher price-sensitivity than business travelers (Garrow, 2010, pp. 18-19).

In equation 2.22 that represents the functional form of utility for our fare class choice problem, we do not account for fare flexibility, amenities like lounge access, seating on board or preferences of customers associated with business or economy class within the specification of v_{nf} . However, these attributes have an important influence on the utility value an individual receives from choosing a particular alternative as well as on substitution patterns between alternatives.

As outlined in section 2.2.3 we assume that these unobserved effects are completely captured by ε_{nf} leading to correlation in alternatives with similar restrictions.

Hence, the alternatives of the considered fare class choice problem are assigned to $m = 1, \dots, 3$ nests in the following way:

- Nest 1: $f = \{1, 2\}$
- Nest 2: $f = \{3, 4\}$
- Nest 3: $f = \{5\}$.

The corresponding nesting structure is displayed in figure 2. It reflects the assumption that business fares and economy fares are closer substitutes among each other than are fares from other nests. Substitution patterns in the NL model are by definition constant between fare classes in the same nest but not constant across nests. Hence, the independence of irrelevant alternatives (IIA) assumption holds within each nest but not in general for alternatives that are assigned to different nests (Train, 2009, p.81). This is justified by the fact that more price sensitive leisure passengers book their trips way in advance and tend to choose the least expensive option. On the contrary, business travelers are known to choose fares that allow for additional amenities at the airport and on board. As business travelers tend to be more time sensitive they also prefer the possibility to cancel or rebook a flight on a short notice if appointments change (Garrow et al., 2007).

To achieve the desired correlation structure for our fare class choice problem, the nest coefficients μ_m for $m = 1, \dots, 3$ are provided for the generation of the NL error term. These coefficients allow for alternatives within the same nest to be closer substitutes than alternatives from different nests resulting in flexible substitution patterns. They are the same for all alternatives in one nest. As alternative $f = 5$ is solely assigned to the third nest, leading to a degenerate nesting structure, we choose $\mu_3 = 1$ for identification purposes.

3.1. Data Sets

As proposed by Garrow et al. (2010) a synthetic data set with 10.000 observations and correlation structure as proposed in the previous section is generated. Both an NL model and an MNL model are estimated from the data. We hereby assume that the behavioral process according to an NL model represents the true choice behavior of airline passengers. Estimates as displayed in table 1 are obtained by estimation with BIOGEME version 2.2 (Bierlaire, 2003). Besides true and estimated coefficients β_{fk} for both the NL and MNL models, the corresponding t-statistics against zero and significance levels are provided. Except for the NL and MNL estimate of the alternative-specific constant β_{10} the null hypothesis is rejected at the 99% level of confidence for the remaining estimates.

Regarding the nest coefficients the null hypothesis of the t-statistic is $\mu_m = 1$ for all nests m confirming that the nest coefficients are statistically significant on a 99% level of confidence.

Column '*t-statistic against true value*' displays the value of the t-statistic for each estimate and nest coefficient against its corresponding true coefficient value. Results show that the true coefficients provided for data generation are well recovered for the NL model. However, we have to reject the null hypothesis that NL estimates β_{30} and β_{40} as well as nest coefficient μ_1 equal their respective true value on a 99 % level of confidence. Although, we cannot be sure whether the estimated coefficients are equal to the true coefficients we can confirm, by the corresponding t-tests against zero, that all three estimates have significant influence on the individual utility values.

As the MNL estimates are obtained from a data set with an NL error structure the value of the t-statistic against the true coefficient value in the MNL yields that the null hypothesis has to be rejected for β_{20} , β_{30} , β_1 , β_{13} and β_{33} on a 99% level of confidence.

Nest coefficients μ_1 , μ_2 and μ_3 are only obtained for the NL model and reflect the degree of correlation between alternatives within the same nest. Correlation for any pair of alternatives in a common nest can be determined by calculating $\kappa_m = 1 - \left(\frac{\mu}{\mu_m}\right)$ with $m = \{1,2,3\}$. Thereby, as proposed in section 2.2, μ is the upper level scale parameter of the NL model and is set to one for identification purposes (Ben-Akiva and Lerman, 1985, p. 287). The correlation matrix associated with our fare class choice problem is

$$\begin{pmatrix} 1 & \kappa_1 & 0 & 0 & 0 \\ \kappa_1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & \kappa_2 & 0 \\ 0 & 0 & \kappa_2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad (3.1)$$

with $\kappa_1 = 0.49$ and $\kappa_2 = 0.59$ being the correlation coefficients of nest 1 and nest 2,

respectively. As $\mu_3 = 1$ we obtain $\kappa_3 = 0$.

3.2. Market Shares

According to the definition of the error terms in section 2.2 we assume demand shifts to be proportional in the MNL model and non-proportional in the NL model between alternatives in different nests, if a fare class is closed. The following analysis of substitution patterns is based on an exemplary comparison of each alternatives' market shares (MS) when

- (i) all alternatives are available
- (ii) the cheapest fare class is closed.

In airline revenue management optimization models similar decisions are part of the seat inventory control aiming on revenue maximization.

Table 2 - MNL market shares in % for situations (i) and (ii).

Fare Class	1	2	3	4	5
Market Share (i)	1.25	7.59	14.60	55.37	21.19
Market Share (ii)	2.54	16.02	31.44	-	50.00

Source: Own calculations

For the simulation of market shares for both situations we apply BIOSIM (Bierlaire, 2003). The market shares corresponding to the MNL estimates of the NL data set for the above defined situations are displayed in table 2. Clearly, the most chosen option in situation (i) is the inflexible economy tariff (alternative 4) and, as expected, the most expensive business fare (alternative 1) is the least chosen option. Now we compare our findings with situation (ii) where only four alternatives remain in the choice set of the individuals. We assume that due to a decision based on a revenue management model fare class four is closed. Thus, ca. 55% of the generated individuals are not able to purchase their first choice and switch to other available alternatives. In table 2 we see that all alternatives gain from the closure of fare class four. The initial market shares of all remaining fare classes increase by approximately 100% in situation (ii). Furthermore, we are able to observe whether the occurring demand shifts from the closed option towards the remaining alternatives are constant by calculating the ratio of substitution for any pair of alternatives. The substitution ratio for two alternatives is then obtained by dividing the respective market shares. According to the IIA assumption of the MNL we expect substitution patterns to be constant for situations (i) and (ii) (Train, 2009, pp. 49-51). Using the example of alternatives one, two and three we obtain the following substitution ratios (SR) for situations (i) and (ii):

$$SR_{13}^{MNL}(i) = \frac{MS_{f=1}}{MS_{f=3}} = \frac{1.25}{14.60} = 0.08$$

$$SR_{13}^{MNL}(ii) = \frac{MS_{f=1}}{MS_{f=3}} = \frac{2.54}{31.44} = 0.08$$

$$SR_{23}^{MNL}(i) = \frac{MS_{f=2}}{MS_{f=3}} = \frac{7.59}{14.60} = 0.51$$

$$SR_{23}^{MNL}(ii) = \frac{MS_{f=2}}{MS_{f=3}} = \frac{16.02}{31.44} = 0.51$$

As expected, substitution patterns are constant if passenger choice behavior is assumed to be best represented by an MNL model. Thus, substitution ratios between all remaining fare classes do not change if a fare class is closed for purchase. This is exemplarily presented here by comparing $SR_{13}^{MNL}(i)$ with $SR_{13}^{MNL}(ii)$ and $SR_{23}^{MNL}(i)$ and $SR_{23}^{MNL}(ii)$. Of course, the same applies for the remaining ratios. However, if we assume that some fare classes share common unobserved attributes this approach does not seem to be correct.

Table 3 - NL market shares in % for situations (i) and (ii).

Fare Class	1	2	3	4	5
Market Share (i)	1.25	7.59	14.5	55.40	21.19
Market Share (ii)	2.54	13.53	43.53	-	40.81

Source: Own calculations

Hence, in the following we examine the case where individual choice behavior is assumed to follow an NL model. For that matter, the market shares are given in table 3. According to the assumptions made above we consider the nesting structure displayed in figure 2. Note, that the displayed NL market shares for situation (i) equal the MNL market shares for situation (i). These values reflect the true market shares of the generated data set as a full set of alternative-specific constants is considered both for data generation and estimation. After closing fare class four we examine demand shifts in the NL model by exemplarily calculating the ratios of substitution for alternatives one, two and three. Thereby, alternatives one and two are members of the same nest, while alternative three belongs to a different nest. By definition, substitution patterns are constant within nests and non-constant across nests. We obtain the following results:

$$SR_{12}^{NL}(i) = \frac{MS_{f=1}}{MS_{f=2}} = \frac{1.25}{7.59} = 0.16$$

$$SR_{12}^{NL}(ii) = \frac{MS_{f=1}}{MS_{f=2}} = \frac{2.12}{13.53} = 0.16$$

$$SR_{13}^{NL}(i) = \frac{MS_{f=1}}{MS_{f=3}} = \frac{1.25}{14.57} = 0.086$$

$$SR_{13}^{NL}(ii) = \frac{MS_{f=1}}{MS_{f=3}} = \frac{2.12}{43.53} = 0.048.$$

Clearly, substitution patterns in the NL model are as expected constant between alternatives one and two (SR_{12}^{NL}) that share a common nest. However, substitution patterns are not constant between alternatives one and three (SR_{13}^{NL}) that belong to different nests.

3.3. Elasticities

Besides market shares elasticities can also be obtained for the synthetic data sets. Elasticities represent the responsiveness of passengers to a change in a certain attribute. As in microeconomic consumer theory the price for a ticket in fare class f is the only attribute of the alternatives in the fare class choice problem. Hence, the following explanations refer to the responsiveness of the passengers of both populations regarding a change in ticket price. Therefore, the only relevant elasticity is the price elasticity. Furthermore, rather than examining disaggregate price elasticities we focus on the corresponding aggregate values (Ben-Akiva and Lerman, 1985, pp. 111-113). Elasticities are calculated based on the true coefficient values provided for data generation. Furthermore, substitution patterns according to the respective choice model are applied to obtain the responsiveness for both the NL and MNL model. The disaggregate direct and cross price elasticities for the MNL model are given by

$$\varepsilon_{z_{nfk}}^{P_{nf}} = [1 - P_{nf}]z_{nfk}\beta_k \quad (3.2)$$

and

$$\varepsilon_{z_{nf'k}}^{P_{nf}} = -P_{nf'}z_{nf'k}\beta_k \cdot \quad \forall f \neq f' \quad (3.3)$$

The values obtained by equations 3.2 and 3.3 refer to the responsiveness of an individual. In contrast, aggregate elasticities provide the responsiveness of some group of decision makers. They are the weighted averages of the individual level elasticities with weighting provided by the choice probabilities. Following Ben-Akiva and Lerman (1985, p. 113) the expected share of a group of decision makers is defined as

$$\bar{P}_{nf} = \frac{\sum_{n=1}^N P_{nf}}{N} \quad (3.4)$$

with N being the total number of decision makers within the respective group. Aggregate direct and cross elasticities for the MNL model are obtained by

$$\bar{\varepsilon}_{z_{nfk}}^{\bar{P}_{nf}} = \frac{\sum_{n=1}^N P_{nf} \varepsilon_{z_{nfk}}^{P_{nf}}}{\sum_{n=1}^N P_{nf}} \quad (3.5)$$

and

$$\bar{\varepsilon}_{z_{nf'k}}^{\bar{P}_{nf}} = \frac{\sum_{n=1}^N P_{nf} \varepsilon_{z_{nf'k}}^{P_{nf}}}{\sum_{n=1}^N P_{nf}} \quad (3.6)$$

Disaggregate and aggregate elasticities for the NL model are obtained in a similar way. For further information on the calculation of NL elasticities we refer to Koppelman and Bhat (2006, pp. 163-165).

Table 4 - NL and MNL aggregate direct price elasticities

Fare Class	1	2	3	4
NL	-2.348	-2.433	-0.743	-0.265
MNL	-3.234	-2.621	-1.129	-0.359

Source: Own calculations

Aggregate elasticities, as considered in the following, refer to the generated population as a whole. Table 4 shows the aggregate direct elasticities for both the MNL and NL model. Price sensitivity clearly seems to be less distinct in the NL model. We suspect this to be a result of the differing definition of stochastic utility as deterministic utility is identical for both models according equation 2.22. Furthermore, price changes in business class fares, as assumed, have a much higher influence on choice probabilities than price changes in economy fares. This conclusion holds for both the MNL and the NL model.

Table 5 - MNL aggregate cross price elasticities.

Fare Class	1	2	3	4
1		0.257	0.373	0.388
2	0.092		0.353	0.399
3	0.074	0.198		0.408
4	0.062	0.179	0.325	
5	0.056	0.170	0.324	0.438

Source: Own calculations

Aggregate values of the cross price elasticities are also obtained. Cross price elasticities reflect the influence on the choice probability of an alternative when the price attribute of another alternative is changed. Thereby, disaggregate cross elasticities of the MNL are uniform, i.e. equal for all alternatives $f \neq f'$ that are affected by the attribute change of alternative f (Ben-Akiva and Lerman, 1985, pp.111-113).

Tables 5 and 6 display aggregate cross price elasticities that are attained for changes in the price of one fare class and the corresponding impact on the choice probabilities of all other fare classes.

Table 6 - NL aggregate cross price elasticities.

Nest	Fare Class	1	2	3	4
1	1		1.211	0.282	0.406
1	2	0.0273		0.248	0.423
2	3	0.061	0.233		0.762
2	4	0.044	0.198	0.374	
3	5	0.040	0.190	0.209	0.467

Source: Own calculations

Thereby, the fare class number stated in the head of each column represents the fare class where a price change occurs. The corresponding elasticities of all other fare classes are displayed in the rows below the respective column.

For interpreting the results we have to keep the substitution patterns of the MNL and NL model in mind. The values of the responses to this price change are similar, though not exactly equal through aggregation, over all remaining fare classes. This indicates that a price change in the first fare class by one unit increases the choice probabilities of all other alternatives by approximately 0.1. The same way, price changes in the other fare classes are analyzed. Clearly, a change in price of the cheapest fare has the largest influence on choice probabilities. The responsiveness to a price change in a certain fare class results in similar changes of choice probabilities of all remaining fare classes. This finding again confirms constant substitution between fare classes if the behavioral process of fare class choice is assumed to be best represented by an MNL model.

Aggregate cross price elasticities of the NL model can be gathered from table 6. As substitution patterns are, by definition, not constant in the NL model the resulting NL cross price elasticities also differ from the ones obtained for the MNL model. For reasons of clarity, we add a column indicating the nest membership of each fare class. In the last four columns we have again the responses to a price change in a certain fare class. Obviously, elasticities of alternatives that are in the same nest with the alternative whose price is increased are larger in value than elasticities of alternatives that are in another nest. This indicates that substitution between fare classes within the same nest is more likely than substitution between alternatives that belong to different nests confirming the existence of the assumed correlation structure in NL error terms.

Both market shares and elasticities obtained for the generated data sets indicate that the consideration of demand dependencies for fare classes with similar characteristics may have an important impact on decisions regarding the seat inventory control in airline revenue management. Airline seat inventory control is an approach where seats are allocated to different fare classes such that revenues are maximized (Williamson, 1992, p. 28). This is done by deciding on the fare classes that are contained in each passenger's choice set. As stated above, in airline revenue management research it is common practice to assume that

demand for alternatives offered at the same time is independent and does not depend on the availability of other fare classes in the choice set of a certain passenger (Talluri and Van Ryzin, 2005, pp. 33-35). Thus, by assuming that demand is best represented by an MNL model demand shifts are assumed to be proportional which may lead to an erroneous estimation of revenues in seat inventory control. Therefore, demand with constant substitution patterns is not an appropriate approach when it comes to fare class choice. As outlined in section 2.2.2 different fare classes exhibit similar restrictions that may lead to non-proportional shifts in demand when fare classes are closed for purchase. Hence, to relax the independent demand assumption demand for the seat inventory control problem in airline revenue management should be represented by a discrete choice model with non-constant substitution patterns.

4. Conclusions

In this article, we have examined the generation of synthetic data sets for fare class choice with non-constant substitution patterns. We were able to show that true coefficients of the choice problem are well recovered and lead to desired substitution patterns. Furthermore, an analysis of market shares and elasticities proves that the generated NL data set exhibits the desired correlation structure. The results provide a basis for overcoming the independent demand assumption that is usually applied in revenue management optimization models. Considering today's possibilities of accurate demand modeling with discrete choice models this assumption clearly seems to be overrun. Although the MNL model already is frequently applied in many revenue management studies it is still lacking the possibility to account for demand dependencies. It is well known that fare class restrictions can be very complex depending on the fencing desired for a particular tariff. With up to 20 different fares on a single flight restrictions might overlap for at least a few fares requiring alternate approaches of demand modeling.

Thus, our findings imply the application of more flexible discrete choice models than MNL. Especially in a combined approach for seat allocation and pricing this allows for potential revenue gains as substitution patterns differ when demand is not assumed to be independent for fare classes with similar characteristics.

One aspect that is not addressed in this article are intertemporal substitutions. In this context, the approach provided in this article can be assumed to give a cross section of passenger's choices at a certain point in time. Hence, dependencies of choice decisions at different time points are not taken into account. However, it has to be assumed that potential passengers check seat availabilities of a desired flight for a certain period of time prior to an actual booking. Passenger's decisions to not book a ticket immediately but wait for a better offer in the future also influences seat availability within the offered fare classes in the days prior to the booking. Thus, modeling intertemporal substitution by applying dynamic choice models could be an interesting approach for future research. Thereby, intertemporal dependencies of passenger choices should be modeled such that recursivity and endogeneity

of passenger behavior are accounted for. Furthermore, the modeler has to keep in mind that passengers' preferences might change over time leading to dynamic inconsistency.

Abstract

The article provides a theoretical framework for the generation of synthetic discrete choice datasets for fare class choice in airline revenue management. The necessity of this research arises from simplifying assumptions regarding demand modeling that are commonly made in airline revenue management optimization models. By applying demand models with more flexible substitution patterns, like the Nested Logit (NL) model, it is possible to account for dependencies between offered fare products and their influence on airline revenue.

Acknowledgements

The author would like to thank an anonymous reviewer and the editor for their very helpful contribution on a first version of this article.

REFERENCES

- Anderson, S., de Palma, A., & Thisse, J.F. (1992), *Discrete Choice Theory of Product Differentiation*, MIT Press, Cambridge, MA.
- Andersson, S.E. (1998), Passenger Choice Analysis for Seat Capacity Control: A Pilot Project in Scandinavian Airlines, *International Transactions in Operational Research* 5(6), 471–486.
- Ben-Akiva, M. & Lerman, S. (1985), *Discrete Choice Analysis: Theory and Application to Travel Demand*, MIT Press, Cambridge, MA.
- Berry, S. T. (1994), Estimating Discrete-Choice Models of Product Differentiation, *The RAND Journal of Economics*, 242–262.
- Bhat, C. (1996), Covariance Heterogeneity in Nested Logit Models: Econometric Structure and Application to Intercity Travel. *Transportation Research Part B: Methodological* 31(1), 11–21.
- Bierlaire, M. (2003), BIOGEME: A Free Package for the Estimation of Discrete Choice Models. In *3rd Swiss Transport Research Conference*.

- Bierlaire, M., Bolduc, D. & McFadden, D. (2008), The Estimation of Generalized Extreme Value Models from Choice-Based Samples, *Transportation Research B*, 42(4), 381–394.
- Brey, R. & Walker, J.L. (2011), Latent Temporal Preferences: An Application to Airline Travel, *Transportation Research Part A: Policy and Practice*, 45(9), 880–895.
- Chiou, L. & Walker, J.L. (2007), Masking Identification of Discrete Choice Models under Simulation Methods, *Journal of Econometrics*, 141(2), 683–703.
- Coles, S., Bawa, J., Trenner, L. & Dorazio, P. (2001), *An Introduction to Statistical Modelling of Extreme Values*, Vol. 208, Springer.
- Fiig, T., Isler, K., Hopperstad, C. & Belobaba, P. (2010), Optimization of Mixed Fare Structures: Theory and Applications, *Journal of Revenue & Pricing Management* 9 (1), 152–170.
- Garrow, L., Bodea, T. & Lee, M. (2010), Generation of Synthetic Datasets for Discrete Choice Analysis. *Transportation*, 37, 183–202.
- Garrow, L. A. (2010), *Discrete Choice Modelling and Air Travel Demand: Theory and Applications*, Ashgate Publishing, Ltd.
- Garrow, L. A., Jones, S.P & Parker, R.A. (2007), How much Airline Customers are Willing to Pay: An Analysis of Price Sensitivity in Online Distribution Channels, *Journal of Revenue & Pricing Management*, 5 (4), 271–290.
- Gilchrist, W. (2000), *Statistical Modelling with Quantile Functions*, CRC Press.
- Hensher, D. A. & Greene, W.H. (2002), Specification and Estimation of the Nested Logit Model: Alternative Normalisations, *Transportation Research Part B: Methodological*, 36, 1-17
- Hess, S., Niemeier, H.-M., Forsyth, P., Müller, J. & Gillen, D. (2010), *Airport Competition: The European Experience*, Chapter: Modelling Air Travel Choice Behaviour, 151–176, Ashgate Publishing, Ltd.
- Hunt, G. L. (2000), Alternative Nested Logit Model Structures and the Special Case of Partial Degeneracy, *Journal of Regional Science* 40 (1), 89–113.
- Koppelman, F. S. & Bhat, C. (2006), A Self Instructing Course in Mode Choice Modeling: Multinomial and Nested Logit Models, *Technical Report*, U.S. Department of Transportation, Federal Transit Administration.

- Marschak, J. (1960), Binary-Choice Constraints and Random Utility Indicators, In: *Proceedings of a Symposium on Mathematical Methods in the Social Sciences*.
- McFadden, D. (1974), Conditional Logit Analysis of Qualitative Choice Behavior. In: Zarembka P., *Frontiers in Econometrics*, 1974, 105-142, Academic Press, New York.
- McFadden, D. (1978), Modelling the Choice of Residential Location, *Institute of Transportation Studies*, University of California.
- McFadden, D. (2001), Economic Choices, *American Economic Review*, 91 (3), 351–378.
- McGill, J. & Van Ryzin, G. (1999), Revenue Management: Research overview and prospects, *Transportation Science*, 33 (2), 233-256.
- Munizaga, M. & Alvarez-Daziano, R. (2002), Evaluation of Mixed logit as a Practical Modelling Alternative. In: Proceedings: *European Transport Conference*, Cambridge, UK.
- Rosenthal, R. E. (2010), *Gams- A User's Guide*, GAMS Development Corporation, Washington, DC, USA.
- Silberhorn, N., Boztuğ, Y. & Hildebrandt, L. (2008), Estimation with the Nested Logit Model: Specifications and Software Particularities, *OR Spectrum*, 30 (4), 635–653.
- Talluri, K. & Van Ryzin, G. (2005), *The Theory and Practice of Revenue Management*, Springer Verlag, New York, NY, USA
- Train, K. E. (2009), *Discrete Choice Methods with Simulation*, Cambridge University Press.
- Walker, J.L. (2001), Extended Discrete Choice Models: Integrated Framework, Flexible Error Structures and Latent Variables, Ph. D. Thesis, Massachusetts Institute of Technology.
- Williams, H. C. W. L. & Ortuzar, J.D. (1982), Behavioural Theories of Dispersion and the Mis-specification of Travel Demand Models, *Transportation Research Part B: Methodological*, 16 (3), 167-219.
- Williamson, E. (1992), Airline Network Seat Inventory Control: Methodologies and Revenue Impacts, Ph. D. Thesis, Massachusetts Institute of Technology.
- Zhang, M. & Bell, P. (2010), Price Fencing in the Practice of Revenue Management: An Overview and Taxonomy, *Journal of Revenue & Pricing Management*, 11 (2), 146–159.

2.4 Choice-Based Revenue Management with Flexible Substitution Patterns

Choice-Based Revenue Management with Flexible Substitution Patterns - Working Paper

Frauke Seidel*

*Hamburg University, Institute for Transport Economics, Moorweidenstraße 18, D-20148
Hamburg, Germany*

Sven Müller

*Karlsruhe University of Applied Sciences, Department of Transport Systems
Management, Moltkestraße 30, D-76133 Karlsruhe, Germany*

Knut Haase

*Hamburg University, Institute for Transport Economics, Moorweidenstraße 18, D-20148
Hamburg, Germany*

Abstract

In this paper, we present a new single-leg choice-based (airline) revenue management (CBRM) optimization model with flexible demand substitution patterns between fare classes. The respective demand model is based on individual utility values that, in sum, represent demand for the choice of airline tickets in certain fare classes. We particularly focus on non-constant substitution between alternatives to capture shifts in demand between alternatives that share common unobserved characteristics from the decision makers perspective. Thus, we are able to relax assumptions applied to revenue management optimization models that employ the multinomial logit demand model. We embed a general random utility model in a simulation-based mixed-integer linear program for revenue maximization. Thereby, we determine the prices for - and the availability of - each fare class and guarantee an optimal allocation of bookings to offered fare classes. We are able to solve instances up to

*Corresponding author

Email addresses: frauke.seidel@uni-hamburg.de (Frauke Seidel),
sven.mueller@hs-karlsruhe.de (Sven Müller), knut.haase@uni-hamburg.de (Knut Haase)

200 bookings (close) to optimality using GAMS/CPLEX. We show by a numerical investigation that the assumption of constant substitution patterns (namely, using the MNL) may yield erroneous predictions of the revenues when the choice behavior is, in fact, more complex. We investigate the loss in revenue due to the application of an inappropriate demand model by a series of numerical studies.

Keywords: choice-based revenue management, discrete optimization, discrete choice, flexible substitution patterns, random utility, logit, nested logit, mixed logit

1. Introduction

The single-resource quantity-based (airline) revenue management (RM) problem is about allocating passengers (demand) to seat capacity (availability). It arises from a special cost structure airlines face. In the short-term, the costs of an airline are largely fixed while transporting an additional passenger adds only little variable costs in comparison to the per unit selling price. This aspect also applies to the hospitality industry and motivated the adoption of revenue management techniques from the airline industry.

Air carriers have, in most situations, an incentive to find booking policies that maximize revenues from sold seats (McGill and Van Ryzin, 1999). To optimally allocate the restricted capacity of a given aircraft to demand, booking policies should ideally exploit each passenger's willingness to pay (WTP) which may vary, amongst other things, according to the reason for travel, demographic characteristics, and individual income. On the other hand, passengers' choices may also depend on the availability of different fare classes that exhibit different prices.

Depending on the control variable that an airline primarily uses to manage demand, revenue management is either qualified as quantity-based and price-based. Thereby, the first uses decisions on capacity allocation and the latter prices as tactical tool. In quantity-based RM, single-resource capacity controls refer to the optimal allocation of seats on a single flight to demand for different fare classes. Therefore, different control types like booking limits, nested booking limits and bid-price controls are available (Talluri et al., 2008). Furthermore, *static* single-resource RM models assume that demands for different fare classes are independent stochastic processes that are not influenced by the availability of other alternatives. This assumption is part

of the so called *independent demand model* (Vulcano et al., 2010; Talluri and Van Ryzin, 2004b, pp. 33) where an endogenization of customer behavior is disregarded. Approaches that incorporate the *independent demand model* are more and more seen as rather unrealistic since the availability of low fare tickets impacts the probability of selling a ticket in a more expensive fare class. Furthermore, the decision of price sensitive customers to buy a ticket at all possibly depends on the availability of the lowest fare (Vulcano et al., 2010).

Traditionally the success of airline revenue management is based on the assumption that cheaper fare classes exhibit combinations of various restrictions like minimum stay or advance purchase requirements and cancellation or rebooking penalties. These restrictions are known as fences between differently priced fare classes (Zhang and Bell, 2010). By defining varying fares for similar products (e.g., tickets for two seats in the same fare class, one with rebooking penalty and one without), potential customers are segmented into groups according to their purchase preferences. Thereby, price differences cannot be completely explained by differences in marginal cost (Stigler, 1987). This instrument is known as price discrimination and offers airlines the potential to increase their total revenues (Williamson, 1992; Talluri and Van Ryzin, 2004b, pp. 352-363). Thus, fencing serves as a justification for the disregard of upsell and downsell between fare classes as customers are discouraged from purchasing a ticket with differing characteristics (Fiig et al., 2010). The ongoing growth of low cost carriers (LCC) has led to the development of fare structures with simplified or no fare restrictions known as restriction free pricing (RFP). The lack of proper fencing results in customers choosing lower priced available fares leading to an increase in load factors while total revenues decline. Customers buy up (buy down) to more expensive (cheaper) fare classes if their preferred choice is unavailable. Therefore, older RM approaches based on the assumption of the independent demand model become less relevant since airlines are more interested in RM tools that are able to manage dependent demands (Weatherford and Ratliff, 2010). However, so called dependent demand RM models are much more complex to manage as the demands used in the optimization are conditional on the RM controls (i.e., price and/or availability). The fares determined in the optimization are conditional on the acquired upsell and recapture for each fare class. The calculation of the according prob-

abilities can be accomplished in different ways such as FRAT5 curves¹ and logit demand models. For example Gallego et al. (2009) utilize a multinomial logit (MNL) model to calculate upsell in a single-leg expected marginal seat revenue (EMSR) model with demand dependencies. However, traditional quantity-based revenue management models mainly focus on the allocation of capacity while prices and demand are considered as exogenously given. Following Kocabiyikoğlu et al. (2013), the partition of pricing and capacity allocation is attributable to the divisional separation of the according organizational functions: marketing (pricing) and operations (revenue management). Another reason is attributed to technical and operational difficulties in implementing a support system for simultaneous price & availability decisions. Although the importance of an integrated approach regarding revenue management and pricing decisions is widely acknowledged their systematic coordination is still at emerging stage (McGill and Van Ryzin, 1999; Garrow et al., 2006).

Recently, choice-based optimization models have become an active research area in airline revenue management and beyond (Haase and Müller, 2014; Gallego and Topaloglu, 2014; Klier and Haase, 2015; Berbeglia and Joret, 2015; Müller and Haase, 2017). Within such modeling frameworks customer demand is usually represented by disaggregate choices via discrete choice models. In contrast to considering deterministic demand, a choice-based approach enables the researcher to account for effects of demand endogenization (Krohn et al., 2017).

Thus, customer preferences can be modeled more accurately and offered products and services can be specified accordingly. In the context of capacity allocation Talluri and Van Ryzin (2004a) apply an MNL model to investigate which subset of fare products should be offered to customers at certain points of time in a single-leg revenue management model. Based on the expectation maximization (EM) method the authors further obtain estimates for both the arrival process and choice model parameters when no-purchase information from customers is unavailable. Gallego et al. (2004) study flexible products in a network RM setting with demand based on a customer choice model. They provide a definition for the customer choice model that

¹FRAT5 stands for 'fare ratio 50 per cent' and represents an approach where fare class price and demand are modelled by utilizing a negative exponential distribution (Weatherford and Ratliff, 2010).

includes the independent demand model, the MNL and "attraction models" such as multinomial probit. Thereby, the authors show that demand for each product depends upon which other products are available. The linear programming (LP) formulation is solved efficiently by a column-generation algorithm. [Vulcano et al. \(2010\)](#) investigate choice-based RM on airline data where choice sets contain flights on a certain day at different flight times. Choices are modelled by an MNL model although the authors raise concern about the known properties of the models constant substitution patterns that might lead to an overestimation of choice probabilities when alternatives share common unobserved characteristics. Alternatively, more flexible discrete choice models should be considered where substitution patterns account for the fact that some alternatives might be closer substitutes than others.

Obviously, studies that incorporate customer choice behavior in RM optimization models mainly utilize the MNL model to account for demand dependencies. Furthermore, the previously mentioned studies do not incorporate simultaneous decisions of capacity allocation and pricing. However, the MNL model exhibits the independence from irrelevant alternatives (IIA) property and hence constant substitution patterns between choice alternatives. The red-bus blue-bus paradox is a prominent demonstration how a model that exhibits the IIA property might overpredict choice probabilities if alternatives in the choice set share common unobserved characteristics. In this case the MNL model predicts false choice probabilities since substitution patterns between any two alternatives within the choice set are required to be constant no matter if additional alternatives exist. Thus, probability ratios are considered for pairs of alternatives while the influence of alternatives that are added or removed from the choice set are *irrelevant* to the calculation of these ratios. This is an issue, that is also recognized by [Vulcano et al. \(2010\)](#). Therefore, in discrete choice theory, the MNL, even though it is very popular and often applied when it comes to demand modeling, is known to be quite restrictive in certain choice situations ([McFadden, 2001](#)). When some of the alternatives in the choice set share unobserved attributes (i.e., similarities) the MNL model does not appropriately reproduce individuals' decision making behavior. This is due to the fact that the stochastic part of utility in the MNL is assumed to be identically and independently (iid) extreme value (EV) distributed resulting in stochastically independent choice decisions between alternatives ([Train, 2009](#), p. 38).

In terms of RM optimization models, these disaggregate choices represent

the demand for fare classes or itineraries. Thereby, some fare classes do exhibit similar characteristics or restrictions, that may lead to non-proportional shifts in demand when fare classes are closed for purchase (i.e., they are not made available for purchase by the airline). Although researchers suggest that more realistic choice models might improve the performance of RM controls, concerns exist with regard to their computational complexity (Vulcano et al., 2010; Etebari and Najafi, 2016).

The explicit incorporation of discrete choice probabilities into the formulation of mathematical programs results in non-linear models (Müller and Haase, 2016). With regard to choice-based optimization problems for locational and assortment decisions, linear reformulations to this problem have been studied by Benati and Hansen (2002), Haase (2009), and Davis et al. (2013). When demand within an optimization model is represented by an MNL, constant substitution patterns (i.e., IIA property) can be used to obtain a linear reformulation of the otherwise non-linear choice probabilities. If we now assume that individual choice behavior follows a more general choice model that exhibits flexible substitution patterns (i.e., not exhibiting IIA property), such linearizations can no longer be determined. In this context, Bierlaire and Azadeh (2016) propose an approach for demand based revenue maximization that is able to incorporate a variety of discrete choice models. They, and Haase and Müller (2013), ensure linear model formulations of the decision variables by applying customers' utilities instead of using the respective choice probabilities. We provide a different mathematical model formulation that is based on customers utilities and the corresponding discrete choice outcomes as presented by Haase (2009) and in the conference talks by Seidel (2011), Seidel (2012) and Seidel (2013). Any substitution pattern can be considered by appropriate specification of the customers' utility function. Therefore, we are able to bypass the problems caused by demand models that exhibit the IIA property. This results in accurate predictions of revenues and the appropriate determination of fare class ticket prices. Our general problem formulation allows for the incorporation of any random utility model (RUM).

Our work mainly contributes to existing research in three points: (i) We provide a formulation for a dynamic (i.e., low-to-high revenue demand arrives in arbitrary order) single-resource choice-based RM model that overcomes two aspects of the independent demand model as stated in Talluri and Van Ryzin (2004b, pp. 33): (1) we relax the assumption that demand for different fare classes are independent random variables and (2) our approach is

thereby able to capture demand dependencies between multiple available fare classes. We achieve this by integrating individual demand for fare classes into a mathematical model for airline revenue management by utilizing a discrete choice model with flexible substitution patterns. (ii) We propose a model formulation that is linear in the decision variables since stochastic utility functions are directly included in the formulation of the optimization model. (iii) Besides providing a quantity based decision instrument for airline RM, our approach also decides on the optimal prices (non-negative variables) for offered/available fare classes.

2. Fare class choice

Our considerations regarding fare class choice behavior are based on the theoretical framework of random utility theory. Following [Marschak \(1960\)](#) a choice model derived under the assumption that a choice maker maximizes its personal utility is called a random utility model (RUM). Thereby, utility is a random quantity from the researchers perspective as some influences affecting the choice decision are not observable. Discrete choice models belong to this category of models ([McFadden, 1974](#)).

For our demand model, we suppose that $n = 1, \dots, N$ airline passengers choose exactly one alternative f out their individual choice sets C_n . According to [Train \(2009, p. 15\)](#) C_n consists of a finite number of mutually exclusive and exhaustive alternatives, here fare classes and an opt-out alternative i.e., to choose not to fly with the airline under consideration.

In line with random utility theory, passengers evaluate each available alternative depending on its attributes (e.g. price) as well as their own individual characteristics (e.g., income and gender). By choosing one alternative f out of C_n each passenger n receives utility u_{nf} . It is formally defined as the decomposition into a deterministic part v_{nf} and a stochastic part ε_{nf} :

$$u_{nf} = v_{nf} + \varepsilon_{nf} \quad \forall n, f \quad (1)$$

with

$$v_{nf} = \sum_k \beta_f^k \cdot z_{nf}^k. \quad (2)$$

Thereby, parameters β_f^k have to be estimated from empirical choice data and z_{nf}^k denote the available alternatives attributes and characteristics of

the choice maker n . ε_{nf} is a stochastic component and utility thereby is a random quantity.

Since in a RUM a choice maker seeks to maximize its utility, the alternative with the highest utility value determines the choice decision. Thus, the choice rule for the fare class choice problem is the following (McFadden, 2001; Ben-Akiva and Lerman, 1985, p. 101): A passenger n chooses alternative f iff

$$u_{nf} > u_{nf'} \quad \forall f \neq f'. \quad (3)$$

According to (1) u_{nf} is stochastic and thereby the probability of choosing f over f' is formally defined by:

$$\begin{aligned} P_{nf} &= \text{Prob}(u_{nf} > u_{nf'}, \forall f' \in C_n, f' \neq f) \\ &= \text{Prob}(v_{nf} + \varepsilon_{nf} > v_{nf'} + \varepsilon_{nf'}, \forall f' \in C_n, f' \neq f) \\ &= \text{Prob}(\varepsilon_{nf'} < v_{nf} - v_{nf'} + \varepsilon_{nf}, \forall f' \in C_n, f' \neq f) . \end{aligned} \quad (4)$$

From (4) we derive any RUM given a specific assumption about the joint distribution of the stochastic utility component ε_{nf} (Ben-Akiva and Lerman, 1985, p. 101). Here, the assumptions about the joint distribution of ε_{nf} determine customer choice behavior in our fare class choice model.

If we suppose that ε_{nf} is independently and identically (iid) type I extreme value (EV) distributed, we derive the MNL model from Equation 4. Thereby, the unobserved stochastic utility components are uncorrelated for different alternatives and each fare class is an equally good substitute for any other alternative in C_n . In other words, substitution patterns between fare classes are constant which results in demand being independent for different available fare classes and their attributes (Train, 2009, p. 39). Total expected demand for fare class f is given by

$$D_f = \sum_n w_n P_{nf} \quad (5)$$

with w_n being a weight for choice maker n . Choice maker n might be a representative for a specific segment of customers. Then, w_n is the number of customers in that segment.

Following the explanations in Seidel (2014) fare classes represent differing products in that they exhibit various combinations of travel restrictions as well as differing prices. We distinguish fare classes by the compartment in which the purchased seat is valid: Business class and economy class. Within

each compartment, fare classes are characterized by advance purchase requirements, length-of-stay requirements, rebooking and cancellation penalties, the possibility to upgrade, the possibility to collect frequent flyer miles, and many more. Thereby, fare restrictions may even vary within the same compartment. Consequently, more complex combinations of restrictions that are imposed on a fare class result in lower prices (Talluri and Van Ryzin, 2004b, pp. 521). Fare class restrictions provide a fencing between low and high fare (price) products that prevents certain customers (i.e., business travellers) from buying down to a cheaper fare class (Zhang and Bell, 2010). A buy down occurs when a customer who is willing to purchase a high fare product in the first place actually chooses a discount fare when both products are available. Thus, fencing aims on discouraging high fare customers from purchasing low fare tickets since fare restrictions reduce the utility of cheaper fares (Fiig et al., 2010). Different fare classes that are sold on a single flight provide passengers with combinations of similar restrictions. These restrictions might be evaluated by a choice maker in a similar way. Therefore, we assume that demand dependencies exist between alternatives with common characteristics. This assumption is supported by Carrier (2003). In such a case, constant substitution patterns, as imposed by the MNL model, are an inappropriate assumption for the considered fare class choice problem. Therefore, we presume, the true choice behavior of an airline customer is best represented by a choice model that allows for correlation in the unobserved factors of utility ε_{nf} .

3. Choice-based revenue management optimization

The underlying idea of the model formulation is based on Monte-Carlo simulation. We consider a population of individuals with inherent utility values for choosing a fare class (or the opt-out alternative) out of a given choice set. Individual utility is defined as a random quantity whose specification depends on the researchers assumptions about the underlying choice behavior of the customers. Individual choices are unknown to the airline (see Section 2). Intuitively, one would employ choice probabilities (4) to determine demand D_f . Unfortunately, no matter which assumptions are made about the stochastic part ε_{nf} , the resulting choice probabilities are non-linear and/or non-closed form. So far, linear reformulations only exist for the restrictive MNL (see Section 2). Since the customer's choice problem is inherent to the choice probabilities (4), we propose to consider the choice problem (3)

instead of the choice probabilities P_{nf} in the optimization problem to come up with a linear formulation.

We assume a predetermined number of scenarios $|S|$ whereas ε_{snf} is a realization of ε_{nf} for scenario s . Then, $U_{snf} = V_{nf} + \varepsilon_{snf}$. Now consider the binary variable Y_{snf} which equals one, iff customer n chooses f in scenario s . That means, $Y_{snf} = 1$, iff $U_{snf} > U_{snf'} \forall f' \neq f$. Since the choice is unique, i.e., $\sum_f Y_{snf} = 1 \forall n, s$, the choice probabilities of (4) can be approximated as

$$\hat{P}_{nf} = \frac{\sum_s Y_{snf}}{|S|} \quad (6)$$

which is an unbiased estimator of P_{nf} by construction (Train, 2009, pp. 115).

As $|S|$ approaches infinity \hat{P}_{nf} approximates P_{nf} .

That way, we are able to overcome the restrictive assumptions made in airline revenue management optimization regarding demand independence between offered fare classes (McGill and Van Ryzin, 1999). We directly incorporate individual utility values and the utility maximizing principle from (3) into the optimization model formulation avoiding non-linear formulations. Of course, this comes at the cost of increasing the dimension of our optimization problem - and $|S|$ might be large. However, Krohn et al. (2017) show that $|S|$ up to 200 might be sufficient in terms of approximating P adequately close.

We assume the price of a fare class f impacts the choice behavior of customer n . Therefore, we consider the non-negative decision variable p_f that denotes the price for fare class f to be element of the deterministic utility $v_{nf}(p_f)$. The stochastic part of utility captures the dependence (i.e., correlation) between fare classes. We consider $|S|$ realizations of the stochastic part and therefore (1) becomes the auxiliary variable

$$U_{snf} = v_{nf}(p_f) + \varepsilon_{snf} \quad \forall s, n, f \quad (7)$$

whereas $U_{snf} \in \mathbb{R}$. From the airlines perspective attributes and characteristics like gender, age, seat space, and travel-time are assumed to be known and are therefore computed in advance for each customer. From the optimization perspective, the realization of the stochastic component can be considered as a parameter as well i.e., ε_{snf} is also computed in advance. Hence, Equation (7) can be written as

$$U_{snf} = -\beta^{\text{price}} \cdot p_f + \tilde{v}_{nf} + \varepsilon_{snf} \quad \forall s, n, f \quad (8)$$

with parameter β^{price} as weight of p_f and parameter \tilde{v}_{nf} containing all pre-determined attributes and characteristics as well as their respective weights. Denoting parameter $\delta_{snf} = \tilde{v}_{nf} + \varepsilon_{snf}$ we can simplify (8) to

$$U_{snf} = -\beta^{\text{price}} \cdot p_f + \delta_{snf} \quad \forall s, n, f. \quad (9)$$

Variable U_{sn}^{max} denotes the maximum utility of individual n for scenario s , i.e.,

$$U_{sn}^{\text{max}} = \max U_{snf} \quad \forall s, n. \quad (10)$$

Following the utility maximization choice rule, choice maker n chooses alternative f if and only if

$$U_{snf} = U_{sn}^{\text{max}}. \quad (11)$$

3.1. Model Formulation

We consider the following additional parameters and variables:

Parameters

\bar{p}_f	upper bound for price variable in fare class f , $\forall f \neq \text{opt-out}$
c	seat limit of considered resource (i.e., aircraft)
L	sufficiently large numbers

Variables

X_{nf}	= 1, if f is offered to n (0, otherwise)
p_f	price for a ticket in fare class f
Π_{snf}	price individual n pays for a ticket in fare class f in scenario s
R	total expected revenue

The objective of our model formulation is to maximize the sum of all prices that are paid by passengers n for tickets on the considered flight-leg. That is, to maximize total expected revenue:

$$\text{maximize } R = \frac{1}{|S|} \cdot \sum_s \sum_n \sum_{f \neq \text{opt-out}} \Pi_{snf} \quad (12)$$

Equations (13) and (14) below define the value of Π_{snf} with $\Pi_{snf} \geq 0$. By Equation (13) we ensure that Π_{snf} is assigned a value no larger than the fare class specific price limit only if fare class f is actually chosen by passenger n in scenario s whereas $Y_{snf} \in \{0, 1\}$.

$$\Pi_{snf} \leq \bar{p}_f \cdot Y_{snf} \quad \forall s, n, f \neq \text{opt-out} \quad (13)$$

By Equation (14) we ensure that Π_{snf} takes the value of price variable p_f :

$$\Pi_{snf} \leq p_f \quad \forall s, n, f \neq \text{opt-out}. \quad (14)$$

The utility of customer n choosing alternative f in scenario s is determined by Equation (15):

$$U_{snf} = \delta_{snf} - \beta^{\text{price}} \cdot p_f - (\delta_{snf} + 1) \cdot (1 - X_{nf}) \quad \forall s, n, f \neq \text{opt-out}. \quad (15)$$

X_{nf} is a binary decision variable with $X_{nf} \in \{0, 1\}$. If fare class f is offered to customer n , i.e., $X_{nf} = 1$, $U_{snf} = \delta_{snf} - \beta^{\text{price}} \cdot p_f$. The higher the value of p_f the lower the utility of a fare class f and hence the likelihood that f is chosen by n . Of course, the objective is to determine p_f such that R is maximized. The higher p_f the higher Π_{snf} while on the contrary U_{snf} declines in p_f . This trade off is acknowledged by interdependencies in Equations (14) and (15). To ensure, that for each passenger n the maximum value of utility U_{snf} over all fare classes f is assigned to variable U_{sn}^{\max} we consider:

$$U_{sn}^{\max} \geq U_{snf} \quad \forall s, n, f \neq \text{opt-out}. \quad (16)$$

The domain of U_{sn}^{\max} is $\mathbb{R}^{\geq \delta_{sn, \text{opt-out}}}$. Equations (17) guarantee that a passenger n chooses the alternative f that maximizes utility in scenario s , i.e., $Y_{snf} = 1$ if $U_{snf} = U_{sn}^{\max}$.

$$U_{sn}^{\max} - U_{snf} \leq L \cdot (1 - Y_{snf}) \quad \forall s, n, f \neq \text{opt-out} \quad (17)$$

otherwise, $Y_{snf} = 0$. Although, Equation (16) allows for $U_{sn}^{\max} > U_{snf}$, it (16) yields $U_{sn}^{\max} = \max_f(U_{snf})$ if at least for one f $U_{snf} > \delta_{sn, \text{opt-out}}$, because otherwise for given s and n all Y_{snf} are zero due to Equation (17) which yields lower R due to Equation (13). According to random utility theory customer n chooses exactly one alternative f in scenario s . If $\sum_f Y_{snf} = 0$ the customer n chooses opt-out.

$$\sum_{f \neq \text{opt-out}} Y_{snf} \leq 1 \quad \forall s, n \quad (18)$$

We further decide on the fare classes offered to customers. The set of offered fare classes to a passenger n is equivalent to the individual choice set $C_n = \{\text{opt-out}\} \cup \{f | X_{nf} = 1\}$

Furthermore, we ensure that only offered fare classes f can be chosen by a certain customer.

$$Y_{snf} \leq X_{nf} \quad \forall s, n, f \neq \text{opt-out} \quad (19)$$

Constraints (16) - (19) map the customers' choice problem (3). As such, demand (5) for fare class f can be determined². In total, the amount of tickets purchased by all passengers must not exceed the resource limit c which is ensured by Equations (20):

$$\sum_n \sum_{f \neq \text{opt-out}} Y_{snf} \leq c \quad \forall s \quad (20)$$

The domains for the considered variables are as follows:

$$\Pi_{snf} \geq 0 \quad \forall s, n, f \quad (21)$$

$$p_f \geq 0 \quad \forall f \quad (22)$$

$$U_{snf} \in \mathbb{R} \quad \forall s, n, f \quad (23)$$

$$U_{sn}^{\max} \in \mathbb{R}^{\geq \delta_{sn, \text{opt-out}}} \quad \forall s, n \quad (24)$$

$$X_{nf} \in \{0, 1\} \quad \forall n, f \quad (25)$$

$$Y_{snf} \in \{0, 1\} \quad \forall s, n, f \quad (26)$$

²Demand for fare class f : $D_f = \sum_n \frac{\sum_s Y_{snf}}{|S|}$

Attributes z_{nf}^k	Min	Mean	Max
<i>Price in fare class f</i>			
z_{n1}^{price}	312.0	1001.0	1737.0
z_{n2}^{price}	256.0	801.6	1350.0
z_{n3}^{price}	68.0	400.5	803.0
z_{n4}^{price}	20.0	199.4	378.0
<i>Gender</i>			
z_n^{gender}	0.0	0.47	1.0
<i>Trip purpose</i>			
z_n^{purpose}	0.0	0.28	1.0

Table 1: Descriptive statistics. Gender and trip purpose are so called dummies, $z_n^{\text{gender}} = 1$, if n is male (0, otherwise) and $z_n^{\text{purpose}} = 1$ if trip purpose of n is business (0, otherwise).

4. Numerical investigation

To verify the applicability of the proposed mathematical model we perform a numerical investigation by applying our approach to a small real world problem. We consider a single-leg continental flight between a city pair that is not further specified. We model demand for tickets in a certain fare class on this flight by simulating individual utility values according to the explanations outlined in Sections 2 and 3. We show the superiority of applying a demand model with flexible substitution patterns (over a demand model with constant substitution patterns) to the mathematical model proposed in Section 3.1. Therefore, we first introduce the specification of the demand model as considered by Seidel (2014). We then perform a validation of the proposed approach by investigating the modeling outcomes for two demand models with differing substitution patterns. Then, we provide a sensitivity analysis followed by the results of the computational study.

We perform the computational investigation by implementing all problems in GAMS 24.7.1 and solving them with CPLEX 12 on a 64-bit Windows 10 Pro Server with 2 Intel Xeon 3.20 GHz processors and 256 GB RAM.

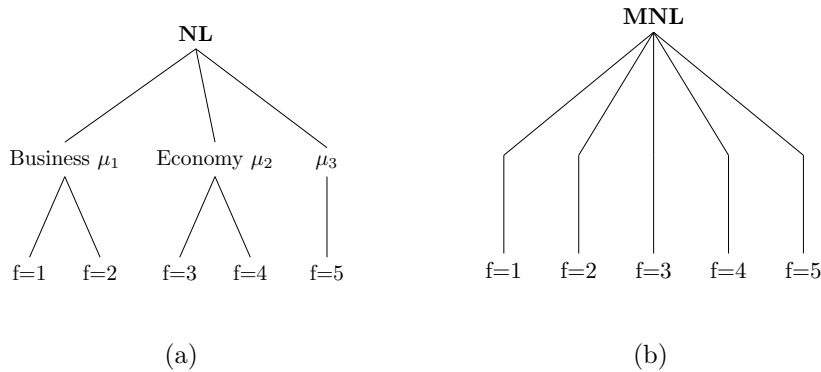


Figure 1: Nesting structures for fare class choice for (a) the NL demand model with a two-level nest structure and corresponding nest coefficients and (b) the MNL demand model for comparison.

4.1. Demand model and substitution patterns

For validation purposes we apply demand models with flexible and constant substitution patterns to the mathematical model from Section 3.1. We thereby compare the modeling outcomes resulting from two strategies that vary according to the assumed choice behavior of customers. We assume, non-constant substitution patterns between fare classes are caused by restrictions that are common to different fare classes. Seidel (2014) showed that for their choice data the nested logit model (NL) which allows for correlation among fare classes performs better than the MNL. Based on this choice data, we specify two demand models: (i) MNL, i.e., constant substitution patterns ignoring correlation between fare classes, and (ii) nested logit (NL) with more flexible substitution patterns, i.e., accounting for correlation. We examine the impact of applying an inappropriate demand model (MNL) on managerial decision making. Furthermore, we provide the summary statistics of the data used for choice model estimation in Table 1. The desired substitution patterns are reflected by the choice decisions that are calculated for each generated individual according to the choice rule as defined in (3). Following Seidel (2014) and the definition of (2) we consider the following functions of the deterministic utility component v_{nf} :

$$v_{n1} = \beta_1^{\text{asc}} + \beta^{\text{price}} \cdot z_{n1}^{\text{price}} + \beta_1^{\text{gender}} \cdot z_n^{\text{gender}} + \beta_1^{\text{purpose}} \cdot z_n^{\text{purpose}} \quad (27)$$

$$v_{n2} = \beta_2^{\text{asc}} + \beta^{\text{price}} \cdot z_{n2}^{\text{price}} + \beta_2^{\text{gender}} \cdot z_n^{\text{gender}} + \beta_2^{\text{purpose}} \cdot z_n^{\text{purpose}} \quad (28)$$

$$v_{n3} = \beta_3^{\text{asc}} + \beta^{\text{price}} \cdot z_{n3}^{\text{price}} + \beta_3^{\text{gender}} \cdot z_n^{\text{gender}} + \beta_3^{\text{purpose}} \cdot z_n^{\text{purpose}} \quad (29)$$

$$v_{n4} = \beta_4^{\text{asc}} + \beta^{\text{price}} \cdot z_{n4}^{\text{price}} + \beta_4^{\text{gender}} \cdot z_n^{\text{gender}} + \beta_4^{\text{purpose}} \cdot z_n^{\text{purpose}} \quad (30)$$

$$v_{n5} = 0 \quad (31)$$

Parameters β_f^{asc} denote the alternative specific constants that measure the average preference of customers for alternative f . The remaining parameters β^{price} , β_f^{gender} and β_f^{purpose} represent the utility contribution for alternative f per unit of attribute k with $k \in \{\text{price, gender, trip purpose}\}$. To obtain the overall individual utility u_{nf} , the stochastic utility component ε_{nf} is added to v_{nf} . The distribution of ε_{nf} is assumed by the researcher. Thus, we obtain the MNL model by assuming that ε_{nf} follows an iid EV distribution while an NL model follows from assuming that the ε_{nf} are generalized extreme value (GEV) distributed (Train, 2009, pp. 80).

Table 2 displays the maximum likelihood parameter estimates. Estimated parameters differ according to the applied demand model. In particular, we find price sensitivity, that is represented by parameter β^{price} , to be higher when we wrongly assume that individual choice behavior follows an MNL demand model. Furthermore, from parameters β_f^{gender} we conclude that for male customers the utility of choosing a ticket in a certain fare class decreases with decreasing fare class price. We observe a similar result with regard to trip purpose. In both models, a customer receives the highest increase in utility on a business trip when the most expensive fare class is chosen (β_1^{purpose}).

For comparison, the least increase in utility has a customer on a business trip who chooses the cheapest ticket (i.e., fare class 4).

For the NL demand model, we assume a two-level nest structure as displayed in Figure 1. Thus, for each nest, we additionally obtain the nest coefficients μ_1 and μ_2 from model estimation. They represent the lower level scale parameters of the NL model and reflect the amount of correlation between alternatives within the same nest allowing for flexible substitution patterns. Since the opt-out alternative $f = 5$ is solely assigned to the third nest, we have a degenerate nesting structure. For identification purposes, we therefore choose $\mu_3 = 1$ for model estimation.

Parameter	NL		MNL	
	Estimate	t-Test	Estimate	t-Test
<i>Alternative-specific constants</i>				
β_1^{asc}	0.340	0.99	0.811	2.23
β_2^{asc}	1.35	4.60	2.07	9.13
β_3^{asc}	1.38	11.74	1.01	7.32
β_4^{asc}	1.85	19.35	2.01	23.49
<i>Price</i>				
β^{price}	-0.00406	-10.81	-0.00512	-19.66
<i>Gender</i>				
β_1^{gender}	0.765	3.55	0.790	3.16
β_2^{gender}	0.568	4.28	0.583	4.26
β_3^{gender}	0.0969	0.99	0.212	1.89
β_4^{gender}	-0.1520	-1.83	-0.181	-2.13
<i>Trip purpose</i>				
β_1^{purpose}	1.76	7.05	1.90	6.63
β_2^{purpose}	1.29	9.41	1.29	9.14
β_3^{purpose}	0.935	8.78	1.16	10.11
β_4^{purpose}	0.437	5.11	0.379	4.39
μ_1	1.29	6.07	–	–
μ_2	1.64	8.52	–	–
Observations	4000		4000	
Parameters	15		13	
$\mathcal{L}(\beta_0)$	-6437.752		-6437.752	
$\mathcal{L}(\hat{\beta})$	-4523.122		-4532.424	
ρ^2	0.297		0.296	
$\bar{\rho}^2$	0.295		0.294	
<i>t</i> -test against 0				
<i>t</i> -test against 1	17			

Table 2: Estimation results from Biogeme for NL and MNL model using Biogeme (Bierlaire, 2003) software package.

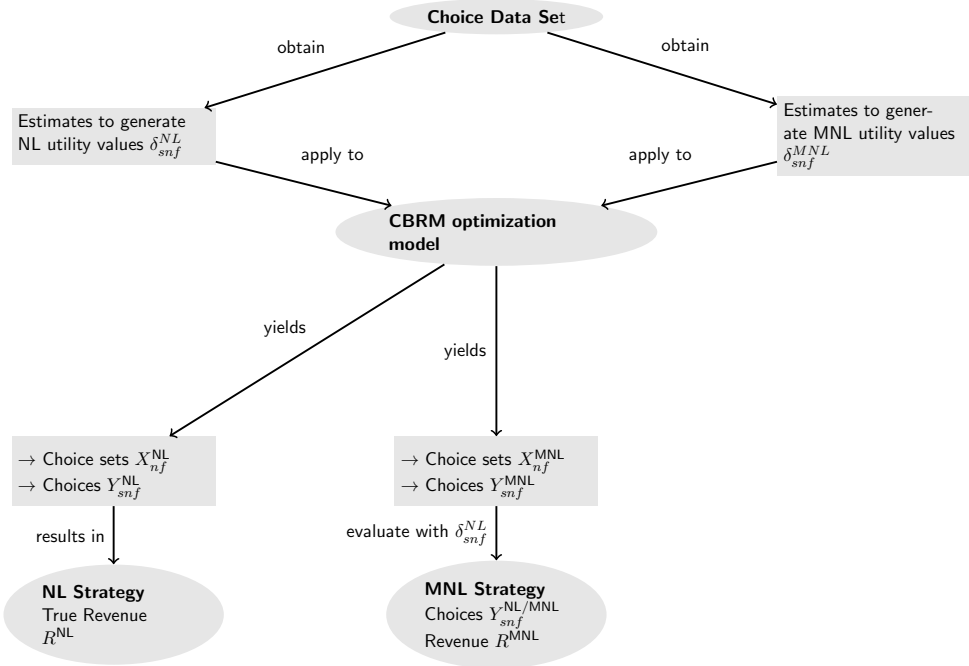


Figure 2: Validation of the NL strategy against the MNL strategy.

With regard to model fit, we observe only minor differences between both models. The final log-likelihood values $\mathcal{L}(\hat{\beta})$ indicate a slightly better model fit of the NL model over the MNL model. This is also confirmed by the $\hat{\rho}^2$ value that is marginally larger for the NL model. A Horowitz-Test rejects the hypothesis that the MNL is the true model.

4.2. Model validation

Despite the fact that we discover only minor differences in estimated parameters and model fit between the NL and MNL models, we can show that revenues do vary remarkably depending on the considered demand model. We choose a small numerical example with $N = 10$ individuals, capacity $c = 10$, and $S = 100$ scenarios. We solve the problem for two strategies with regard to individual choice behavior: (i) assuming an NL demand model

Market Shares					
Strategy	Fare Class 1	Fare Class 2	Fare Class 3	Fare Class 4	Opt-out
NL	0.02	0.180	0.365	0.03	0.405
MNL	0.01	0.204	0.324	0	0.462

Table 3: Average market shares over ten random instances for the NL and MNL strategies.

Fare class prices				
Strategy	Fare Class 1	Fare Class 2	Fare Class 3	Fare Class 4
NL	1255	880	578	120
MNL	1296	809	624	100

Table 4: Average prices paid by customer n per fare class in the NL and MNL strategies for the considered problem set (S=100 x N=10).

and (ii) assuming an MNL model that we consider to be the wrong demand model. We solve each strategy with ten random instances and investigate the outcome of both strategies. For validation we utilize the approach as displayed in Figure 2. First, we apply the estimated coefficients from Table 2 in Section 4.1 to the demand model specific generation of individual utility values δ^{NL} and δ^{MNL} . We then employ these utility values to our mathematical program (Section 3.1) to obtain choice sets and choices that are specific to the respective demand model. Solving the CBRM model yields decision variables $X_{nf}^{\text{NL}}, Y_{snf}^{\text{NL}}$ for the NL demand model and X_{nf}^{MNL} and Y_{snf}^{MNL} for the MNL demand model. The obtained variable values provide the choice sets (i.e., offered fare classes) and choice decisions (i.e., chosen fare classes) for our example. Then, we examine both strategies with regard to the revenues we obtain respectively.

According to Figure 2, we obtain the true revenue and choice decisions for the NL strategy from applying the NL demand model to our optimization model. To compare the true revenue with the revenue obtained from an MNL strategy, we evaluate each solution from the MNL demand model with the

true (NL) demand model. This yields the revenue and choice decisions of the MNL strategy.

Choice decisions and thus market shares per fare class slightly differ by strategy as can be seen in Table 3. In both strategies fare classes 2 and 3 are the most chosen fare options accounting for circa 55% of the total choices in the NL strategy and for 52% in the MNL strategy. In the NL strategy, fare class 3 is the most chosen option accounting for 36% of all choice decisions whereas the same fare class yields a market share of 32% in the MNL strategy.

The figures displayed in Table 3 also reflect the differences in substitution patterns between the considered strategies. Since fare class 4 is hardly made available to arriving customers in the NL strategy (i.e., the market share of fare class 4 is 3%) more choices go to fare class 3 which is the closest substitute to fare class 4. This is reasonable since in the NL strategy fare class 3 is a closer substitute to fare class 4 than are the remaining fare classes (see Figure 1).

On the contrary, in the MNL strategy the market share of fare classes 1 and 2 is higher than in the NL strategy while it is less for fare class 3. Furthermore, the *opt-out* alternative is chosen less frequently in the NL strategy. The differences in individual choice decisions as well as optimal prices (Table 4) for each strategy result in higher revenues for an NL strategy compared to an MNL strategy. Although price sensitivity is lower in the NL strategy than in the MNL strategy (see β^{price} in Table 2) we do not generally observe higher average prices that are paid by individuals for the respective fare classes in the NL strategy. Rather, the combination of market shares and ticket prices yields that MNL revenues are always below NL revenues.

In our example, revenues average over all instances at 4178.50 (NL strategy) and 3692.50 (MNL strategy) monetary units. Figure 3 displays the revenues for the NL and MNL strategies for each instance as well as the average revenues over all instances. Allowing for flexible substitution patterns - NL strategy - leads to higher revenues compared to the revenues earned in a more restrictive MNL strategy. Figure 3 displays the revenues of the NL and MNL strategies for 10 random instances. Thus, based on the results from the CBRM optimization model, we can show that for our example revenues are increased by considering flexible substitution patterns for passenger demand. Furthermore, we investigate the impact of changes in customers' price sensitivity on revenues. Therefore, we vary the price coefficient β^{price} within its standard error (see Table 2). We learn from Figure 4 that revenues increase with a decrease in price sensitivity. However, the increase is not strictly

monotone which can be seen from slightly declining revenue values for decreasing price sensitivity. Declining revenues for instances 4, 6, 8 and 10 result from variations in offered choice sets and individual choice decisions. With decreasing price sensitivity fare class 4 is offered less. In the first instance where $\beta^{price} = -0.00446$ 7 out of 10 individuals are offered tickets in fare class 4. In the sixth instance ($\beta^{price} = -0.00396$) fare class 4 tickets are offered to only two individuals. Thus, in our model a decreasing price sensitivity encourages potential passengers to buy up to more expensive fare classes by closing cheaper ones. On the choice side, this results in an increase of ticket sales of fare class 3, for example. While in the first instance on average 1.4 out of 10 individuals choose fare class 3 this number raises to 4.84 in the last instance ($\beta^{price} = -0.00356$). By investigating choice sets and individual choices for the provided example, declines in revenues between single instances can be attributed to variations in demand for more expensive fare classes in combination with slight increases in the number of individuals choosing the opt-out alternatives. For example, in the seventh instance ($\beta^{price} = -0.00386$) 3.36 out of 10 individuals choose opt-out on average, while in the next instance this number increases to 4.04. At the same time demand for tickets in fare classes 1, 2, and 3 declines slightly from instance 7 to 8 resulting in a decline in revenue.

In Table 6 we provide the results of the computational study where we investigate the solvability of the proposed mathematical program. Therefore, we choose three problem sets with a varying number of scenarios, simulated individuals (demand) and varying resource capacity (seats). While limiting the maximum computational time to 7200 seconds, we are able to solve all problems close to optimality with a relative gap of less than 6%. However, for two problem sets with $|S| = 50$ scenarios, and demand of $|N| = 200$ the relative gap is at 10% for a resource capacity of 50 and at 90% for a resource capacity of 100. Obviously, with an increasing number of scenarios the allocation of bookings to available seats becomes more difficult when capacity is smaller than the number of individuals $|N|$.

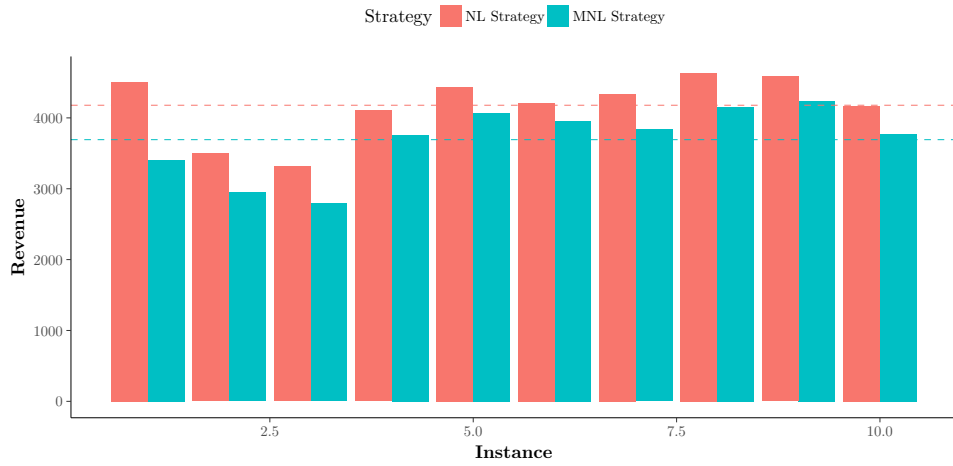


Figure 3: Comparison of revenues for the NL and MNL strategies for 10 random instances each with $|S| = 100$, $|N| = 10$ and $c = 10$. Dashed lines represent the respective averages over all instances.

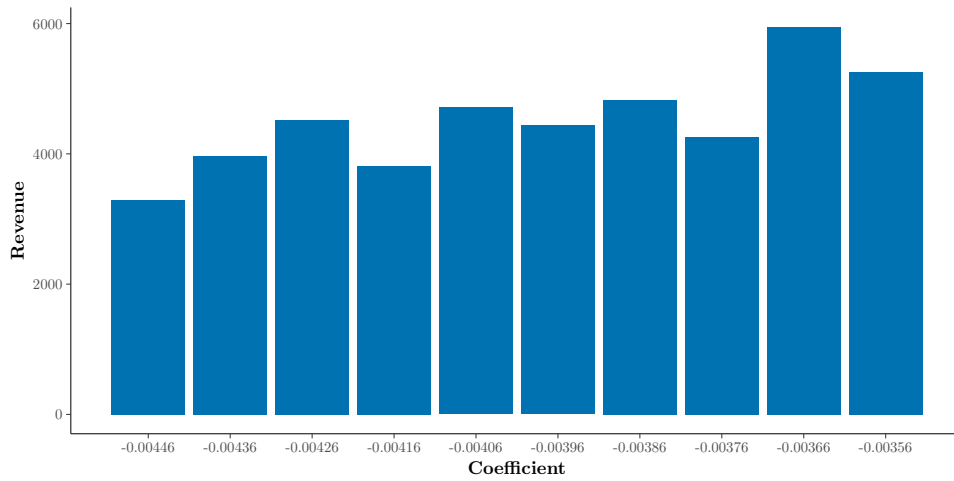


Figure 4: Revenue for variation of the price coefficient β^{price} within its standard error.

Market Shares					
β^{price}	Fare Class 1	Fare Class 2	Fare Class 3	Fare Class 4	Fare Class 5
-0.00446	0.002	0.092	0.12	0.364	0.422
-0.00436	0	0.062	0.274	0.256	0.408
-0.00426	0.012	0.208	0.236	0.134	0.41
-0.00416	0.004	0.096	0.306	0.194	0.40
-0.00406	0.01	0.258	0.282	0.04	0.41
-0.00396	0.014	0.236	0.162	0.11	0.478
-0.00386	0.02	0.236	0.346	0.062	0.336
-0.00376	0.008	0.188	0.33	0.07	0.404
-0.00366	0.032	0.164	0.488	0.07	0.246
-0.00356	0.03	0.19	0.484	0	0.296

Table 5: Market shares per fare class for decreasing price sensitivity.

$ S $	$ N $	c	R^*	CPU	GAP
10	100	50	43,392.75	64.15	0.0
		100	48,650.25	5.89	0.0
		200	48,650.25	5.91	0.0
	200	50	55,820.31	26.17	0.0
		100	88,749.03	2288.57	0.0
		200	99,031.25	16.61	0.0
20	100	50	42,520.26	7200.22	0.0
		100	47,141.77	17.28	0.0
		200	47,141.77	17.38	0.0
	200	50	54,908.97	6071.55	0.05
		100	86,551.71	7200.42	0.58
		200	94,929.50	47.29	0.0
50	100	50	39,448.77	7200.51	5.11
		100	44,342.89	93.48	0.0
		200	44,342.89	89.22	0.0
	200	50	48,887.20	7201.07	10.18
		100	7,425.12	2015.76	90.53
		200	91,696.20	250.27	0.0

Table 6: The values for R^* , CPU, GAP are average values over ten randomly generated instances. CPU denotes the time in seconds used by CPLEX to solve the respective problem. We set the maximum computational time to 7200 seconds. GAP denotes the gap reported by CPLEX in %.

5. Conclusion

With this article, we provide a formulation of a choice-based revenue management optimization model with flexible demand substitution patterns. We utilize a demand model that is based on individual utility values in the context of an airlines' fare class choice problem. The formulation of the mathematical model that we provide is general in that it allows for the incorporation of any discrete choice model. However, our approach particularly focuses on a demand model that captures non-constant substitution between fare classes. Thereby, we are able to overcome the shortcomings of the multinomial logit demand model that is frequently applied in revenue management optimization models. By applying a general random utility model in a simulation-based mixed-integer linear program for revenue maximization we are able to validate our approach. We find that the consideration of demand dependencies between fare classes improves revenues. Differences between the NL and MNL strategies arise in particular from fare classes offered to the generated individuals in combination with ticket prices. We observe that buy up's to higher priced fare classes occur according to the respective strategy. In the NL strategy, this leads to a higher market share for fare class 3 which serves as the closest substitute for fare class 4 while in the MNL strategy substitution patterns are constant among all remaining fare classes. Furthermore, the opt-out alternative is less frequently chosen in the NL strategy resulting in an overall higher number of ticket sales than in the MNL strategy. Thus, in combination with fare class prices set by the CBRM model we observe higher revenues by assuming flexible demand substitution patterns between fare classes.

However, for further research we suggest to address the following issues: (i) The obtained validation results should be confirmed by applying non-artificial demand data to the optimization model to confirm our findings. (ii) The proposed numerical example provides valuable insights into the mechanisms of our optimization model. However, in terms of revenue management applications large problem sets need to be solved to account for a large number of flights within an airlines' network. In particular, computational complexity needs to be taken into account when multiple sets of utility values have to be generated for large numbers of individuals. As shown in the numerical investigation, solving problem sets where $|S| \geq 50$ and $|N| > c$ is more complex. Therefore, further research should be dedicated solving problem sets with a large number of generated sets of utility values S where $|N| \gg c$.

Appendix A. CBRM Model formulation

$$\text{maximize } R = \frac{1}{|S|} \cdot \sum_s \sum_n \sum_f \Pi_{snf}$$

$$\begin{aligned} \Pi_{snf} &\leq \bar{p}_f \cdot Y_{snf} && \forall s, n, f \neq \text{opt-out} \\ \Pi_{snf} &\leq p_f && \forall s, n, f \neq \text{opt-out} \\ U_{snf} &= \delta_{snf} - \beta^{\text{price}} \cdot p_f - (\delta_{snf} + 1) \cdot (1 - X_{nf}) && \forall s, n, f \neq \text{opt-out} \\ U_{sn}^{\max} &\geq U_{snf} && \forall s, n, f \neq \text{opt-out} \\ U_{sn}^{\max} &\leq L \cdot (1 - Y_{snf}) + U_{snf} && \forall s, n, f \neq \text{opt-out} \\ \sum_n \sum_{f \neq \text{opt-out}} Y_{snf} &\leq c && \forall s \\ \sum_{f \neq \text{opt-out}} Y_{snf} &\leq 1 && \forall s, n \\ Y_{snf} &\leq X_{nf} && \forall s, n, f \neq \text{opt-out} \\ \Pi_{snf} &\geq 0 && \forall s, n, f \\ p_f &\geq 0 && \forall f \\ U_{snf} &\in \mathbb{R} && \forall s, n, f \\ U_{sn}^{\max} &\in \mathbb{R}^{\geq \delta_{n, \text{opt-out}, s}} && \forall s, n \\ X_{nf} &\in \{0, 1\} && \forall n, f \\ Y_{snf} &\in \{0, 1\} && \forall s, n, f \end{aligned}$$

References

- Ben-Akiva, M. E., Lerman, S. R., 1985. *Discrete Choice Analysis, Theory and Application to Travel Demand*. MIT Press, Cambridge, MA.
- Benati, S., Hansen, P., 2002. The maximum capture problem with random utilities: Problem formulation and algorithms. *European Journal of Operational Research* 143, 518–530.
- Berbeglia, G., Joret, G., 2015. Assortment optimisation under a general discrete choice model: A tight analysis of revenue-ordered assortments. Available at SSRN: <https://ssrn.com/abstract=2620165> or <http://dx.doi.org/10.2139/ssrn.2620165>.
- Bierlaire, M., 2003. BIOGEME: A Free Package for the Estimation of Discrete Choice Models. In: 3rd Swiss Transport Research Conference.
- Bierlaire, M., Azadeh, S. S., 2016. Demand-based discrete optimization. Tech. rep., École Polytechnique Fédérale de Lausanne.
- Carrier, E., 2003. Modeling airline passenger choice: Passenger preference for schedule in the passenger origin-destination simulator (pods). Ph.D. thesis, Citeseer.
- Davis, J., Gallego, G., Topaloglu, H., 2013. Assortment planning under the multinomial logit model with totally unimodular constraint structures. Department of IEOR, Columbia University. Available at http://www.columbia.edu/~gmg2/logit_const.pdf.
- Etebari, F., Najafi, A., 2016. Intelligent choice-based network revenue management. *Scientia Iranica. Transaction E, Industrial Engineering* 23 (2), 747–756.
- Fiig, T., Isler, K., Hopperstad, C., Belobaba, P., 2010. Optimization of mixed fare structures: Theory and applications. *Journal of Revenue & Pricing Management* 9 (1), 152–170.
- Gallego, G., Iyengar, G., Phillips, R., Dubey, A., 2004. Managing flexible products on a network. Department of Industrial Engineering and Operations Research, Columbia University, CORC Technical Report TR-2004-01.

URL <http://www.corc.ieor.columbia.edu/reports/techreports/tr-2004-01.pdf>

- Gallego, G., Li, L., Ratliff, R., 2009. Choice-based EMSR methods for single-leg revenue management with demand dependencies. *Journal of Revenue and Pricing Management*, 8 2 (3), 207–240.
- Gallego, G., Topaloglu, H., 2014. Constrained assortment optimization for the nested logit model. *Management Science* 60 (10), 2583–2601.
- Garrow, L., Ferguson, M., Keskinocak, P., Swann, J., 2006. Expert opinions: Current pricing and revenue management practice across us industries. *Journal of revenue and pricing management* 5 (3), 237–247.
- Haase, K., 2009. Discrete location planning. Tech. Rep. WP-09-07, Institute for Transport and Logistics Studies, University of Sydney.
- Haase, K., Müller, S., 2013. Management of school locations allowing for free school choice. *Omega* 41 (5), 847–855.
- Haase, K., Müller, S., 2014. A comparison of linear reformulations for multinomial logit choice probabilities in facility location models. *European Journal of Operational Research* 232 (3), 689–691.
- Klier, M., Haase, K., 2015. Urban public transit network optimization with flexible demand. *OR Spectrum* 37 (1), 195–215.
- Kocabiyıkoğlu, A., Popescu, I., Stefanescu, C., 2013. Pricing and revenue management: The value of coordination. *Management Science* 60 (3), 730–752.
- Krohn, R., Müller, S., Haase, K., 2017. Preventive health care facility location planning problem (working paper). Tech. rep., Hamburg Business School.
- Marschak, J., 1960. Binary-choice constraints and random utility indicators. In: *Mathematical Methods in the Social Sciences*. Stanford University Press, pp. 312–329.
- McFadden, D., 1974. Conditional logit analysis of qualitative choice behaviour. In: Zarembka, P. (Ed.), *Frontiers of Econometrics*. Academic Press, New York, pp. 105–142.

- McFadden, D., 2001. Economic choices. *American Economic Review* 91 (3), 351–378.
- McGill, J., Van Ryzin, G., 1999. Revenue management: Research overview and prospects. *Transportation science* 33 (2), 233–256.
URL <http://ben-israel.rutgers.edu/711/McGill-VanRyzin.pdf>
- Müller, S., Haase, K., 2016. On the product portfolio planning problem with customer–engineering interaction. *Operations Research Letters* 44 (3), 390–393.
- Müller, S., Haase, K., 2017. Revenue maximization tariff zone planning in public transportation. working paper. Tech. rep., Hamburg Business School.
- Seidel, F., 2011. Simulating Fare Class Choice Behavior with Flexible Substitution Patterns in Airline Revenue Management, Conference Talk, International Conference on Operations Research, Zürich.
- Seidel, F., 2012. Choice-Based Airline Revenue Management with Flexible Substitution Patterns, Conference Talk, International Annual Conference of the German OR Society, Hannover.
- Seidel, F., 2013. Fare Class Choice with Flexible Substitution Patterns, Conference Talk, EURO-INFORMS, 26th Conference on Operational Research, Rome.
- Seidel, F., 2014. Synthetic data sets with non-constant substitution patterns for fare class choice. *Zeitschrift für Verkehrswissenschaft* 85 (1), 32–55.
- Stigler, G. J., 1987. *The theory of price*. Macmillan, London, UK.
- Talluri, K., Van Ryzin, G., 2004a. Revenue management under a general discrete choice model of consumer behavior. *Management Science* 50 (1), 15–33.
- Talluri, K., Van Ryzin, G., 2004b. *The Theory and Practice of Revenue Management*. Springer Verlag.
- Talluri, K. T., Van Ryzin, G. J., Karaesmen, I. Z., Vulcano, G. J., 2008. Revenue management: models and methods. In: *Proceedings of the 40th*

Conference on Winter Simulation. Winter Simulation Conference, pp. 145–156.

Train, Kenneth, E., 2009. Discrete choice methods with simulation. Cambridge University Press.

Vulcano, G., Van Ryzin, G., Chahr, W., 2010. Choice-based revenue management: An empirical study of estimation and optimization. *Manufacturing & Service Operations Management* 12 (3), 371–392.

Weatherford, L., Ratliff, R., 2010. Review of revenue management methods with dependent demands. *Journal of Revenue & Pricing Management* 9 (4), 326–340.

Williamson, E., 1992. Airline network seat inventory control: Methodologies and revenue impacts. Ph.D. thesis, Massachusetts Institute of Technology.

Zhang, M., Bell, P., 2010. Price fencing in the practice of revenue management: An overview and taxonomy. *Journal of Revenue & Pricing Management* 11 (2), 146–159.

A Eidesstattliche Versicherung

Hiermit erkläre ich, **Frauke Korfmann geb. Seidel**, an Eides statt, dass ich die Dissertation mit dem Titel

“Essays on Advanced Discrete Choice Applications”

selbständig - und bei einer Zusammenarbeit mit anderen Wissenschaftlern - gemäß der beigefügten Darlegung -

nach §6 Abs. 3 der Promotionsordnung der Fakultät für Wirtschafts- und Sozialwissenschaften vom 24. August 2010 -

verfasst und keine anderen als die von mir angegebenen Hilfsmittel benutzt habe. Die den herangezogenen Werken wörtlich oder sinngemäß entnommenen Stellen sind als solche gekennzeichnet.

Ich versichere, dass ich keine kommerzielle Promotionsberatung in Anspruch genommen habe und die Arbeit nicht schon in einem früheren Promotionsverfahren im In- oder Ausland angenommen oder als ungenügend beurteilt worden ist.

Frauke Korfmann