Retail Demand Estimation and Merger Screening with Local Choice Sets

I N A U G U R A L - D I S S E R T A T I O N

to obtain the degree of
Doktor der Wirtschaftswissenschaften
submitted to

the Faculty of Business Administration and Economics
at the Heinrich Heine University Düsseldorf

presented by:
Lars Marten Zeigermann

Supervisor: Prof. Dr. Tomaso Duso
Head of Firms and Markets Department
DIW Berlin

5 September 2025

Acknowledgement

I owe a huge debt of gratitude to my supervisor, Professor Dr. Tomaso Duso, for his support and encouragement to complete this dissertation. I would also like to extend my thanks to Professor Dr. Justus Haucap for kindly taking on the role of my second supervisor.

Finally, I would like to thank my family and friends for enjoying the good moments of my doctoral studies with me, and for supporting me through the remaining 95 per cent of the time.

Contents

List of Figures

List of Tables

Chapter 1

General Introduction

Grocery retailing has gained much attention in both the economic and marketing literature and is among the most intensively researched industries. This is particularly true for empirical research. Factors that have contributed to this are: a) the availability of extensive and highly detailed data, b) increasingly sophisticated models of demand (and supply) to study consumer behaviour, compute counterfactuals and perform policy simulations, and c) the advent of advanced econometric software and powerful computers.

Although researchers have put tremendous effort into refining the underlying demand models - often some form of discrete choice model[1] - and the structural models reflecting the specifics of grocery retailing markets, one important aspect has been widely neglected: The existing literature generally relies on the often implicit assumption that consumers have access to all supermarkets and their products, ignoring that the local retail landscape is often very heterogeneous.

As an illustration, consider the distribution of supermarkets in the German federal state of Rhineland-Palatinate, displayed in Figure 1.1 below. Each dot represents a supermarket, and the colour indicates the retailer the supermarket belongs to.[2]

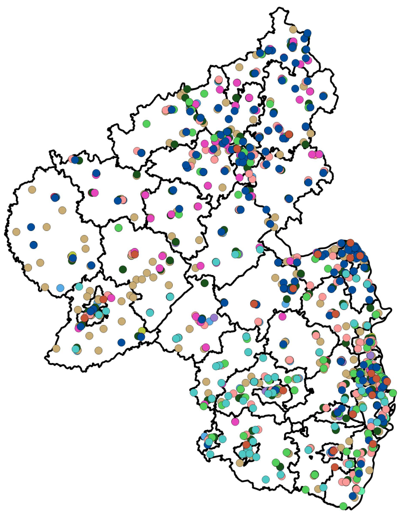


Figure 1.1: Supermarket locations in Rhineland-Palatinate by retailer[3]

[1] For a comprehensive overview of discrete choice models most often used in empirical studies of the retail industry, see e.g. Train (2009) and Greene (2012), chapter 21.

[2] The dataset was obtained by from Wer-zu-Wem GmbH, a company address database provider.

[3] The shapefile was downloaded from `https://www.suche-postleitzahl.org`.

Households living in the more densely populated areas in the north and south-east have access to a larger number of different retailers. In more rural areas in central Rhineland-Palatinate and the west, markedly fewer retailers are available. While it may be realistic to assume that the leading national retailers are available to more or less all consumers (left panel of Figure 1.2), this assumption is hardly justified for their smaller competitors (right panel of Figure 1.2).

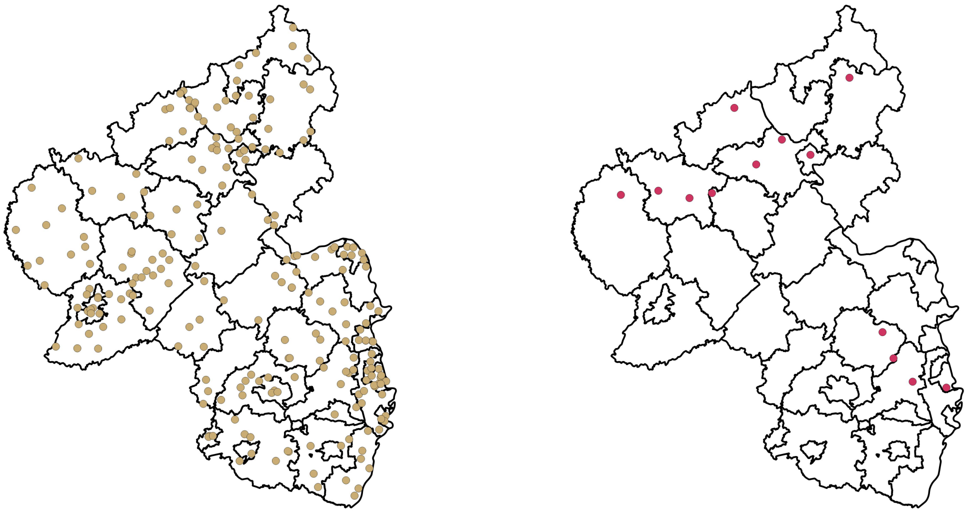


Figure 1.2: Supermarket locations of a leading national retailer (left panel) and a smaller competitor (right panel) in Rhineland-Palatinate[4]

The contribution of this dissertation is twofold: In Chapters 2 and 3, I demonstrate how ignoring heterogeneity in the local retail landscape leads to biased demand estimates and distorts popular measures such as price elasticities and upward-pricing pressure (UPP) - a popular merger screening tool - inferred from structural models of demand and supply. Chapters 4, 5, and 6 present routines written for the widely used statistical software package Stata. The first allows users to compute various forms of elasticities and marginal effects for mixed logit models. The second enables users to obtain geographical coordinates from addresses and to translate coordinates into addresses. The third routine is for computing various distance measures.

Chapter 2, **Estimating Retail Demand and Price Elasticities with Local Choice Sets**, demonstrates how relaxing the common - and often implicit -

[4]The shapefile was downloaded from `https://www.suche-postleitzahl.org`.

assumption of a global choice set containing all products in favour of local choice sets reflecting the local retail landscape affects logit demand estimation results and price elasticities based thereon.

To address this issue, I use two main data sources: the 2010 GfK (Gesellschaft für Konsumforschung) consumer panel, which comes with additional information on household locations and a comprehensive address list of all supermarkets in Germany provided by the company address database provider Wer-zu-Wem GmbH.

Local choice sets are constructed following a simple rule: All products offered by retailers with at least one supermarket within a 10-kilometre radius of the approximated household location are included. As a consequence, the average number of retailers available to a household drops from 13 to 8, and the number of products to choose from goes down by on average more than 30 per cent.

With local choice sets, demand estimates and price elasticities change substantially. Retailer fixed-effects, indicating the popularity of a retailer, increase for smaller retailers. All cross-price elasticities between products offered by the same retailer rise. The difference between choice set specifications is the largest for the smallest retailer with the fewest supermarkets, and it is close to 700 per cent. Elasticities between products offered at two competing retailers change in either direction, ranging from a decrease of 50 per cent to an increase of 95 per cent. Important factors driving the result are the number of local choice sets each of the two retailers is contained in and the number of choice sets both retailers share.

Chapter 3, **Retail Merger Screening with Local Choice Sets**, builds on the previous chapter and compares predicted merger outcomes obtained under the common assumption of global choice sets with those derived with local choice sets. To illustrate the effect, I use the popular and widely used upward-pricing pressure (UPP) that indicates which mergers are likely to give rise to anticompetitive outcomes. This fairly simple screening tool does not require full-blown merger simulation and, therefore, allows for an evaluation of all 78 possible mergers between the 13 retail chains in the dataset.

Comparing both choice set specifications reveals that UPP can be smaller or larger with local choice sets. The percentage difference ranges from -47 per cent to 119 per cent. One striking result is that the difference between both choice set scenarios is strongest when UPP is small, and the merger is unlikely to be detrimental to competition. In cases where UPP is high and the merger is more likely to raise concerns, UPP results are fairly similar with both choice set specifications.

Chapter 4, **Mixlelast: Stata Module for Mixed Logit Elasticities and Marginal Effects**, describes a post-estimation command for `mixlogit`, a user-written Stata routine for the estimation of mixed logit models (Hole 2007). While `mixlogit` provides a straightforward solution for estimating mixed logit models, no routine existed to compute elasticities and marginal effects. Since the computation of mixed logit elasticities and marginal effects requires simulation techniques and hence advanced programming skills, the lack of a post-estimation routine presented a significant obstacle for practitioners and researchers alike to make full use of `mixlogit`. `mixlelast` fills this gap and provides an easy-to-use solution for computing various forms of marginal effects and elasticities, including an option for calculating bootstrapped standard errors.

Chapter 5, **Opencagegeo: Stata Module for Geocoding**, describes a geocoding routine for Stata. Geocoding is the computational process of converting addresses into geographic coordinates (e.g., latitudes and longitudes).[5] For example, `opencagegeo` would tell us that the Düsseldorf Institute for Competition Economics (DICE) of the Heinrich-Heine-University Düsseldorf located at the Oeconomicum, Universitätsstraße, 40225 Düsseldorf - where the better part of this dissertation was written - is precisely located at 51.1888412, 6.7934205. `opencagegeo` is further able to reverse this process, i.e. it converts geographic coordinates into a readable address, a process usually referred to as reverse geocoding. E.g. `opencagegeo` would convert 50.7405974, 7.1881607 into the address where this dissertation was finalised.

Chapter 6, **Lhaversine: Mata Library for Distance Functions**, presents several functions for Mata, Stata's matrix programming language, to compute distances and distance-related measures. All functions contained in `lhaversine` use the haversine formula for great circle distances[6] and take latitudes and longitudes as inputs. `lhaversine` can produce a distance matrice for a large number of locations, a matrix indicating which locations are within a specified range and a vector containing the number of locations within a given distance.

Chapter 7 summarises the main findings and concludes.

[5] `opencagegeo` uses the application programming interface (API) of OpenCage Data, a professional web-based geocoding service accessing different open data sources such as OpenStreetMap (OSM).

[6] See van Brummelen (2013), chapter 9 for an illustrative derivation of the haversine formula.

References

Greene, W. H. (2012). *Econometric Analysis.* Pearson, 7th edition.

Hole, A. R. (2007). Estimating mixed logit models using maximum simulated likelihood. *Stata Journal*, 7(3):388–401.

Train, K. (2009). *Discrete Choice Methods with Simulation.* Cambridge University Press, 2nd edition.

van Brummelen, G. (2013). *Heavenly Mathematics: The Forgotten Art of Spherical Trigonometry.* Princeton University Press.

Chapter 2

Estimating Retail Demand and Price Elasticities with Local Choice Sets

2.1 Introduction

A common and often implicit assumption in discrete choice modelling is that all decision-makers choose from an identical set of alternatives, which is sometimes referred to as the *global* or *universal* choice set. Using a global choice set in the context of grocery retailing is equivalent to assuming that all retailers - and hence all products they offer - are available to all consumers. This assumption is particularly unrealistic in sparsely populated rural areas, where consumers can - at best - choose between a few different retailers unless they are willing to travel vast distances. Even in urban areas, where supermarket density and variety are generally higher, not all retailers are necessarily present in all local areas. For example, there are smaller retail chains that focus their business activities on some regional or even local markets where they are a relevant player but do not have any supermarkets in others.

Ignoring heterogeneity in choice sets in demand estimation, we incorrectly model the underlying choice process by falsely including alternatives that were not available or considered by the decision maker. It is well established that this kind of misspecification can give rise to severely biased estimation results, see e.g. Bronnenberg and Vanhonacker (1996); Bajari and Benkard (2005); Goeree (2008); Draganska and Klapper (2011). These effects necessarily carry over to the wide range of popular measures (such as elasticities and marginal effects) and structural (supply) models (e.g. merger and policy simulations) used by researchers and practitioners alike.[1]

The economics literature on retailing demand and supply has largely ignored heterogeneity in choice sets. Some examples from grocery retailing include Bonnet and Dubois (2010), Bonnet et al. (2022), Haucap et al. (2021) and Villas-Boas (2007). However, a growing body of literature rooted in marketing has dealt with choice set heterogeneity in several ways, especially in grocery retailing. Hickman and Mortimer (2016) provide a comprehensive overview of the literature that deviates from the common global choice set assumption in favour of choice sets that contain only a subset of the global set of alternatives. They distinguish two main strands of the literature.

The first strand recognises that consumers usually do not *consider* all of the potentially available products in their individual choice process, but limit the global set of alternatives to an individual subset of alternatives between which they choose. Reasons for reducing the alternatives to choose from could be limited memory or knowledge or costly information acquisition. These reduced choice sets are com-

[1] Reiss and Wolak (2007) provide a detailed overview of structural economic modelling.

monly referred to as consideration sets. As consideration sets are usually unobserved, researchers have tried to model the formation process. This is often done in a two-stage approach which models consideration set formation in the first stage and the product choice conditional on the consideration set in the second stage.[2]

The second strand of the literature considers that the alternatives actually *available* for choice can vary. Assortment variations may occur across regional markets, between different retailers, or over time. Examples in the literature that address assortment variation in the choice model are e.g. Draganska et al. (2009) and Shah et al. (2015). Another source of variation of available products is temporary stock-outs, i.e. choice situations in which one or several alternatives are not in stock and, therefore, not relevant in the choice process. The effects of stock-outs on demand estimation results are studied by Conlon and Mortimer (2013).

Although it is well documented that consumers usually do their grocery shopping where they live and travel distances are fairly short[3], geo-referenced data on household and retailer locations to construct choice sets has gained little attention in models of retailing demand (and supply).[4] The main goal of this paper is to illustrate how relaxing the global choice set assumption in favour of local choice sets that reflect actual local retailing structure impacts demand estimates and own- and cross-price elasticities in a traditional discrete choice framework.

For this purpose, I combine two main datasets. Choice data is taken from the German GfK (Gesellschaft für Konsumforschung) consumer panel, including information on the approximate household location. Address data for all German supermarkets is provided by Wer-zu-Wem GmbH. To construct local choice sets, I use a simple rule: a decision-maker can choose between all products offered at supermarkets that are within 10 km of the approximated household location.[5]

Applying this rule significantly affects the composition of choice sets. While large retailers that maintain a tight net of supermarkets are included in almost all local

[2]Examples are Barroso Ludena and Llobet (2012), Draganska and Klapper (2011), Goeree (2008), Nierop et al. (2010) and Pancras (2010). Alternative approaches adjust the utility by including observables or doing structural research on the formation process of the set of alternatives considered.

[3]For Germany, see the detailed analysis of distances travelled for grocery shopping by Krüger et al. (2013).

[4]A notable exception is Lu (2017), who uses travel distances in the consideration set formation stage of a two-stage approach.

[5]It is possible to construct more advanced local choice sets taking into account that travel distances vary by type of supermarket and depend on e.g. age, the availability of a car in the household, etc. However, as the main goal of this paper is to systematically compare and contrast results derived under the global choice sets assumption with those from local choice sets, it is sufficient to us a simple rule for choice set formation.

choice sets, smaller retailers with fewer supermarkets are only present in a small fraction of all local choice sets. Equally importantly, the location of supermarkets also has a significant impact on the number of choice sets in which two competing retailers are present and consumers can choose between them: if two competing retailers have supermarkets in similar areas, they will necessarily have more choice sets in common than two retailers whose business activities focus on different areas.

The logit demand estimation results indicate that consumer preferences for supermarkets are systematically biased if the local retail structure is ignored. Under the global choice set assumption, smaller retailers are deemed less popular than their larger competitors. With local choice sets, we see that they can be equally popular as big national chains where available; the reason why fewer consumers choose their products is simply because they are not locally available for many consumers. Substitution patterns measured by cross-price elasticities change dramatically once we use local choice sets. With local choice sets, cross-price elasticities between products held by the same retailer are uniformly higher. This result intuitively makes sense, as two products held by the same retailer are always available together and hence are an available alternative. The increase is stronger for smaller retailers, which are included in a small number of local choice sets.

The effects on substitution patterns between retailers are even stronger when comparing local and global choice sets, and can go either way. Further inspection of these results reveals two key factors driving the differences in cross-price elasticities. First, the percentage change in cross-price elasticities resulting from local choice sets is c.p. decreasing in the number of choice sets a retailer is contained in. For this reason, it is most pronounced for smaller retailers. Second, the percentage change increases c.p. with the number of common local choice sets. In other words, the more frequently both supermarkets are included in the same choice set, the greater the percentage difference in elasticities between choice set specifications.

The remainder of this paper is organised as follows. In Section 2.2, the conditional logit demand model and the resulting elasticities are derived for global and local choice sets. Section 2.3 provides a detailed description of the data. Local choice sets are constructed in Section 2.4. Demand estimation results and elasticities are presented and compared in Section 2.5. Section 2.6 concludes.

2.2 Demand Model and Elasticities

This section presents the conditional logit model and derives expressions for the predicted own- and cross-price elasticities. This is done first under the common assumption of a global choice set, and then for local choice sets.

Although there exist more sophisticated discrete choice models, e.g. the now widely used random coefficient models[6], the conditional logit model (McFadden 1973) is used in what follows. The simplicity of this traditional choice model and its mathematical properties significantly reduce the complexity of the mathematical expressions and hence facilitate pinning down the, thus facilitating the identification of differences between both choice sets specifications. In addition, avoiding the complexities (and caveats) of more advanced discrete choice models makes it markedly easier to illustrate the effects of relaxing the global choice set assumption in favour of (more realistic) local choice sets in the empirical application presented in Section 2.5.

2.2.1 Global Choice Set

Demand Consider a population of consumers indexed by $n \in \mathcal{N} = \{1, \ldots, N\}$ making T repeated choices. At each choice occasion t, the decision-makers choose one alternative $j \in \mathcal{J} = \{1, \ldots, J\}$. The choice set $\mathcal{J}$ contains all J alternatives and is identical across all N decision makers and all T choice occasions. Following the convention in the literature, $\mathcal{J}$ will be referred to as the global choice sets in what follows. Further assume there is a set of retailers indexed by $r \in \mathcal{R} = \{1, \ldots, R\}$. Each retailer r offers a subset of the products contained in the global choice set, i.e. $\mathcal{J}^r \subset \mathcal{J}$. Further, assume that each product j is sold by one unique retailer, hence $\mathcal{J}^r \cap \mathcal{J}^s = \emptyset, \forall r \neq s$. Thus, assuming all decision-makers can choose from the global choice set implies that all decision-makers have access to all retailers.

The indirect utility of consumer n obtains from buying product i at choice occasion t is given by

$$U_{nit} = \underbrace{\alpha p_{nit} + x_{nit}\beta_i}_{V_{nit}} + \varepsilon_{nit}, \tag{2.1}$$

where x_{nit} is a vector of observed product characteristics and p_{nit} denotes the price of product i at choice occasion t. ε_{nit} is a mean-zero, iid extreme value I distributed individual-specific random shock. For ease of notation, I will denote the deterministic portion of the n's utility from buying product i at choice occasion t by V_{nit}.

[6]See Train (2009) for a detailed description.

The probability for consumer n of buying product i at choice occasion t, i.e. $U_{nit} > U_{njt} \; \forall \; j \neq i$, can then be written as

$$P_{nit}^{G} = \frac{\exp(V_{nit})}{\sum_{j=1}^{J} \exp(V_{njt})}, \tag{2.2}$$

where the superscript G indicates global choice sets. Note that the expression in (2.2) is always positive, i.e. we obtain positive choice probabilities for all J alternatives in each choice situation. The sum over all J choice probabilities for each individual at each choice occasion necessarily adds up to one.[7]

The (predicted) market share of an alternative i is then the average choice probability over the sample of all N individual decision-makers over all T choice occasions. Formally,

$$s_i^G = \frac{1}{N}\frac{1}{T}\sum_{n=1}^{N}\sum_{t=1}^{T} P_{nit}^{G}. \tag{2.3}$$

Elasticities Having established an expression for (predicted) market shares, we can now derive elasticities.

In general, elasticities of market shares describe the percentage change in the market share following a one per cent increase in one of the product characteristics of one alternative. Elasticities can be computed for any product characteristic[8], but elasticities with respect to price are the most relevant both in the academic literature and in practice. Therefore, only price elasticities are considered in what follows. Nonetheless, the insights gained apply equally to any other product characteristic contained in x.

Own-price elasticities measure the percentage change in the market share of an alternative if that product's own price goes up by one per cent, ceteris paribus. They are, hence, a measure of how sensitive consumers react to increasing prices. As demand usually falls in prices, we expect own-price elasticities to be negative. The higher the own-price elasticity in absolute terms, the stronger the drop in market share of the affected product.

[7]Sometimes, an outside option is included, which captures the fact that the decision-maker might choose not to buy any available alternative. Since including an outside option creates further challenges and adds complexity, I refrain from doing so in this paper. This simplification is also justified because I am using data on fluid milk, a low-priced, fast-moving consumer and staple good, where potential buyers are unlikely to opt against purchasing because of product characteristics.

[8]Note that it is not meaningful to compute the impact of a one per cent change in a dummy variable. With dummy variables, a discrete change from one to zero (or vice versa) needs to be evaluated instead, see Chapter 4.3.

Formally, the own-price elasticity of the market share of product i is defined as

$$E_{ii} = \frac{\bar{p}_i}{s_i} \frac{\partial s_i}{\partial p_i}, \tag{2.4}$$

where $\bar{p}_i$ denotes the average price of alternative i across all choice occasions and decision makers. Plugging in the expression for the predicted market share of i derived in (2.3) yields

$$E_{ii}^G = \frac{\bar{p}_i}{s_i^G} \frac{1}{N} \frac{1}{T} \sum_{n=1}^{N} \sum_{t=1}^{T} \frac{\partial P_{nit}^G}{\partial p_i}. \tag{2.5}$$

That is, the own-price market share elasticity of i is simply the average over the population's individual marginal effects, $\partial P_{nit}^G / \partial p_i$, multiplied by the mean price and the market share. As Train (2009) shows, taking the derivative in (2.5) gives

$$E_{ii}^G = \frac{\bar{p}_i}{s_i^G} \frac{1}{N} \frac{1}{T} \sum_{n=1}^{N} \sum_{t=1}^{T} \alpha P_{nit}^G (1 - P_{nit}^G). \tag{2.6}$$

Since the choice probability P_{nit}^G is always positive, individual marginal effects, $\alpha P_{nit}(1 - P_{nit}^G)$, will be negative for all decision-makers and products, given the price coefficient, α, is negative as expected. That is, all decision-makers are affected by the price change in i, and the change in the market share of i following the price is thus driven by all N decision-makers. Note that the individual marginal effects increase in P_{nit} in absolute terms and reach their maximum at $P_{nit} = 0.5$. Provided that the number of available alternatives J is sufficiently large and the choice probabilities are below 0.5, we find the demand for products with a high market share to be more elastic.

In contrast, cross-price elasticities of market shares give the percentage change of the market share of a product if the price of a rival's product price increases by one per cent, ceteris paribus. Here, we would expect positive values as consumers will, to some degree, substitute the now relatively more expensive product for an alternative. The cross-price elasticity of i's market share with respect to a price increase in j is

$$E_{ij}^G = \frac{\bar{p}_j}{s_i^G} \frac{\partial s_i^G}{\partial p_j} \tag{2.7}$$

$$= \frac{\bar{p}_j}{s_i^G} \frac{1}{N} \frac{1}{T} \sum_{n=1}^{N} \sum_{t=1}^{T} \frac{\partial P_{nit}^G}{\partial p_j}. \tag{2.8}$$

Taking the first derivative yields

$$E_{ij}^G = \frac{\bar{p}_j}{s_i^G}\frac{1}{N}\frac{1}{T}\sum_{n=1}^{N}\sum_{t=1}^{T} -\alpha P_{nit}^G P_{njt}^G. \tag{2.9}$$

The individual marginal effects, $-\alpha P_{nit}^G P_{njt}^G$, are always non-zero and positive if the price coefficient α is negative as expected. Hence, the effect of a price increase in j on i's market share is driven by all N decision makers. The expression in (2.9) is increasing in the market share of product j. This result is intuitive, as we would expect the increase in the demand for i to be larger if the rival product, which is now more expensive, has a large market share.

2.2.2 Local Choice Sets

As before, there are N decision makers choosing one alternative per choice occasion. However, we no longer maintain the assumption of a global choice set, but allow the available alternatives to differ between decision-makers. I.e., a consumer may only have access to a subset of retailers, $\mathcal{R}^n \subseteq \mathcal{R}$. Recalling that each of the J alternatives is offered by only a single retailer, the set of products to choose from is determined by the set of available retailers $\mathcal{R}^n$, i.e. $\mathcal{J}^n = \cup_{r\in\mathcal{R}^n}\mathcal{J}^r$. Consequently, all alternatives offered by the retailers not included in $\mathcal{R}^n$ cannot be chosen. Because choice sets will reflect the local retail landscape in the empirical example, I refer to local choice sets in what follows.[9]

The utility function given in (2.1) remains unchanged, however, n's choice probability of alternative i at choice occasion t is now

$$P_{nit}^L = \begin{cases} \dfrac{\exp(V_{nit})}{\sum_{j\in\mathcal{J}^n}\exp(V_{njt})} & \text{if } i \in \mathcal{J}^n \\ 0 & \text{otherwise,} \end{cases} \tag{2.10}$$

where the superscript L indicates local choice sets.

If alternative i is available to n, i.e. $i \in \mathcal{J}^n$, the choice probability resembles the expression for the logit model maintaining the assumption of a global choice set, see (2.2). The difference lies in the denominator, which now only contains the products in n's local choice set, $j \in \mathcal{J}^n$, rather than all J alternatives. The choice probability for all products in a choice set is, ceteris paribus, increasing in the number

[9] A more general term sometimes used in the literature is *heterogeneous* or *individual* choice set.

of competing alternatives excluded from the local choice sets. This intuitively makes sense as it is more likely that i is chosen the fewer competing alternatives there are. The sum over the choice probabilities of all alternatives in decision-maker n's choice set $\mathcal{J}^n$ add up to one. For all remaining alternatives $i \notin \mathcal{J}^n$, the choice probability is necessarily zero.[10] As before, the market share of alternative i is defined as the average choice probability over all N decision-makers and T choice situations, i.e.

$$s_i^L = \frac{1}{N}\frac{1}{T}\sum_{n=1}^{N}\sum_{t=1}^{T} P_{nit}^L. \tag{2.11}$$

The fundamental difference is that only the subset of individuals with i in their choice sets, denoted by $\mathcal{N}^i$, contribute to the market share. For all other decision-makers $n \in \mathcal{N}^{-i}$, the individual choice probability is zero. This has far-reaching consequences for elasticities.

Elasticities The own-price elasticity of the i's market share with local choice sets is

$$E_{ii}^L = \frac{\bar{p}_i}{s_i^L}\frac{\partial s_i^L}{\partial p_i}. \tag{2.12}$$

Plugging in (2.11) and taking the derivative yields

$$E_{ii}^L = \frac{\bar{p}_i}{s_i^L}\frac{1}{N}\frac{1}{T}\sum_{n=1}^{N}\sum_{t=1}^{T} \alpha P_{nit}^L(1 - P_{nit}^L). \tag{2.13}$$

The individual marginal effects, $\alpha P_{nit}^L(1 - P_{nit}^L)$, are no longer different from zero for all decision makers[11], but only for those that can actually choose i, i.e. $n \in N^i$. For the remaining $n \in N^{-i}$, the effect is necessarily zero as they are unaffected by the price increase in i.

[10] Depending on the research question, it would only make sense to study the effect on decision-makers affected, i.e. those that have the respective alternative in their local choice set. However, as I study the impact on the market, it is necessary to include those decision-makers who do not have the relevant alternative in their local choice sets.

[11] Technically, the individual marginal effect is not defined for those individuals that do not have alternative i in their choice sets. For convenience, the individual effect is assumed to be zero in those cases.

Equivalently, cross-price elasticities are then

$$E_{ij}^L = \frac{\bar{p}_j}{s_i^L}\frac{\partial s_i^L}{\partial p_j} \tag{2.14}$$

$$= \frac{\bar{p}_j}{s_i^L}\frac{1}{N}\frac{1}{T}\sum_{n=1}^{N}\sum_{t=1}^{T} -\alpha P_{nit}^L P_{njt}^L. \tag{2.15}$$

For individual marginal effect, -$\alpha P_{nit}^L P_{njt}^L$, to be non-zero, both products i and j need to be available. Otherwise, substitution between the two alternatives is impossible; hence, a price change in j does not affect a decision-maker's choice concerning alternative i. Let's denote the subset of relevant decision-makers satisfying this condition by $\mathcal{N}^{ij}$, i.e. $\mathcal{N}^{ij} = \mathcal{N}^i \cap N^j$.

In the context of local choice sets depending on the availability of retailers, it is insightful to distinguish intra- and inter-retailer cross-price elasticities: intra-retailer cross-price elasticities are between two products offered by the same retailer (i.e. $r^i = r^j$) and inter-retailer cross-price elasticities are between two products sold by two competing retailers (i.e. $r^i \neq r^j$).

First, consider intra-retailer cross-price elasticities. Because the availability of a product in the choice set is determined by the presence (or absence) of a particular retailer, any pair of products i and j held by the same retailer, i.e. $r^i = r^j$, is consequently available to the same subset of consumers, $\mathcal{N}^{ij} = \mathcal{N}^i = \mathcal{N}^j$.

This does not hold for inter-retailer cross-price elasticities, i.e. $r^i \neq r^j$. In the one extreme case, either the products of retailer r^i or those of r^j are contained in a choice set, but never both at the same time. Then, no decision maker can choose between the two, i.e. $N^{ij} = \emptyset$, and there is consequently no substitution. This would be the case when the supermarkets of both retailers are located in distinct geographic areas. The price increase in j does not affect the market share of i, and the cross-price elasticity is zero. In the other extreme case, products i and j are either both present or both absent even though they are held by two competing retailers. This would be the case when both retailers have supermarkets in the same locations. Then, the same what was stated about intra-retailer elasticities applies here.

In intermediate cases, products i and j are available to some decision-makers, while others can only buy i or j or neither of the two. The effect of the price increase in j on the market share of i is then solely driven by the decision-makers with access to both, i.e. those contained in the subset N^{ij}. Consequently, inter-retailer cross-price elasticities are determined not only by the individual choice probabilities of products i and j and their overall availability in choice sets, but also crucially by the number of decision-makers with access to both alternatives.

In addition, the size of the local choice sets will, ceteris paribus, also affect the cross-price elasticity between two alternatives. The fewer alternatives there are in the local choice set, the higher are the choice probabilities and consequently the elasticities. This intuitively makes sense as the effect of a price rise in one product on another is stronger when fewer rival alternatives are available.

The consequences of relaxing the assumption of a unique global choice set for all decision-makers in favour of (more realistic) local choice sets are illustrated in an empirical example in what follows.

2.3 Data

This section describes the two main data sources, i.e. the 2010 German GfK (Gesellschaft für Konsumforschung) household panel and the Wer-zu-Wem supermarket address dataset. Furthermore, it explains why fluid milk[12] was chosen as the product category and why the analysis was restricted to German federal state of Rhineland-Palatinate.

Fluid milk was selected among various alternative products in the panel dataset for several reasons.[13] First, fluid milk is a fast-moving consumer good and a stable good which is frequently purchased by many households. For this reason, the number of observed purchases will be sufficiently large even if the analysis is restricted to one region. Second, fluid milk is a product which is differentiated based on a few discrete categories such as brand, fat content, whether is is fresh or UHT[14]. This guarantees the correct identification of products even if product identifiers in the dataset are inconsistent, changing over time or even missing (all of which is the case in the GfK consumer panel). Third, because the degree of product differentiation is fairly limited, the number of relevant products in terms of sales is relatively low compared to other product categories with a markedly higher variety of products. For this reason, restricting the products considered in the demand model to a reasonable number (which is commonly done in discrete choice modelling) does not lead to too many observations being dropped.

[12]Fluid milk is a commonly used industry term for milk processed for beverage use.

[13]Data was also available on coffee, yoghurt, deserts, toilet paper and diapers.

[14]UHT stands for ultra-high temperature processing, a procedure to increase shelf-life and allow for storage at room temperature.

Limiting the analysis to one federal state has two major advantages. First, the number of observations and the computational burden is reduced substantially. Second and more importantly, many products are offered only regionally. This is particularly true for fluid milk that is commonly sold close to the dairy. Limiting the dataset to a region ensures that the products considered are those available in the supermarket.

In the 2010 GfK panel, Rhineland-Palatinate is represented by 1,406 households that made 21,592 shopping trips where fluid milk was purchased.[15] Why Rhineland-Palatinate which has a population of about 4 million representing about 5 per cent of the German population[16] was selected for this study will be discussed at some length in the section on household location data below.

Choice data The GfK consumer panel dataset comprises actual purchases made by the panel households who scan their receipts using home scanning devices. The data contains information on the different products purchased, the actual price paid (including discounts and promotions), the quantity, and the retailer[17]. In addition, a number of product characteristics are provided. For fluid milk, these are the fat content and dummies for UHT, private labels[18] and organically sourced products.

In order to keep the number of products manageable, I only consider the 100 products that are chosen most frequently. Following the convention in the literature, a product is defined as the combination of a physical product and the retailer where it is sold.[19] That is, two physically identical products sold at different retailers are considered two distinct products. The total number of sales of the top 100 products considered is 21,592. This is about 97 per cent of all fluid milk purchases in Rhineland-Palatinate in 2010; only 3 per cent of all observations are dropped.

The main product characteristics of the top 100 products are summarised in Table 2.1 below.

Of the 100 products considered, 69 are private labels. The average price is approximately 0.70 Euros, however, it varies strongly ranging from 0.39 to almost

[15]The number of total sales is significantly larger as the number of cartons or bottles of milk bought per shopping trip is usually greater than one. This is particularly true for UHT products, which can be stored (unrefrigerated) over a long period of time. However, as discrete choice models generally allow us to model *what* is chosen, but not *how much* of it, I treat multiple units purchases of the same product as a single choice.

[16]`https://en.wikipedia.org/wiki/Rhineland-Palatinate`

[17]Only the retailer's name is given. The exact supermarket where the purchase was made is not disclosed.

[18]Private label products are produced for and sold under a retailer's brand.

[19]Technically, the retailer selling the product is treated as a product characteristic.

Variable	Mean	Std. Dev.	Min.	Max.
price	69.61	21.302	39	149
privlab (0=brand, 1=private label)	0.690	0.463	0	1
fat (per cent)	2.512	1.17	0.1	3.8
fresh (0=UHT, 1=fresh)	0.560	0.497	0	1
organic (0=conventional, 1=organic)	0.13	0.336	0	1
N		1200		

Table 2.1: Characteristics of top 100 products

1.50 Euros.[20] The mean fat content is 2.5 per cent. There are 56 fresh and 44 UHT products. 13 products are organic.

The top 100 products are held by 13 different retail chains; five discount stores and eight full-line distributors.[21] Retail format, number of products and market shares by retailer are displayed in Table 2.2 below.

	Format	# products	Market share
Retailer 1	D	5	29.78
Retailer 2	FL	12	6.96
Retailer 3	FL	11	6.45
Retailer 4	FL	6	1.10
Retailer 5	FL	9	5.30
Retailer 6	D	6	19.98
Retailer 7	FL	2	.31
Retailer 8	D	10	9.12
Retailer 9	D	4	2.66
Retailer 10	FL	4	4.39
Retailer 11	D	15	7.06
Retailer 12	FL	12	5.94
Retailer 13	FL	4	.95
Total		100	100

Table 2.2: Retail format, number of products and market share by retailer

Retailer market shares range from just 0.3 to almost 30 per cent. The aggregate market share of the five discount supermarkets is approximately 65 per cent and illustrates their strong position in the German grocery retailing market. The number of products in the top 100 per retailer ranges from 2 to 15.

[20] The GfK consumer panel only reports the price paid of the alternative the household actually purchased. The prices for the non-chosen alternatives in the choice set are monthly averages over the prices paid by other consumers. In very few cases where no price for a product for a given month was observed, yearly averages were computed instead.

[21] Due to the confidentiality agreement with GfK, the retailers cannot be named.

Household locations The GfK consumer panel comes with an additional dataset containing various household characteristics, including the five-digit postcode[22] of the place of residence of the household.[23] Due to the lack of more detailed information, households are assumed to be located at the geographic centre (so-called centroids) of their respective postcode areas. Therefore, the accuracy of the approximation depends on the size of the postcode area: the smaller the area, the more accurate the approximation of the true household location.

In Rhineland-Palatinate, the average postcode area covers 30.5 km^2, making it the German federal state with the smallest average postcode areas except for the city states Berlin, Hamburg and Bremen (see left panel of Figure A2.1 in the appendix).[24]

A further advantage of studying Rhineland-Palatinate is that the difference in size between rural and urban postcode areas is smaller than in other states. Therefore, even in rural areas where postcode areas are generally larger, the approximation will still be fairly accurate (see right panel of Figure A2.1 in the appendix). Visual comparison of the geographical centres of the postcode areas (i.e. the approximated household location) and the population centre of a postcode area (i.e. where the household is most likely to be located) confirms that they often coincide.

Of the 651 postcode areas in Rhineland-Palatinate, 439 (i.e. 67.4 per cent) are represented by panel households. To obtain the coordinates of the geographical centres of those postcode areas, i.e. the longitude and latitude, I use opencagegeo, an OpenStreetMap-based geocoding routine for Stata (Zeigermann 2016).

The locations of the 1,406 panel households are plotted in Figure 2.1.

[22]Five-digit postcodes are the most granular postcode unit used in Germany.

[23]For reasons of confidentiality, the exact household location is not disclosed.

[24]Because the three German city-states are very densely populated and have no rural areas where the number of competing retailers is limited, they are no good examples for studying the impact of local choice sets on demand estimates and (price) elasticities.

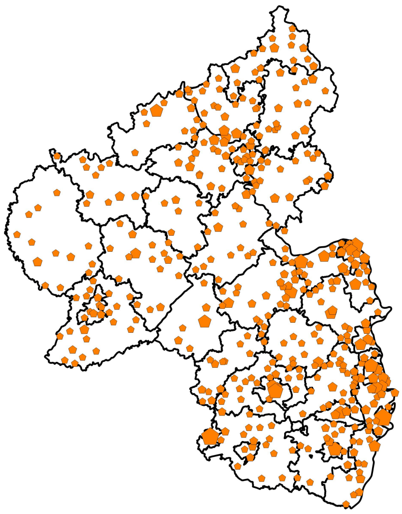


Figure 2.1: Locations of panel households in Rhineland-Palatinate[25]

The size of the symbol indicates the number of households in the respective postcode area. The map clearly shows that the more densely populated areas in the east and especially the south-east are represented by a larger number of panel members than the rural areas in the centre and the west of Rhineland-Palatinate. However, all parts of Rhineland-Palatinate are represented in the panel.

Retailer locations Address data for supermarkets is obtained from Wer-zu-Wem GmbH, a professional address database provider.[26] The addresses in the dataset come from various sources, including the German Federal Gazette (Bundesanzeiger), company registers and retailer websites, and are randomly being cross-checked manually. The dataset comprises around 27,000 retailer addresses of discounters and full-line distributors in Germany as of early 2011.[27] Completeness of the address data is particularly important for the smaller retailers operating only very few supermarkets in Rhineland-Palatinate. Here, missing some supermarkets can severely bias the results. For this reason, I have cross-checked - and, in rare cases, complemented - the Wer-zu-Wem data with supermarkets found in the 2010 and 2011 editions of the German telephone book and the Yellow Pages.

[25] The shape-file was downloaded from `https://www.suche-postleitzahl.org`.

[26] I am indebted to Stefan A. Duphorn of Wer-zu-Wem GmbH for granting me access to the data.

[27] While the data is primarily used for marketing purposes, it has been used by researchers in the field of retailing geography, e.g. Burgdorf et al. (2014), Jürgens (2012) and Neumeier (2014). Burgdorf et al. (2014) make spot checks of the completeness of the data and find very few missing addresses for the larger retail chains.

The address dataset lists a total of 1,218 supermarkets in Rhineland-Palatinate for the 13 retailers considered in this paper. About 58 per cent of them are discounters, the remaining 42 per cent are full-line distributors. The number of supermarkets per retailer in Rhineland-Palatinate ranges from 3 to 228, see Table 2.3 below.

	Freq.	**Percent**	**Cum.**
Retailer 1	190	15.60	15.60
Retailer 2	228	18.72	34.32
Retailer 3	11	0.90	35.22
Retailer 4	13	1.07	36.29
Retailer 5	21	1.72	38.01
Retailer 6	175	14.37	52.38
Retailer 7	3	0.25	52.63
Retailer 8	166	13.63	66.26
Retailer 9	75	6.16	72.41
Retailer 10	100	8.21	80.62
Retailer 11	22	1.81	82.43
Retailer 12	156	12.81	95.24
Retailer 13	58	4.76	100.00

Table 2.3: Number of supermarkets by retailer

Households living close to the border of Rhineland-Palatinate may do their grocery shopping in neighbouring German states. These are Northrhine-Westfalia to the north, Hesse in the east, Baden-Württemberg in the south-east and Saarland in the south. Shopping across the border in neighbouring German states is potentially strong as the border areas are fairly densely populated and exhibit a high supermarket density. To account for this, retailers located in the 2-digit postcode areas along the border of Rhineland-Palatinate are added. The total number of supermarkets included in the analysis thereby increases to 3,601.

Moreover, Rhineland-Palatinate has a common border with Belgium and Luxembourg in the west and France in the south. In contrast to the densely populated border areas within Germany, the national border is sparsely populated except for the Trier and the Zweibrücken area. Thus, one can safely assume that ignoring fluid milk purchases in Belgium, Luxembourg and France will have little impact on the empirical results.

The supermarket addresses are geocoded using the Stata command opencagegeo (Zeigermann 2016). For 86.9 per cent of the 3,601 supermarkets, the exact location was found. For the remaining 13.1 per cent, the street was identified. All supermarket locations in Rhineland-Palatinate are displayed in Figure 2.2. Supermarkets in neighbouring 2-digit postcode areas bordering Rhineland-Palatinate are not shown.

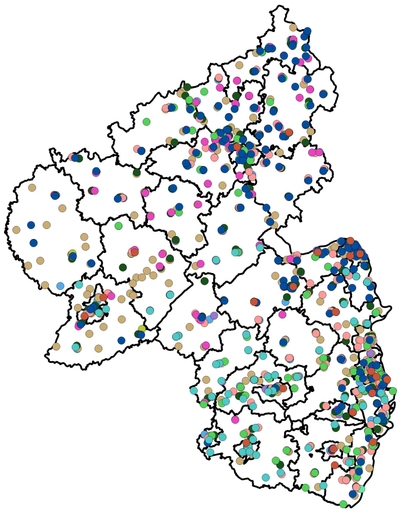
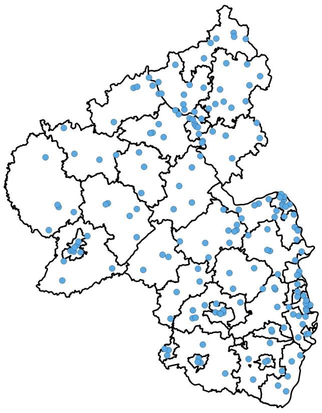


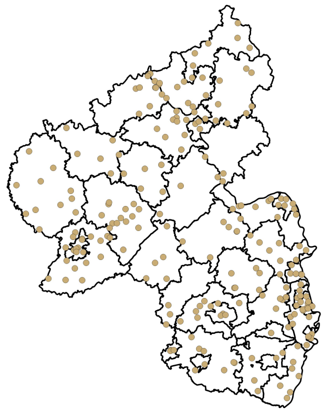
Figure 2.2: Supermarket locations in Rhineland-Palatinate[28]

The map shows a considerably higher density and variety of retailers in the south, particularly the south-east. In western and central Rhineland-Palatinate, markedly fewer supermarkets can be found. This directly corresponds with the population density and the distribution of panel members, see Figure 2.1. To get a clearer picture of local differences in the retailer structure, Figure 2.3 depicts supermarket locations by retailer in separate panels.

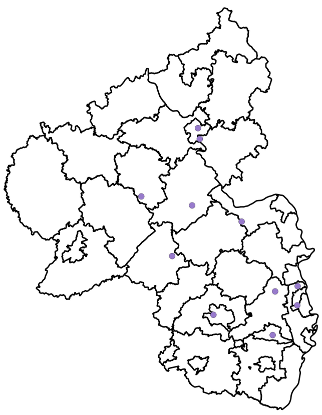
[28]The shape-file was downloaded from `https://www.suche-postleitzahl.org`.



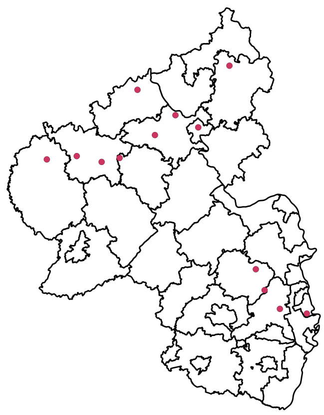
(a) Retailer 1



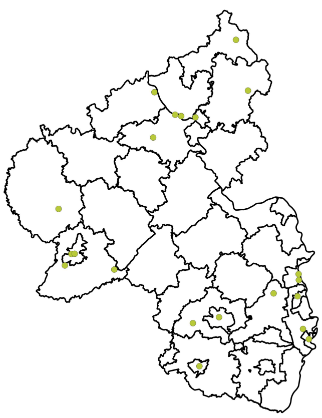
(b) Retailer 2



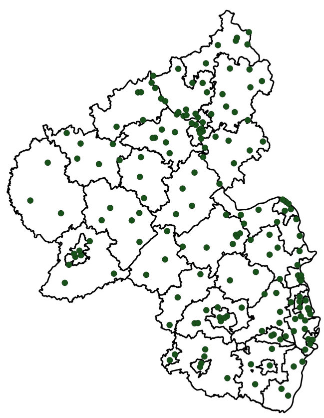
(c) Retailer 3



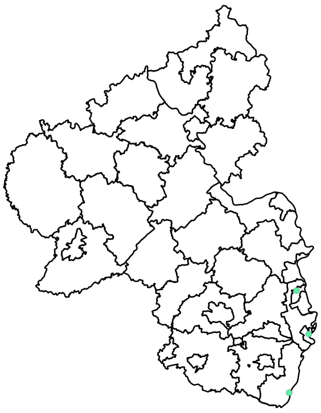
(d) Retailer 4



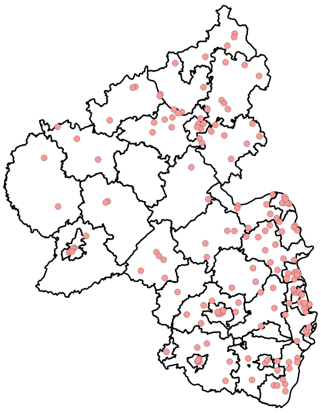
(e) Retailer 5



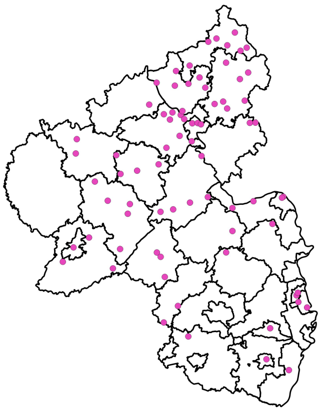
(f) Retailer 6



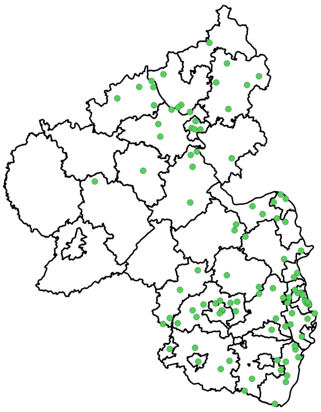
(g) Retailer 7



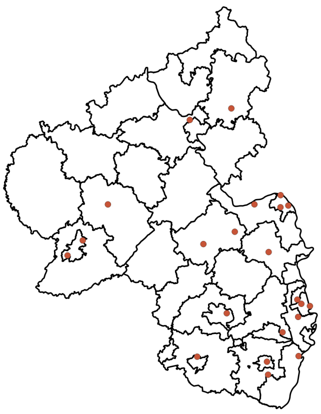
(h) Retailer 8



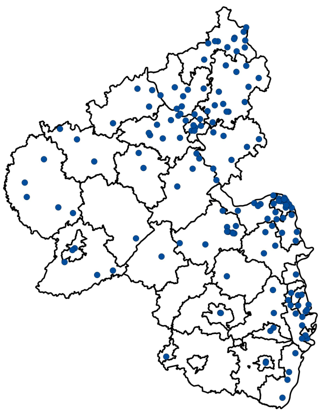
(i) Retailer 9



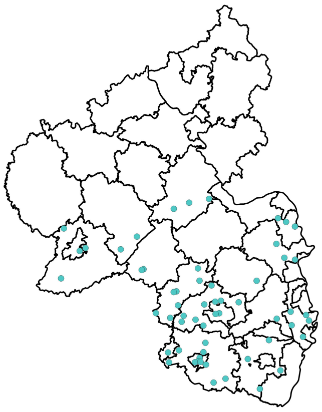
(j) Retailer 10



(k) Retailer 11



(l) Retailer 12



(m) Retailer 13

Figure 2.3: Supermarket locations in Rhineland-Palatinate by retailer[29]

The 13 retailers can be divided into three groups in terms of their regional availability. Group one comprises retailers 1,2 and 6, which are the largest in terms of the number of supermarkets. They are fairly evenly distributed across Rhineland-Palatinate (see panels a, b and f of Figure 2.3). In the second group are those retailers that still have a considerable number of supermarkets, but which are not as evenly spread across the state. These are retailers 8, 9, 10, 12 and 13 (see panels h, i, j, l and m of Figure 2.3). Finally, retailers 3, 4, 5, 7 and 11 have significantly fewer supermarkets and are present in only a number of selected locations (see panels c, d, e, g and k of Figure 2.3).

2.4 Construction of Local Choice Sets

This section describes the construction of local choice sets and compares them to the global choice set.

The maps in Figure 2.3 show that the retailer landscape is heterogeneous and characterised by very strong local differences. Modelling the choice process with a global choice set ignores this and implicitly assumes that consumers have access to all 13 retailers. This is particularly unrealistic for smaller retailers operating only a few of supermarkets in some locations. But even some relatively large retailers focus their business activities on some parts of Rhineland-Palatinate and are not available to consumers living in other areas.

[29]The shape-file was downloaded from `https://www.suche-postleitzahl.org`.

To capture the presence (and absence) of retailers and their products when modelling the choice process, I construct local choice sets containing only the products of retailers with a supermarket within a reasonable distance from the household's location. Research on shopping behaviour has shown that the travel distances for grocery shopping depend on numerous factors such as the type and size of the supermarket, the means of transport used/available, the population density and household characteristics such as age.[30] Taking all these aspects into account would complicate the construction of local choice sets tremendously and is beyond the scope of this paper. Recalling that the main purpose of this paper is to illustrate the overall effects of relaxing the global choice set assumption in favour of local choice sets, rather than computing very precise elasticities and choice parameters, it is in order to construct local choice sets according to a simple rule: Only the products of retailers with at least one supermarket within a 10-kilometre radius of the approximated household location are included in the household's local choice set.[31]

The radius of 10 kilometres is, however, not arbitrary and reflects the research results on actual distances travelled for grocery shopping. Moreover, it seems to be a good compromise between including too many retailers by drawing too large circles and excluding too many retailers by drawing too narrow circles.[32]

Constructing local choice sets according to the 10-kilometre-rule[33], I find that approximately 96 per cent of the observed product choices were actually made at retailers which are included in the choice set. I.e. 4 per cent of the purchases were made at retailers that where not captured by the 10-kilometre-rule. Because the data reveals that the household visited a supermarket of a certain retailer, all other products held by this retailer were added to the choice set for this particular choice occasion.

The global choice set contains all top 100 products held by the 13 retailers studied. The number of retailers and products with local choice sets is summarised in Table 2.4 below. The households have, on average, access to 8 rather than all 13 retailers. The average number of products to choose from drops from 100 to 68.

[30]See Krüger et al. (2013) for a very detailed research on factors driving supermarket choice.

[31]I have also constructed choice sets with a 5 km and 20 km radius. Not surprisingly, the results with 5 km deviate even more strongly from global choice sets, and those with a 20 km radius are closer to them compared to the 10 km rule used here.

[32]See the results on distances travelled for grocery shopping in Krüger et al. (2013).

[33]To identify all retailers within a 10-kilometre radius of a household, I use the Haversine formula for great-circle distances, which computes the distances between locations based on their latitudes and longitudes. The all-to-all matrix of distances between the 1406 panel households and the 3601 retailer locations in Rhineland-Palatinate and neighbouring 2-digit postcode areas was obtained using the haversine library for Mata (Zeigermann 2025).

High standard deviations on both further indicate a strong variation between the local choice sets of the panel households.

	Mean	**Std. Dev.**	**Min.**	**Max.**
# retailers	8.471	2.095	1	13
# products	67.702	17.77	4	100
N		21592		

Table 2.4: Average number of products and retailers in local choice sets

The heterogeneity in the local supplier structure is also reflected in the total number and the percentage of local choice sets a retailer is contained in, see Table 2.5.

	Freq.	**Percent**
Retailer 1	21219	98.27
Retailer 2	21069	97.58
Retailer 3	6935	32.12
Retailer 4	6856	31.75
Retailer 5	8932	41.37
Retailer 6	21386	99.04
Retailer 7	2793	12.94
Retailer 8	19686	91.17
Retailer 9	16088	74.51
Retailer 10	17250	79.89
Retailer 11	11760	54.46
Retailer 12	18609	86.18
Retailer 13	10315	47.77

Table 2.5: Presence of retailers in local choice sets

A retailer is, on average, present in approximately 65 per cent of the local choice sets. The large retailers 1, 2 and 6, which have by far the most supermarkets, are almost always present. The smallest, i.e. retailer 7, is contained in only 13 per cent of the local choice sets.

As discussed earlier, another important factor driving cross-price elasticities between products held by two competing retailers is the number of choice occasions where both are present at the same time. Table A2.1 in the appendix gives the percentage of common local choice sets for all retailer pairs. The average retailer pair is present in only 43.4 per cent of the choice sets. However, there is variation across retailers. The lowest number of choice sets share retailers 4 and 7. In only 6.5 per cent of the 21,592 choice occasions, decision-makers can choose between both. The largest number of shared choice occasions is between retailers 1 and 6 (97.7

per cent). These figures reflect the uneven distribution of supermarkets by retailers across Rhineland-Palatinate. For an illustrative example, consider retailers 3 and 4. They are very similar in terms of presence in choice sets (32.15 vs 31.75 per cent), but retailer 4 focusses its business on the north while retailer 3 has a similar number of supermarkets in the north and the south. As a consequence, retailer 13, which is not at all present in the north, shares around 17 per cent of the choice sets with retailer 3, but only 10 per cent with retailer 4.

2.5 Results

This section compares estimation results and price elasticities between both choice set specifications, and analyses the factors driving the differences.

2.5.1 Demand Estimation

The conditional logit demand model is estimated by maximum likelihood using Stata's clogit command. Product characteristics entering the utility function are the price[34], the fat content, two dummy variables for fresh and organically sourced products, and a set of retailer and brand fixed effects[35].

The demand model is estimated conditional on buying, i.e. I model the observed choices, but do not consider that a household might have opted not to buy any of the fluid milk products. I refrain from modelling the possibility of not buying, which is commonly referred to as an outside option in the literature, for the following reasons. Allowing for an outside option adds further complexity, requires strong assumptions on consumer demand, raises the requirements regarding the completeness of the data (which is potentially not satisfied by the GfK consumer panel) and - most importantly - does not affect the key results of this paper[36].

[34]Some unobserved product characteristics might be correlated with the price, e.g. product advertising, and hence give rise to biased estimates. A common approach to tackle this issue would be the widely used control function approach proposed by Petrin and Train (2010). In this paper, I do not correct for a potential bias in the price coefficient for the following reasons: the control function approach hinges on a number of critical assumptions, raises data requirements and adds further complexity. Most importantly, a potential bias would be similar for global and local choice sets and thus not systematically distort the results relevant in this paper.

[35]Retailer 1 acts as the baseline for the retailer fixed effects and private labels act as the baseline for the brand fixed effects.

[36]Not allowing for an outside option similarly affects the results obtained under both types of choice sets, so it does not substantially affect the relative differences between them.

The logit demand estimates with global (left) and local (right) choice sets are presented in Table 2.6.

	(1) global CS		(2) local CS	
chosen				
price	-0.0589***	(-49.21)	-0.0590***	(-49.20)
fat	0.0311***	(4.29)	0.0314***	(4.33)
fresh	-0.386***	(-26.17)	-0.386***	(-26.19)
organic	0.436***	(8.57)	0.441***	(8.66)
ret_2	-1.896***	(-64.36)	-1.923***	(-64.77)
ret_3	-1.862***	(-16.60)	-0.535***	(-4.74)
ret_4	-3.604***	(-47.80)	-2.377***	(-31.21)
ret_5	-2.197***	(-65.78)	-1.175***	(-33.84)
ret_6	-0.146***	(-7.21)	-0.171***	(-8.35)
ret_7	-3.029***	(-24.29)	-0.762***	(-6.04)
ret_8	-1.718***	(-65.40)	-1.614***	(-61.05)
ret_9	-2.382***	(-54.71)	-2.102***	(-47.95)
ret_10	-1.882***	(-54.04)	-1.628***	(-46.38)
ret_11	-2.010***	(-67.18)	-1.291***	(-41.79)
ret_12	-1.819***	(-58.85)	-1.684***	(-53.80)
ret_13	-3.338***	(-46.34)	-2.589***	(-35.73)
brand_1	0.620**	(3.08)	0.634**	(3.15)
brand_2	2.537***	(15.57)	2.552***	(15.64)
brand_3	1.174***	(16.17)	1.183***	(16.26)
brand_4	0.963***	(5.33)	0.982***	(5.43)
brand_5	0.634***	(5.78)	0.637***	(5.80)
brand_6	0.124	(1.07)	0.136	(1.18)
brand_7	-0.395	(-1.85)	-0.379	(-1.78)
brand_8	0.375***	(4.46)	0.399***	(4.72)
brand_9	3.253***	(17.76)	3.263***	(17.80)
brand_10	-1.884***	(-11.44)	-1.873***	(-11.37)
brand_11	0.0989	(1.07)	0.104	(1.13)
brand_12	-1.823***	(-8.19)	-1.811***	(-8.13)
N	2159200		1461814	

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2.6: Conditional logit demand estimates with global and local choice sets

In both specifications, almost all variables entering the model are highly statistically significant. The price coefficient is negative, as expected. Consumers value a higher fat content and organically sourced products, and prefer UHT products to fresh milk. All retailer dummies are negative indicating that retailer 1, the baseline, is the most popular. Moreover, there is a considerable degree of heterogeneity in the magnitude of the retailer fixed effects, showing strong differences in the popularity of retailers. The brand fixed effects, which are mainly positive, reveal that branded products are generally preferred over private labels, which serve as the baseline. Here again, we observe substantial variation between the different brands offered.

Comparing parameter estimates across specifications, we make two observations. First, the coefficients on price, fat, fresh, organic and the brand fixed effects are almost identical across both specifications.[37] Second, retailer fixed effects decrease in absolute terms and are severely affected by how choice sets are formed. This result makes sense intuitively because the differences in choice sets between both specifications are on retailer level and are therefore fully captured by the retailer fixed effects.

For closer inspection, retailer fixed effects for both specifications and the difference between them are given in Table 2.7.

	coeff global	coeff local	difference
Retailer 2	-1.896	-1.923	-0.027
Retailer 3	-1.862	-0,535	1.327
Retailer 4	-3.604	-2.377	1.2270
Retailer 5	-2.197	-1.175	1.022
Retailer 6	-0.146	-0.171	-0.025
Retailer 7	-3.029	-0.762	2.267
Retailer 8	-1.718	-1.614	0.104
Retailer 9	-2.382	-2.102	0.280
Retailer 10	-1.882	-1.628	0.254
Retailer 11	-2.010	-1.291	0.719
Retailer 12	-1.819	-1.684	0.135
Retailer 13	-3.338	-2.589	0.749

Table 2.7: Retailer fixed effects with global and local choice sets

With global choice sets, the estimated fixed effects indicate a generally higher popularity of larger retailers which are present in a larger fraction of the 21,592 choice sets, see left panel of Figure A2.2 in the appendix. I.e. the model explains the lower market shares of smaller retailers to a large degree by their lower popularity among the decision-makers. This relationship disappears with local choice sets. The right panel of Figure A2.3 in the appendix illustrates that the positive correlation between a retailer's fixed effect and its choice set presence no longer exists with local choice sets. I.e., consumers generally do not dislike smaller retailers, but

[37]Re-estimating both specifications with product prices artificially set to the yearly average yields identical parameter estimates for all variables but the retailer fixed effects confirming that the small differences observed are solely driven by slight variations in price over time.

limited availability prevents many consumers from buying their products. In Figure A2.4 in the appendix, the absolute differences in fixed effects between choice sets specifications are plotted against the percentage of choice sets the respective retailers is included in. The graph shows that the differences in the fixed effect between choice set specifications is strongest for the smallest retailers which are included in the fewest choice sets and zero for large retailers.

2.5.2 Predicted Probabilities and Market Shares

The parameter estimates presented above are used to compute predicted probabilities for all 100 alternatives on each of the 21,592 choice occasions, which are then aggregated into market shares. Summary statistics are given in Table 2.8. The average market share, i.e. the average predicted choice probability over the entire sample of 21,592 choice occasions by retailer, is identical across both specifications.[38]

The individual choice probabilities, however, differ strongly between the two scenarios. With global choice sets, the variance of predicted individual probabilities is very low.[39]

This changes dramatically with local choice sets. As we have seen before, local choice sets are very heterogeneous both in size and composition, which leads to strong variations in the choice probability of an alternative across choice sets. For illustration, consider retailer 1. The mean choice probability of its products is about 5.96 per cent and identical across specifications, but the standard deviation is larger by a factor of 40 if choice sets are local. With global choice sets, the average choice probabilities of retailer 1's products lie in a narrow range between 5.87 and 6.03. With local choice sets, we observe choice probabilities between zero (if retailer 1 and its products were not in the local choice set) and almost 17 per cent.

2.5.3 Elasticities

Using the predicted parameters from both specifications and the individual choice probabilities, I compute market share elasticities as described above. With 100 products, we will obtain 10,000 values in total for each specification: 100 own-price elasticities and 9,900 cross-price elasticities.

[38] Running a two-sided t-test on the market shares of the top 100 products confirms that there is no statistically significant difference between the predicted market shares of both specifications.

[39] The variation is caused by the slight variation in the monthly prices. With prices constant over time, predicted choice probabilities would be identical across all 21,592 choice occasions.

	Variable	**Mean**	**Std. Dev.**	**Min.**	**Max.**
Retailer 1:	share_global	5.957	0.045	5.874	6.03
	share_local	5.957	1.612	0	16.825
Retailer 2:	share_global	0.58	0.034	0.548	0.678
	share_local	0.58	0.281	0	4.586
Retailer 3:	share_global	0.586	0.016	0.559	0.613
	share_local	0.586	0.867	0	5.048
Retailer 4:	share_global	0.183	0.019	0.161	0.24
	share_local	0.183	0.278	0	1.357
Retailer 5:	share_global	0.589	0.064	0.544	0.779
	share_local	0.589	0.724	0	5.955
Retailer 6:	share_global	3.331	0.037	3.265	3.394
	share_local	3.331	1.052	0	16.667
Retailer 7:	share_globall	0.153	0.01	0.14	0.172
	share_local	0.153	0.404	0	1.945
Retailer 8:	share_global	0.912	0.019	0.884	0.954
	share_local	0.912	0.351	0	4.268
Retailer 9:	share_global	0.666	0.007	0.648	0.679
	share_local	0.666	0.463	0	7.483
Retailer 10:	share_global	1.096	0.009	1.075	1.106
	share_local	1.096	0.638	0	25
Retailer 11:	share_global	0.471	0.031	0.428	0.554
	share_local	0.471	0.449	0	3.464
Retailer 12:	share_global	0.495	0.006	0.483	0.506
	share_local	0.495	0.3	0	4.313
Retailer 13:	share_global	0.237	0.005	0.226	0.244
	share_local	0.237	0.273	0	4.261
	N		21592		

Table 2.8: Predicted market shares by retailer with global and local choice sets

To make the results more comprehensible and comparable between choice set specifications, I proceed in three steps. First, I present own-price elasticities. Second, I show and compare intra-retailer cross-price elasticities between products held by the same retailer. Third, I discuss inter-retailer cross-price elasticities between products sold by two competing retailers.

Own-price elasticities Summary statistics for own-price elasticities of all top 100 products with global and local choice sets are given in Table 2.9. Average own-price elasticities by retailer are provided in Table A2.2 in the appendix.

Variable	Mean	Std. Dev.	Min.	Max.
own_global	-4.071	1.252	-8.449	-2.626
own_local	-4.046	1.255	-8.428	-2.601
N		100		

Table 2.9: Average own-price elasticities with global and local choice sets

In both specifications, own-price elasticities average around -4, ranging between -2.6 and -8.4. Pairwise comparison shows that own-price elasticities are uniformly larger in absolute terms with global choice sets.[40] However, the difference is negligible and almost zero.

Cross-price elasticities More interesting and insightful is the comparison of cross-price elasticities. First, consider intra-retailer cross-price elasticities, i.e. for product pairs sold by the same retailer.

The average intra cross-price elasticity over the 864 product pairs rises strongly from 0.026 to 0.042, see Table 2.10. Average intra-retailer cross-price elasticities by retailer are given on the diagonal of the Tables A2.3 and A2.4 in the appendix for global and local choice sets, respectively.

[40] Assume for illustration that a product i has a market share of 1 per cent. Let there be $N = 100$ household making a single choice each. With global choice sets, there are identical choice probabilities for all consumers. The marginal effect is then $\frac{1}{N}\sum_{n=1}^{N} P_{nit}^{G}(1-P_{nit}^{G}) = \frac{1}{100}(100*0.01*0.99) = 0,0099$. With local assume now that only 50 households can choose product i. To obtain an overall market share of 1 per cent, the individual choice probabilities of those 50 consumers must be 2 per cent. The direct marginal effect is then $\frac{1}{N}\sum_{n=1}^{N} P_{nit}^{G}(1-P_{nit}^{G}) = \frac{1}{100}(50*0.02*0.98+50*0*0) = 0,0098$.

Variable	Mean	Std. Dev.	Min.	Max.
cross_intra_global	0.026	0.037	0.001	0.285
cross_intra_local	0.042	0.044	0.001	0.305
N		864		

Table 2.10: Average intra-retailer cross-price elasticities with global and local choice sets

Pairwise comparison shows that intra-retailer cross-price elasticities uniformly increase with local choice sets. This result is driven by two factors. The first is technical, and results from the logit model used. Recall the expression for market share elasticities $E_{ij} = \frac{\bar{p}_j}{s_i}\frac{1}{N}\frac{1}{T}\sum_{n=1}^{N}\sum_{t=1}^{T} -\alpha P_{nit}P_{njt}$. As we have seen before, the predicted market share, s_i, the average price, $\bar{p}_j$, and the estimated price coefficient α are identical across specifications. The difference must thus result from the multiplicative term of individual choice probabilities, i.e. $P_{nit}P_{njt}$. Note that although the average choice probability (i.e. the market share) is identical, this term is larger when the individual choice probabilities are high for some decision makers and zero for others (which is the case with local choice sets) compared to the situation, where all decision-makers have almost the same choice probability (which is the case with global choice sets).[41]

The second factor directly results from how local choice sets are constructed. Because products sold by the same retailer are always present in the same local choice sets, while products of other retailers are only present in some, the former necessarily become a relatively more important alternative following a price increase in one product compared to the choice situation with the global choice set.

The average percentage increase in intra cross-price elasticities is as high as 85.4 percent and varies strongly between just 7.1 per cent and 693.3 percent, see Table 2.11. The average percentage changes of intra-retailer cross-price elasticities by retailer are displayed on the diagonal of Table A2.5 in the appendix.

[41] For illustration, assume two products i and j held by the same retailer, which have a market share of 3 and 5 per cent, respectively. Let there be $N = 100$ households making one choice each. In our setting with global choice, choice makers have individual choice probabilities that are identical across all consumers. In this case, $\frac{1}{N}\sum_{n=1}^{N} P_{nit}^{G}P_{njt}^{G} = \frac{1}{100}(100 * 0.03 * 0.05) = 0.0015$. Now assume that with local choice sets, only 50 households have access to both products. For the overall market share to be 3 and 5 per cent, respectively, their individual choice probability must be 6 and 10 per cent. Then $\frac{1}{N}\sum_{n=1}^{N} P_{nit}^{L}P_{njt}^{L} = \frac{1}{100}(50 * 0.06 * 0.1 + 50 * 0 * 0) = 0.003$.

Variable	**Mean**	**Std. Dev.**	**Min.**	**Max.**
cross_intra_change	85.376	77.602	7.084	693.299
N		864		

Table 2.11: Percentage change in average intra-retailer cross-price elasticities between choice set specifications

To better understand the factors driving the difference in intra-retailer cross-price elasticities, I regress the percentage difference in elasticities on the share of local choice sets in which the respective retailer is contained. To allow for a non-linear relation, I also include a squared term. The regression results are presented in Table 2.12 below.

	%-change cross_intra	
ishare	-11.647***	(-58.80)
isharesq	.067***	(44.86)
_cons	527.391***	(90.39)
N	864	
R^2	0.948	

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2.12: Regression results for intra-retailer cross-price elasticity differences between choice set specifications

The results show that the percentage change in intra-retailer cross-price elasticities decreases in the share of choice sets a retailer is contained in. I.e. the change is largest for small retailers with few supermarkets. The positive coefficient on the squared term indicates convexity, i.e. the percentage change decreases at a decreasing rate. The very high R^2 of 0.95 indicates that almost the entire variation in the percentage change between specifications can be explained by the share of choice sets in which a retailer is present.

In the next step, I turn to the remaining 9,036 inter-retailer cross-price elasticities computed for product pairs held by competing retailers. Under both specifications, the average inter cross-price elasticity is almost identical at approximately 0.03, see Table 2.13. Average inter-retailer cross-price elasticities by retailer are given in the off-diagonal elements of the Tables A2.3 and A2.4 in the appendix.

Variable	Mean	Std. Dev.	Min.	Max.
cross_inter_global	0.034	0.050	0.001	0.286
cross_inter_local	0.032	0.048	0.001	0.295
N		9036		

Table 2.13: Average inter-retailer cross-price elasticities with global and local choice sets

While the average decrease in inter-retailer cross-price elasticities is only about 4.2 per cent, the percentage change between different retailer pairs ranges from -50 to 95 per cent. The average percentage change of inter-retailer cross-price elasticities by retailers is given in the off-diagonal elements of Table A2.5 in the appendix.

Variable	Mean	Std. Dev.	Min.	Max.
%-change cross_inter	-4.244	12.194	-49.203	94.991
N		9036		

Table 2.14: Percentage change in average inter-retailer cross-price elasticities between choice set specifications

In stark contrast to intra-retailer effects, inter-retailer cross-price elasticities between different retailers do not change uniformly and go in either direction. To understand this result, it is insightful to highlight two counteracting factors.

For the multiplicative in the expression for cross-price elasticities, $E_{ij} = \frac{\bar{p}_j}{s_i}\frac{1}{N}\frac{1}{T}\sum_{n=1}^{N}\sum_{t=1}^{T} -\alpha P_{nit}P_{njt}$, to be non-zero, both products need to be present in a decision maker's choice set. Contrary to the global choice sets scenario, this is not always the case with local choice sets. For decision makers who have access to only one of the two relevant retailers, or to neither, there is no substitution between products i and j. Therefore, allowing for local choice sets has a decreasing effect on inter-retailer cross-price elasticities. The fewer the number of supermarkets that retailers r_i and r_j have, the less often they should be included in the same choice set. In the most extreme case with no local choice set containing both retailers, inter-retailer cross-price elasticities drop to zero.

However, for those individuals whose choice sets contain both alternatives i and j, the multiplicative term will be higher for local choice sets compared to the global choice set scenario for the same reasons discussed in the context of intra-retailer cross-price elasticities. Therefore, allowing for local choice sets will result in stronger inter-retailer substitution for those consumers that have access to both retailers.

The overall effect will depend on the combined impact of both factors. To shed light on these opposing effects driving the differences in inter-retailer cross-price elasticities, I regress the percentage change both on the share of choice sets retailer i is contained in and the share of choice sets rival j is contained in. To account for the fact that substitution between two products sold by competing retailers is only possible in the context of local choice sets when both are contained in the same choice set, I also include the percentage of local choice sets both retailers considered have in common. As before, I add squared terms of all variables entering the model to allow for non-linearity. The estimation results are given in Table 2.15.

	(1)	
	%-change cross_inter	
ishare	-3.052***	(-66.13)
isharesq	0.014***	(49.18)
jshare	-3.052***	(-66.14)
jsharesq	0.014***	(49.22)
ijshare	3.928***	(70.37)
ijsharesq	-0.018***	(-53.12)
_cons	133.7468***	(72.37)
N	9036	
R^2	0.5344	

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2.15: Regression results for inter-retailer cross-price elasticity differences between choice set specifications

As expected, the difference in inter-retailer corss-price elasticities is, ceteris paribus, decreasing in the share of local choice sets an alternative is contained in. Hence, the increase in elasticities is strongest for the smallest retailers that are represented in the fewest local choice sets. The results further confirm that, ceteris paribus, the percentage change between choice sets specifications increases in the share of common choice sets of the two respective retailers. In other words, the percentage change between specifications is rising in the share of choice sets the two respective retailers share.

The R^2 indicates that a substantial amount of the variation between choice sets is explained by the choice set presence of the two respective retailers and the fraction of choice sets they share. However, there are additional factors that are driving the effects that are not captured. These are, for example, the size of the choice sets both retailers are represented in and their composition, i.e. the other alternatives to choose from.[42]

2.6 Conclusion

In this paper, I demonstrate how relaxing the common - and often implicit - assumption of a global choice set for all decision-makers in favour of local choice sets reflecting the retailer landscape impacts the demand estimation results and elasticities based thereon.

For this purpose, I combine the 2010 GfK consumer panel, which includes information on the approximate household location, and a comprehensive list of supermarket addresses in Germany. The product category under study is fluid milk. As a fast-moving consumer good with high purchasing frequencies and quantities, and clear product characteristics, it is ideal for modelling the choice process. To lower the computational burden and to capture that fluid milk products are often regional products, I restrict the analysis to the German federal state of Rhineland-Palatinate. Its key advantage over other German federal states it the fairly small size of its five-digit postcode areas which helps to obtain accurate approximations of household locations, as they are only disclosed at postcode level.

While the global choice sets contains the top 100 products for all decision makers, local choice sets are constructed by a simple rule: only products held by retailers with a supermarket within a radius of 10 kilometres of the approximated household location are contained in the local choice set. Applying this rule, the average number of retailers available drops from 13 to 8, and the number of products to choose from decreases by 32 per cent on average.

Comparing the logit demand estimates between both specifications shows that the differences are fully captured by the retailer fixed effects. With global choice sets, fixed effects for large retailers that operate many supermarkets all over Rhineland-Palatinate are the largest. This indicates that, ceteris paribus, consumers prefer these large retail chains to their smaller competitors. However, once the assumption of global presence of all retailers and their products is relaxed, this result changes fundamentally. Now, smaller retailers can be just as popular, if not more so, among consumers who have access to them. The lower market share of smaller retailers is thus to a large extent simply due to their lower availability.

[42]E.g. if the two retailers are both contained in large choice sets with many rival retailers, the multiplicative term $P_{nit}P_{njt}$ is smaller than if the two smaller choice sets with higher predicted choice probabilities for both relevant products.

These effects on demand estimates carry over to elasticities of demand, which are computed on the basis of predicted choice probabilities. Own-price elasticities uniformly fall, but only by a negligible amount. The impact on cross-price elasticities is, however, substantial.

Intra-retailer cross-price elasticities, i.e. between the products held by the same retailer, rise uniformly with local choice sets. The increase between the different product pairs ranges from just 7 per cent to almost 700 per cent in the example. This result is driven by two factors. First, two products held by the same retailer are present in the same choice sets. It is therefore always possible to substitute one for the other in the case of a price increase, while products of rival retailers are not necessarily available. Second, because there are generally fewer alternatives in local choice sets, the choice probability of each product naturally increases, resulting in larger cross-price effects. The increase in intra-retailer cross-price elasticities is largest for retailers which are contained in the lowest number of local choice sets, while the effect on large retailers with many supermarkets is small.

The second effect giving rise to higher elasticities with local choice sets also holds for inter-retailer cross-price elasticities between products held by competing retailers. However, inter-retailer cross-price elasticities with local choice sets also depend on the number of choice sets both alternatives are contained in. As smaller retailers are included in fewer choice sets, their products are less likely to be substitutes for each other. Hence, introducing local choice sets also has a decreasing effect on inter-retailer cross-price elasticities. Therefore, the overall effect of introducing local choice sets on cross-price elasticities between alternatives offered by different retailers is ambiguous and goes in either direction. In the empirical example, the change in inter-retailer cross-price elasticities following the introduction of local choice sets ranges from a decrease of almost 50 per cent to an increase of 95 per cent. The percentage change is c.p. largest for smaller retailers that are contained in fewer choice sets and increases c.p. in the number of choice sets common to both retailers.

These results demonstrate that, despite their efforts to develop increasingly sophisticated logit demand models, researchers risk obtaining biased demand estimates and elasticity when ignoring choice set heterogeneity in their empirical approach.

Acknowledgements

First and foremost, I would like to thank Anna Lu for long and intense discussions about discrete choice models with choice set heterogeneity. I am also indebted to Tomaso Duso for his repeated feedback and suggestions, which have helped to improve this paper. The paper has also gained from discussions with Florian Heiss, Michelle Sovinsky, Philip Schmidt-Dengler and the participants of EARIE in Milan.

References

Bajari, P. and Benkard, L. (2005). Demand estimation with heterogeneous consumers and unobserved product characteristics: A hedonic approach. *Journal of Political Economy*, 113(6):1239–1276.

Barroso Ludena, A. and Llobet, G. (2012). Advertising and consumer awareness of new, differentiated products. *Journal of Marketing Research*, 49:773–792.

Bonnet, C., Bouamra-Mechemache, Z., and Klein, G. J. (2022). National brands in hard discounters: market expansion and bargaining power effects. *European Review of Agricultural Economics*, 49(5):1113–1145.

Bonnet, C. and Dubois, P. (2010). Inference on vertical contracts between manufacturers and retailers allowing for nonlinear pricing and resale price maintenance. *The RAND Journal of Economics*, 41(1):139–164.

Bronnenberg, B. J. and Vanhonacker, W. R. (1996). Limited choice sets, local price response and implied measures of price competition. *Journal of Marketing Research*, 33(2):163–173.

Burgdorf, M., Krischausky, G., and Müller-Kleissler, R. (2014). Berechnung und Visualisierung von Indikatoren zur Nahversorgung. In Strobl, J., Blaschke, T., and Griesebner, G., editors, *Angewandte Geoinformatik 2014*. Wichmann Verlag.

Conlon, C. T. and Mortimer, J. H. (2013). Demand estimation under incomplete product availability. *American Economic Journal: Microeconomics*, 5(4):1–30.

Draganska, M. and Klapper, D. (2011). Choice set heterogeneity and the role of advertising: An analysis with micro and macro data. *Journal of Marketing Research*, 48(4):653–669.

Draganska, M., Mazzeo, M., and Seim, K. (2009). Beyond plain vanilla: Modeling joint product assortment and pricing decisions. *Quantitative Marketing and Econonmics*, 7:105–146.

Goeree, M. (2008). Limited information and advertising in the U.S. personal computer industry. *Econometrica*, 76(5):1017–1074.

Haucap, J., Heimeshoff, U., Klein, G. J., Rickert, D., and Wey, C. (2021). Vertical relations, pass-through, and market definition: Evidence from grocery retailing. *International Journal of Industrial Organization*, 74:102–693.

Hickman, W. and Mortimer, J. H. (2016). Demand Estimation with Availability Variation. SocArXiv qe69j, Center for Open Science.

Jürgens, U. (2012). Standortmuster von Lebensmitteldiscoutern - eine systematische Betrachtung am Beispiel Schleswig-Holsteins. *Berichte des Arbeitskreises Geographische Handelsforschung*, 31:35–41.

Krüger, T., Anders, S., Walther, M., Klein, K., and Segerer, M. (2013). Qualifizierte Nahversorgung im Lebensmitteleinzelhandel - Bericht im Auftrag von HDE und BVL. Technical report, HafenCity University Hamburg and Universität Regensburg.

Lu, A. (2017). Inference of consumer consideration sets. Dissertation, Deutsches Institut für Wirtschaftsforschung (DIW).

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In *Frontiers in Economics*. Academic Press.

Neumeier, S. (2014). Modellierung der Erreichbarkeit von Supermärkten und Discountern: Untersuchung zum regionalen Versorgungsgrad mit Dienstleistungen der Grundversorgung. Thünen Working Papers 16, Johann Heinrich von Thünen Institute, Federal Research Institute for Rural Areas, Forestry and Fisheries.

Nierop, E. V., Bronnenberg, B., Paap, R., Wedel, M., and Franses, P. H. (2010). Retrieving unobserved consideration sets from household panel data. *Journal of Marketing Research*, 47(1):63–74.

Pancras, J. (2010). A framework to determine the value of consumer consideration set information for firm pricing strategies. *Computational Economics*, 35:269–300.

Petrin, A. and Train, K. (2010). A control function approach to endogeneity in consumer choice models. *Journal of marketing research*, 47(1):3–13.

Reiss, P. C. and Wolak, F. A. (2007). Structural econometric modeling: Rationales and examples from industrial organization. In Heckman, J. J. and Leamer, E. E., editors, *Handbook of Econometrics*, volume 6, pages 4277–4415. Elsevier.

Shah, D., Kumar, V., and Zhao, Y. (2015). Diagnosing brand performance: Accounting for the dynamic impact of product availability with aggregate data. *Journal of Marketing Research*, 52(2):147–165.

Train, K. (2009). *Discrete choice methods with simulation.* Cambridge University Press.

Villas-Boas, S. (2007). Using retail data for upstream merger analysis. *Journal of Competition Law and Economics*, 3(4):689–715.

Zeigermann, L. (2016). OPENCAGEGEO: Stata module for geocoding. Statistical Software Components, Boston College Department of Economics.

Zeigermann, L. (2025). LHAVERSINE: Stata module to compute great-circle distances with haversine formula. Statistical Software Components, Boston College Department of Economics.

Appendix

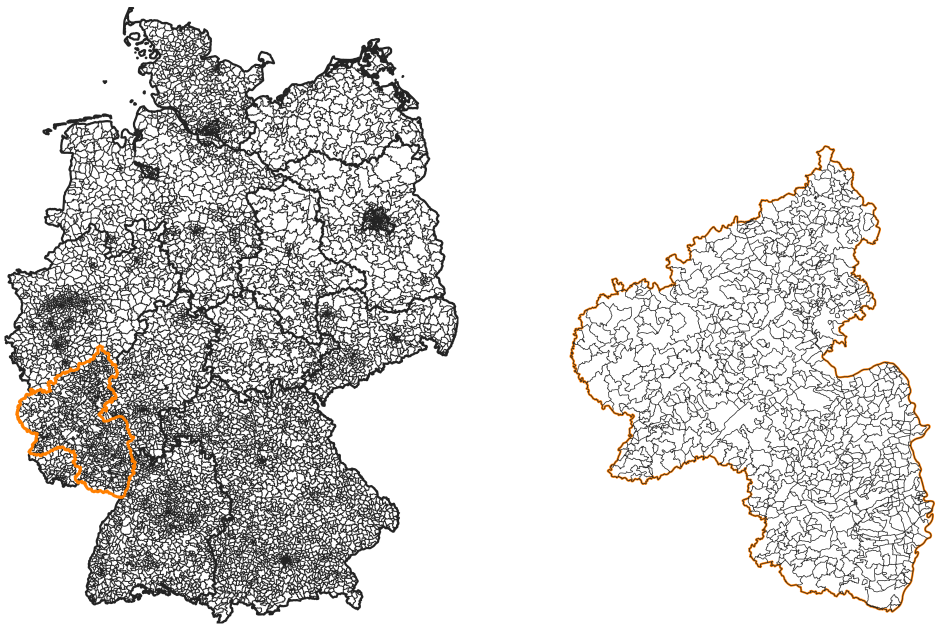


Figure A2.1: Five-digit postcode areas in Germany (left panel) and Rhineland-Palatinate (right panel)[43]

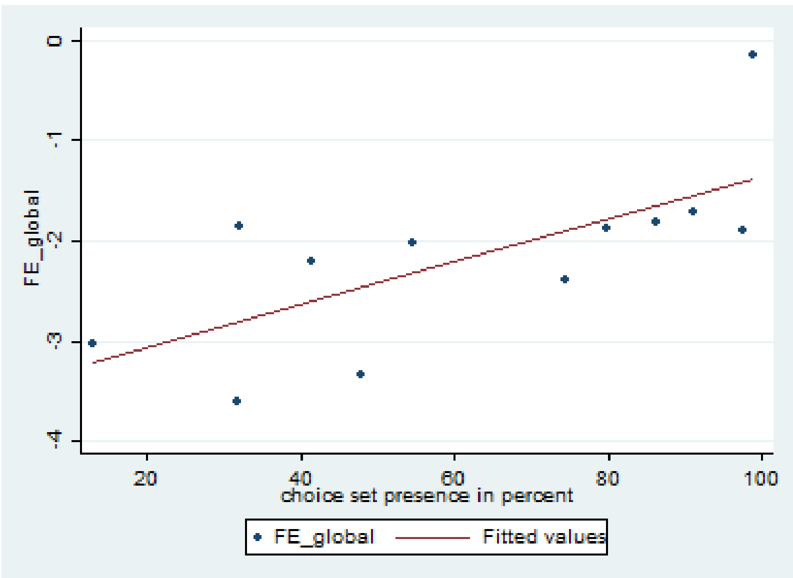


Figure A2.2: Retailer fixed effects with global choice sets and retailer presence

[43]The shape-file was downloaded from `https://www.suche-postleitzahl.org`.

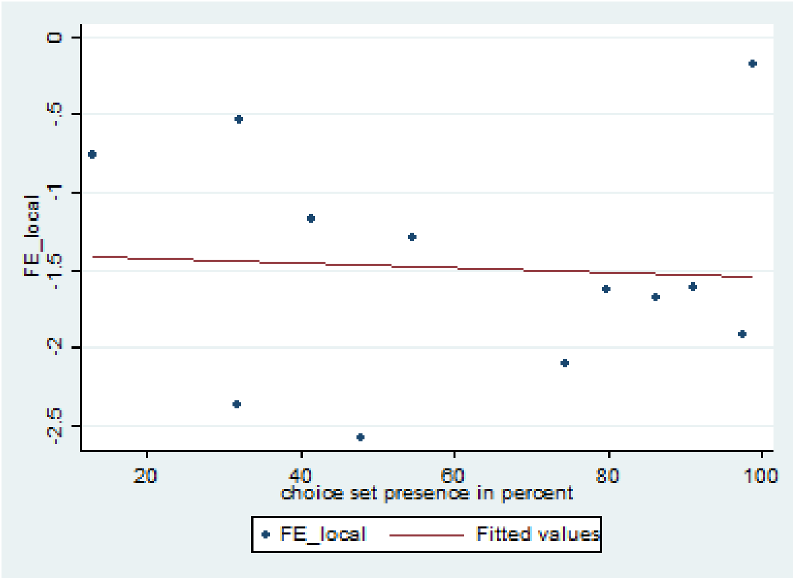


Figure A2.3: Retailer fixed effects with local choice sets and retailer presence

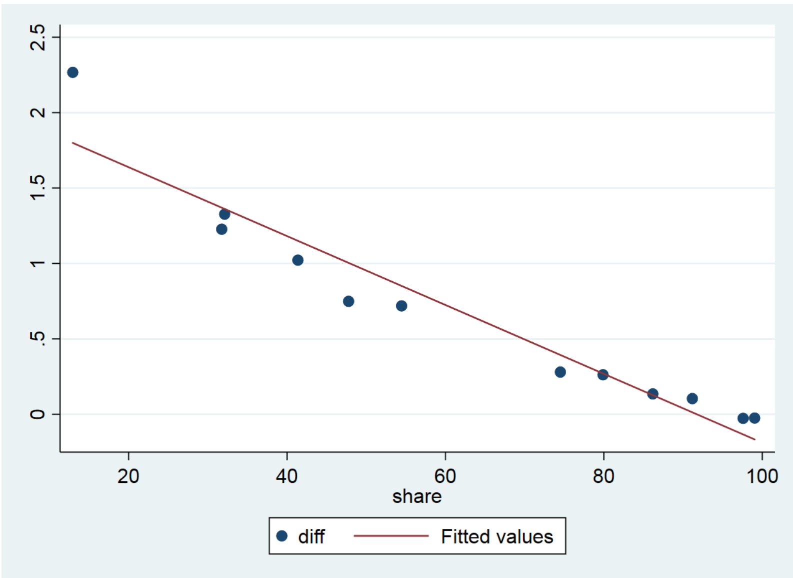


Figure A2.4: Differences in retailer fixed effects between choice set specifications and retailer presence

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13
R1	.983	.961	.32	.317	.413	.977	.129	.909	.738	.797	.543	.849	.477
R2	.961	.976	.314	.316	.413	.967	.129	.889	.728	.781	.544	.854	.463
R3	.32	.314	.321	.111	.165	.321	.072	.304	.242	.29	.226	.297	.176
R4	.317	.316	.111	.318	.204	.318	.065	.314	.276	.288	.195	.307	.1
R5	.413	.413	.165	.204	.414	.414	.12	.407	.292	.373	.27	.382	.236
R6	.977	.967	.321	.318	.414	.99	.129	.909	.743	.797	.544	.859	.476
R7	.129	.129	.072	.065	.12	.129	.129	.129	.113	.129	.108	.129	.101
R8	.909	.889	.304	.314	.407	.909	.129	.912	.688	.768	.537	.809	.45
R9	.738	.728	.242	.276	.292	.743	.113	.688	.745	.625	.437	.667	.3
R10	.797	.781	.29	.288	.373	.797	.129	.768	.625	.799	.477	.716	.409
R11	.543	.544	.226	.195	.27	.544	.108	.537	.437	.477	.545	.502	.31
R12	.849	.854	.297	.307	.382	.859	.129	.809	.667	.716	.502	.862	.381
R13	.477	.463	.176	.1	.236	.476	.101	.45	.3	.409	.31	.381	.478

Table A2.1: Share of common choice sets for all retailer pairs

	global CS	local CS
R1	-3.33	-3.32
R2	-4.69	-4.68
R3	-4.10	-4.04
R4	-3.97	-3.95
R5	-3.87	-3.83
R6	-3.95	-3.94
R7	-4.53	-4.48
R8	-3.90	-3.89
R9	-3.07	-3.05
R10	-3.05	-3.04
R11	-4.30	-4.28
R12	-4.68	-4.67
R13	-3.34	-3.32

Table A2.2: Average own-price elasticities with global and local choice sets

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13
R1	0.1872	0.0207	0.0200	0.0067	0.0196	0.1150	0.0058	0.0299	0.0204	0.0336	0.0174	0.0173	0.0075
R2	0.1871	0.0205	0.0200	0.0067	0.0194	0.1149	0.0058	0.0299	0.0203	0.0335	0.0174	0.0173	0.0075
R3	0.1872	0.0207	0.0200	0.0067	0.0196	0.1150	0.0058	0.0299	0.0204	0.0336	0.0174	0.0173	0.0075
R4	0.1872	0.0207	0.0200	0.0067	0.0195	0.1150	0.0058	0.0300	0.0204	0.0336	0.0173	0.0173	0.0075
R5	0.1872	0.0205	0.0200	0.0067	0.0196	0.1149	0.0058	0.0299	0.0204	0.0336	0.0173	0.0173	0.0075
R6	0.1872	0.0207	0.0200	0.0067	0.0196	0.1150	0.0058	0.0299	0.0204	0.0336	0.0174	0.0173	0.0075
R7	0.1872	0.0205	0.0200	0.0067	0.0196	0.1151	0.0058	0.0299	0.0204	0.0336	0.0174	0.0173	0.0075
R8	0.1872	0.0207	0.0200	0.0068	0.0196	0.1150	0.0058	0.0299	0.0204	0.0336	0.0173	0.0173	0.0075
R9	0.1872	0.0207	0.0200	0.0067	0.0196	0.1150	0.0058	0.0299	0.0204	0.0336	0.0174	0.0173	0.0075
R10	0.1872	0.0207	0.0200	0.0067	0.0196	0.1150	0.0058	0.0299	0.0204	0.0336	0.0174	0.0173	0.0075
R11	0.1871	0.0207	0.0200	0.0067	0.0195	0.1150	0.0058	0.0299	0.0204	0.0336	0.0174	0.0173	0.0075
R12	0.1872	0.0207	0.0200	0.0067	0.0196	0.1150	0.0058	0.0299	0.0204	0.0336	0.0174	0.0173	0.0075
R13	0.1872	0.0207	0.0200	0.0067	0.0196	0.1150	0.0058	0.0299	0.0204	0.0336	0.0174	0.0173	0.0075

Table A2.3: Average cross-price elasticities between retailer pairs with global choice sets

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13
R1	0.2005	0.0207	0.0165	0.0065	0.0173	0.1167	0.0048	0.0304	0.0204	0.0334	0.0156	0.0166	0.0078
R2	0.1876	0.0252	0.0157	0.0062	0.0168	0.1196	0.0046	0.0285	0.0211	0.0319	0.0157	0.0197	0.0076
R3	0.1547	0.0162	0.0636	0.0053	0.0185	0.0935	0.0078	0.0248	0.0168	0.0301	0.0178	0.0144	0.0067
R4	0.1794	0.0192	0.0156	0.0220	0.0264	0.1076	0.0065	0.0311	0.0213	0.0346	0.0164	0.0175	0.0039
R5	0.1656	0.0177	0.0189	0.0091	0.0483	0.0996	0.0112	0.0285	0.0167	0.0333	0.0173	0.0157	0.0078
R6	0.1900	0.0215	0.0163	0.0063	0.0170	0.1262	0.0047	0.0297	0.0210	0.0326	0.0153	0.0177	0.0074
R7	0.1540	0.0164	0.0269	0.0076	0.0379	0.0923	0.0460	0.0273	0.0191	0.0356	0.0218	0.0162	0.0097
R8	0.1901	0.0198	0.0166	0.0070	0.0187	0.1142	0.0053	0.0342	0.0201	0.0344	0.0170	0.0170	0.0075
R9	0.1873	0.0215	0.0166	0.0070	0.0161	0.1188	0.0054	0.0295	0.0302	0.0350	0.0166	0.0171	0.0066
R10	0.1861	0.0197	0.0180	0.0070	0.0195	0.1116	0.0062	0.0307	0.0212	0.0448	0.0171	0.0168	0.0077
R11	0.1686	0.0186	0.0205	0.0064	0.0196	0.1015	0.0073	0.0293	0.0195	0.0330	0.0330	0.0160	0.0079
R12	0.1797	0.0235	0.0167	0.0068	0.0178	0.1175	0.0054	0.0294	0.0202	0.0327	0.0161	0.0236	0.0060
R13	0.1933	0.0209	0.0178	0.0034	0.0203	0.1135	0.0075	0.0296	0.0177	0.0343	0.0181	0.0138	0.0175

Table A2.4: Average cross-price elasticities between retailer pairs with local choice sets

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13
R1	7.12	0.43	-17.33	-4.20	-11.67	1.44	-17.82	1.48	0.07	-0.61	-9.91	-4.00	3.25
R2	0.28	23.10	-21.75	-7.45	-13.63	4.02	-20.00	-4.78	3.86	-4.81	-10.17	13.87	0.94
R3	-17.35	-21.66	218.14	-21.87	-5.44	-18.72	34.04	-17.24	-17.26	-10.22	2.37	-16.61	-11.05
R4	-4.19	-7.30	-21.83	226.79	35.23	-6.49	12.27	3.36	4.43	3.16	-5.26	1.56	-48.85
R5	-11.55	-13.39	-5.28	35.39	146.85	-13.38	93.07	-4.66	-17.99	-0.69	0.56	-9.28	3.54
R6	1.47	4.20	-18.67	-6.47	-13.47	9.73	-19.95	-0.68	3.31	-2.97	-11.74	2.17	-1.35
R7	-17.72	-19.77	34.25	12.39	93.05	-19.87	691.38	-8.58	-6.22	6.04	25.19	-6.08	28.33
R8	1.51	-4.62	-17.20	3.37	-4.76	-0.68	-8.68	14.55	-1.33	2.53	-1.89	-1.64	-1.00
R9	0.07	4.01	-17.24	4.41	-18.10	3.27	-6.34	-1.36	48.13	4.33	-4.24	-0.84	-12.90
R10	-0.62	-4.68	-10.20	3.14	-0.83	-3.01	5.90	2.50	4.33	33.61	-1.64	-2.59	2.13
R11	-9.92	-10.05	2.38	-5.29	0.41	-11.78	25.02	-1.92	-4.24	-1.64	89.64	-7.45	4.21
R12	-3.97	14.06	-16.57	1.57	-9.37	2.16	-6.18	-1.64	-0.80	-2.56	-7.41	36.58	-20.33
R13	3.25	1.09	-11.02	-48.86	3.41	-1.38	28.17	-1.02	-12.89	2.14	4.23	-20.35	131.97

Table A2.5: Percentage differences in cross-price elasticities for all retailer pairs between global and local sets

Chapter 3

Retail Merger Screening with Local Choice Sets

3.1 Introduction

Grocery retailing is one of the most intensively researched industries in the empirical economic literature. Especially structural econometric models (of demand and supply) have become very popular over the past decades. The reason for their popularity is that structural econometric models are very powerful and versatile tools that can be used to i) compute unobserved parameters (such as demand elasticities, marginal costs, bargaining power), ii) conduct counterfactual analysis (such as merger or policy simulations) or iii) test the predictive performance of rival theories (e.g. different models of competition).[1]

The empirical literature on grocery retailing has addressed a wide range of research questions. Summarising the entire literature of structural econometric models in retailing is beyond the scope of this paper and the following contributions are just a selection of examples for the three different uses of structural models identified by Reiss and Wolak (2007).[2] For example, Draganska et al. (2010), Bonnet and Bouamra-Mechemache (2016), Richards et al. (2011) and Berto Villas-Boas (2007) use structural models to compute (unobserved) measures of bargaining power and profit sharing between retailers and manufacturers. Examples of counterfactual analysis are Bonnet et al. (2013) and Bonnet and Requillart (2013) who analyse the effects of potential upstream cost shocks and taxes on consumer prices. Examples of studies that use structural models to infer the pricing strategy used in vertical relations between manufacturers and retailers are Bonnet and Dubois (2010) and Haucap et al. (2021).

These selected examples show that the empirical literature on grocery retailing has made tremendous efforts to advance structural modelling by using increasingly sophisticated demand models, adding more and refined structure on the supply side and using richer datasets that have become available in recent years. However, the literature estimating some form of discrete choice demand model relies - often implicitly - on the strong global choice set assumption[3]: All consumers are assumed to choose from an identical choice set that contains all alternatives.

In the context of grocery retailing, the global choice set assumption is only satisfied when all consumers have access to all retailers and can choose between all

[1]This classification of the use cases of structural econometric models goes back to Reiss and Wolak (2007).

[2]Bonnet et al. (2020) review the relevant literature focusing on important issues such as understanding of vertical relationships in food markets, price formation, competition issues, environmental and nutritional policies.

[3]The literature also refers to the universal choice set.

products they offer. This is, however, very often not the case. Studies have shown that consumers visit supermarkets close to where they live and usually do not travel far for grocery shopping.[4][5] For this reason, the global choice set assumption is violated for people in rural areas with access to a very limited number of different retailers. But even in urban areas, not necessarily all retailers are available. This is the case because some retailers focus their business activities on some regions and are not present in others.

It is well established that ignoring variations in choice sets can give rise to biased demand estimates and measures derived from them, see e.g. Bronnenberg and Vanhonacker (1996); Bajari and Benkard (2005); Goeree (2008); Draganska and Klapper (2011). Hence, falsely assuming access to all retailers and their products in the demand estimation independent of where the consumers live, we risk obtaining biased demand estimation results that will carry over to the outcome of structural models based on them.

The goal of this paper is to demonstrate how relaxing the global choice set assumption in favour of more realistic local choice sets reflecting the retail landscape impacts the results of a structural model of supply and demand. In particular, I show how the outcome of potential retail mergers is affected by the definition of choice sets. Rather than performing a full-blown merger simulation, I make use of the popular and widely used upward pricing pressure (UPP) proposed by Farrell and Shapiro (2010a) as a screening tool for the initial price effects of a merger.

I use two datasets: the GfK (Gesellschaft für Konsumforschung) consumer panel data on milk purchases for the federal German state of Rhineland-Palatinate that also contains the approximate household locations and a comprehensive address dataset comprising all supermarket addresses provided by Wer-zu-Wem GmbH. Local choice sets are then constructed by a simple rule. A household's local choice set only contains the products offered by retailers that have a supermarket within a distance of 10 kilometres of the household's approximate place of residence. To keep the model tractable, I restrict my analysis to federal state of Rhineland-Palatinate.

I estimate two logit demand models, one with global and one with local choice sets. The first insight is that demand estimates are severely affected by choice set specification. With global choice sets, retailer fixed effects indicate that smaller retailers are less popular than their larger rivals. This result vanishes once local

[4]Krüger et al. (2013) study travel distances for grocery shopping in Germany and show that most consumers buy close to their homes.

[5]Other reasons for deviations from a global choice set are limited availability of products due to assortment decisions, temporary stock-out, or limited knowledge about available alternatives, see Hickman and Mortimer (2016).

choice sets are being used. Then smaller retailers are equally or even more popular, and their lower market share is simply due to their limited availability.

Using the demand estimation results, I compute UPP with both global and local choice sets for all potential mergers between the 13 retailers included in the data. Under both specifications, UPP is higher when the merging partner is large in terms of market shares. However, closer inspection of the results reveals that ignoring local heterogeneity in choice sets by assuming a global choice set severely affects the results. Percentage changes in UPP between retailer pairs resulting from introducing local choice sets range between -46 and 117 per cent. The relative differences in UPP between choice set specifications are largest for mergers involving smaller retailers, i.e. when the absolute UPP is fairly small, and the potential anticompetitive effect is limited. For potential mergers with large competitors that give rise to competitive concerns, UPP is similar under both choice set specifications.

The remainder of this paper is structured as follows. In Section 3.2, I derive upward pricing pressure (UPP) from pre- and post-merger first-order conditions. Section 3.3 presents the logit demand model, derives diversion ratios - the first building block of the UPP formula - and highlights the differences between global and local choice sets. In the following Section 3.4, I demonstrate how mark-ups - the second building block of UPP - can be inferred from the underlying supply model and the demand estimation results. Sections 3.5 and 3.6 summarise the data and show how local choice sets deviate from the global choice set. The demand estimation results and UPP for all potential mergers obtained under both choice set specifications are presented and compared in Section 3.7. Section 3.8 concludes.

3.2 UPP as a Merger Screening Tool

In this section, I outline the concept underlying upward-pricing pressure (UPP) and provide a formal definition.

There is a long tradition in the economic literature to simulate the (price) effects of mergers. However, a full-blown merger simulation is a challenging task that is non-trivial to implement and has fairly high data requirements. For antitrust authorities it is necessary to predict the outcomes of potential mergers or at least identify potentially harmful mergers without having the time or the data to perform a merger simulation.[6] For this reason, a number of simpler alternative approaches for evaluating potential merger effects have been developed. One of them is upward-

[6]Budzinski and Ruhmer (2009) review the merger simulation literature.

pricing pressure or UPP, formally introduced by Farrell and Shapiro (2010a).[7] UPP has gained considerable relevance in antitrust analysis[8] and is included in merger guidelines of several countries, including the US, the UK, France and Brazil, see Dutra and Sabarwal (2020).

The idea behind UPP is simple. It quantifies the incentive of the merging firm to raise prices after the merger by comparing two countervailing effects: its incentives to raise prices due to lost competition and its incentives to lower prices because of cost reductions through synergies.

Since the release of the first version of Farrell and Shapiro (2010a) in 2008, UPP has been modified and extended in a number of papers, see e.g. the gross upward pricing pressure index (GUPPI) proposed by Shapiro (2010) and Hausman et al. (2011). Others advocate modifying the UPP by relaxing some of its restrictive assumptions, thus bringing it closer to merger simulations and increasing its complexity and data requirements. Examples are Simons and Coate (2010) and Jaffe and Weyl (2013) who introduce pass-through terms, or Schmalensee (2009) who suggests accounting for cost-synergies in both merging parties' products. Initially constructed for Bertrand competition with single-product firms, UPP has been generalised to allow for multi-product firms (Jaffe and Weyl 2013), Cournot competition (Moresi 2009a), bidding competition (Moresi 2009b), and two-sided markets (Affeldt et al. 2013).

The seminal paper by Farrell and Shapiro (2010a) has spurred a vivid and controversial debate as to whether UPP itself is indicative of the magnitude of price change following a merger or merely of the sign of the expected price effect (Bailey et al. 2010; Farrell and Shapiro 2010b; Schmalensee 2009; Epstein and Rubinfeld 2010). The predictive accuracy of UPP has also been addressed empirically in a number of Monte Carlo studies. E.g. Miller et al. (2017) conclude that - depending on the demand system - UPP provides fairly precise post-merger prices. Cheung (2016) also finds that with structurally estimated price elasticities, UPP has high predictive accuracy, while simpler demand-side inputs give rise to mispredictions. Dutra and Sabarwal (2020) show that modifications of the initial UPP that incorporate cost efficiencies guided by theory can substantially improve the prediction of post-merger prices.

[7] Farrell and Shapiro (2010a) build on earlier works using first-order approaches to assess merger effects, see e.g. Hausman et al. (1994) and Werden (1996).

[8] See for example Baltzopoulos et al. (2015) who review five cases of the Swedish competition authority that rely on some form of UPP analysis.

To avoid the discussion of whether UPP can or cannot give precise predictions of the price increases following a merger, I take a conservative approach in this paper. I compute UPP without taking potential merger synergies into consideration and interpret the magnitude of UPP as an indicator of how much anti-competitive concern the merger should raise.[9]

3.2.1 Formal Derivation of UPP

Assume there are $r = 1, \ldots, R$ multi-product retailers. Each retailer r holds a set of J^r differentiated products denoted $\mathcal{J}^r$ which is a subset of all products $\mathcal{J} = \{1, \ldots, J\}$ in the market. Each product j is sold by a unique retailer, i.e. $\mathcal{J}^r \cap \mathcal{J}^s = \emptyset, \forall r \neq s$ Finally, let the market size be M.

Pre-merger maximisation problem The profit function of retailer r is given by

$$\pi_r = \sum_{j \in \mathcal{J}^r} (p_j - c_j) s_j(\mathbf{p}) M, \tag{3.1}$$

where $s_j(\mathbf{p})$ is the market share of product j for a vector $\mathbf{p}$ of all product prices.[10] Marginal costs of product j are denoted c_j and assumed to be constant. Retailers compete in prices à la Bertrand-Nash. Retailer r's first-order conditions are

$$s_j(\mathbf{p}) + \sum_{k \in \mathcal{J}^r} (p_k - c_k) \frac{\partial s_k(\mathbf{p})}{\partial p_j} = 0, \forall j \in \mathcal{J}^r. \tag{3.2}$$

For convenience, the expression can be rewritten in matrix notation:

$$\underbrace{\mathbf{s}_r(\mathbf{p})}_{J^r \times 1} + \underbrace{\left(\frac{\partial \mathbf{s}_r(\mathbf{p})^T}{\partial \mathbf{p}_r} \right)}_{J^r \times J^r} \underbrace{(\mathbf{p}_r - \mathbf{c}_r)}_{J^r \times 1} = \mathbf{0}, \tag{3.3}$$

where $\mathbf{s}_r$, $\mathbf{p}_r$ and $\mathbf{c}_r$ denote the market share, price and marginal cost vectors of r, respectively. For convenience, I rearrange the first-order conditions:

$$\mathbf{g_r}(\mathbf{p}) \equiv \underbrace{- \left(\frac{\partial \mathbf{s}_r(\mathbf{p})^T}{\partial \mathbf{p}_r} \right)^{-1}}_{J^r \times J^r} \underbrace{\mathbf{s}_r(\mathbf{p})}_{J^r \times 1} + \underbrace{(\mathbf{p}_r - \mathbf{c}_r)}_{J^r \times 1} = \mathbf{0} \tag{3.4}$$

[9] The magnitude of the UPP gives an indication of how large merger-induced efficiency gains need to be to offset the incentives for price increases.

[10] Vectors and matrices are in bold face.

Post-merger maximisation problem Now assume retailer r merges with its rival s offering the set of products $\mathcal{J}^s$. Let the merged retailer be denoted by m. Often, efficiency gains due to reductions in marginal cost resulting from the merger are also included. For simplicity, I do not consider possible merger-induced efficiencies in what follows.[11] The post-merger profit function of retailer m is then

$$\pi_m = \pi_r + \pi_s \tag{3.5}$$
$$= \sum_{j \in \mathcal{J}^r} (p_j - c_j) s_j(\mathbf{p}) M + \sum_{l \in \mathcal{J}^s} (p_l - c_l) s_l(\mathbf{p}) M. \tag{3.6}$$

The profit maximising post-merger prices of the products initially held by retailer r satisfy the following first-order conditions:

$$s_j(\mathbf{p}) + \sum_{k \in \mathcal{J}^r} (p_k - c_k) \frac{\partial s_k(\mathbf{p})}{\partial p_j} + \sum_{l \in \mathcal{J}^s} (p_l - c_l) \frac{\partial s_l(\mathbf{p})}{\partial p_j} = 0, \forall j \in \mathcal{J}^r \tag{3.7}$$

The third term reflects that now a pricing decision for a product j initially held by retailer r no longer only affects the profits earned from the products in $\mathcal{J}^r$, but also through additional profits gained from consumers that switch to products in $\mathcal{J}^s$ initially held by retailer s. The first-order conditions for the products in $\mathcal{J}^s$ are analogue. Rewriting the post-merger first-order conditions in matrix notation yields

$$\underbrace{\mathbf{s_r}(\mathbf{p})}_{J^r \times 1} + \underbrace{\left(\frac{\partial \mathbf{s_r}(\mathbf{p})^T}{\partial \mathbf{p}_r} \right)}_{J^r \times J^r} \underbrace{(\mathbf{p}_r - \mathbf{c}_r)}_{J^r \times J^r} + \underbrace{\left(\frac{\partial \mathbf{s}_s(p)^T}{\partial \mathbf{p}_r} \right)}_{J^r \times J^s} \underbrace{(\mathbf{p}_s - \mathbf{c}_s)}_{J^s \times 1} = 0, \tag{3.8}$$

which can be rearranged as follows:

$$\mathbf{h_r}(\mathbf{p}) \equiv \underbrace{- \left(\frac{\partial \mathbf{s_r}(\mathbf{p})^T}{\partial \mathbf{p}_r} \right)^{-1}}_{J^r \times J^r} \underbrace{\mathbf{s}_r(\mathbf{p})}_{J^r \times 1} - \underbrace{\left(\frac{\partial \mathbf{s_r}(\mathbf{p})^T}{\partial \mathbf{p}_r} \right)^{-1}}_{J^r \times J^r} \underbrace{\left(\frac{\partial \mathbf{s_s}(\mathbf{p})^T}{\partial \mathbf{p}_r} \right)}_{J^r \times J^s} \underbrace{(\mathbf{p_s} - \mathbf{c_s})}_{J^s \times 1} + \underbrace{\mathbf{p}_r \mathbf{c}_r}_{J^r \times 1} \tag{3.9}$$

Upward pricing pressure The system of first-order conditions the merging and the non-merging firms given by (3.4) and (3.9), respectively, could be used to simulate post-merger prices given sufficient information on marginal cost and a demand system (Miller et al. 2017). The complexity of full-blown merger simulation arises from the fact that not only the prices of the merging but also of the non-merging firms re-equilibrate.

[11] The UPP values can hence be interpreted as the required marginal cost reductions needed to offset the price increase following a merger.

Using the logit demand model and choice data described in section 3.5, it would be possible to simulate the post-merger equilibrium. However, as the purpose of this paper is to illustrate the effects of local choice sets on expected merger outcomes, it is useful to refrain from the complexities of merger simulation in favour of a more straightforward approach. This is particularly true because it would be infeasible to simulate all potential mergers between the 13 retailers considered in the empirical example provided later.

A simpler alternative proposed by Shapiro (2010) is to compute upward-pricing pressure, which is the difference between pre-merger (3.4) and post-merger first-order conditions (3.9).[12] UPP for retailer r is defined as follows

$$\mathbf{UPP_r} \equiv \mathbf{h_r} - \mathbf{g_r} = \underbrace{-\left(\frac{\partial \mathbf{s_r}(\mathbf{p})}{\partial \mathbf{p_r}}^T\right)^{-1}\left(\frac{\partial \mathbf{s_s}(\mathbf{p})}{\partial \mathbf{p_r}}^T\right)}_{\substack{J^r \times J^s \\ \text{Matrix of diversion from r to s}}} \underbrace{(\mathbf{p_s} - \mathbf{c_s})}_{\substack{J^s \times 1 \\ \text{Mark-ups of s}}} . \tag{3.10}$$

UPP_s is analogue.

The UPP formula for retailer r in (3.10) has two main building blocks: the diversion matrix between retailers r and s and the mark-ups of the merging partner s. The diversion matrix measures which proportion of sales lost due to a price increase by r is recaptured by increased sales of retailer s. Multiplying the diversion matrix by the mark-ups of retailer s yields the value of the sales shifted from r to s. I.e., the incentive for r to raise its prices increases in the value of the sales diverted to the merging partner s (Miller et al. 2013).

In what follows, I will first show how diversion matrices and mark-ups can be derived from a logit demand model and the supply model of retail competition. For both diversion ratios and mark-ups, I will highlight how relaxing the assumption of a global choice set containing all $j \in \mathcal{J}^r$ products in favour of local choice sets containing only a subset of all products affects the results.

3.3 Demand Model and Diversion Ratios

In this section, I will briefly present the conditional logit demand model (McFadden 1973) and derive diversion matrices from it. First, I will do so under the common global choice set assumption, i.e. all decision-makers can choose from the exact same

[12] Cheung (2016) establishes the theoretical relationship between UPP and merger simulations. She shows that UPP can be interpreted as a merger simulation solving for the change in the price of one merging firm holding all other endogenous variables fixed, i.e. ignoring re-equilibration.

set of products. Then, I will relax that assumption and allow for local differences in choice sets reflecting the local retail landscape.

The conditional logit model has long been superseded by more complex logit demand models that provide more realistic demand estimates and measures of substitution that are fundamental to structural modelling.[13] However, to demonstrate how relaxing the common global choice set assumption in favour of more realistic local choice sets impacts UPP, it is advantageous to use the traditional logit model. Its simplicity and mathematical properties help tremendously to identify and illustrate the effects resulting from different ways of constructing choice sets.

3.3.1 Conditional Logit Model

Assume there is a population of N decision-makers choosing one alternative in each of the T periods.[14] This implies that the size of the market M introduced earlier equals $N \cdot T$.

Let the indirect utility that consumer n obtains from buying product i in period t be

$$U_{nit} = \underbrace{\alpha p_{nit} + \mathbf{x}_{nit}\boldsymbol{\beta}}_{V_{nit}} + \varepsilon_i, \tag{3.11}$$

where x_{nit} is a vector of observed product characteristics and p_{nit} denotes the price of product i at time t. The coefficients α is the disutility from price, and $\boldsymbol{\beta}$ is a vector of taste parameters. Finally, ε_{nit} is a mean-zero, iid extreme value I distributed individual-specific random shock. For ease of notation, I will denote the deterministic portion of the n's utility from buying product i at time t by V_{nit}.

Global choice sets In a situation with global choice sets, all N decision makers can choose from all J alternatives in all T periods. Recalling from above that any product j can only be purchased from one retailer, the global choice set assumption amounts to assuming that all R retailers are available to all decision makers at any time.

[13]Today, random coefficient or mixed logit models are frequently used. See Train (2009) for a detailed description.

[14]Here, for simplicity, I assume the same number of repeated choices for all decisions. The number of repeated choices may vary between decision-makers and does so in the application.

The probability for consumer n of buying product i in period t, i.e. $U_{nit} > U_{njt} \; \forall \; j \neq i$, can then be written as

$$P^G_{nit} = \frac{\exp(V_{nit})}{\sum_{j=1} \exp(V_{njt})}, \tag{3.12}$$

where the superscript G indicates global choice sets. Note that the expression in (3.12) is always positive, i.e. we obtain positive choice probabilities for all J alternatives in each choice situation. The sum over all J choice probabilities for each individual at each choice occasion necessarily adds up to one.

The (predicted) market share of an alternative i is defined as the average choice probability over the sample of all N individual decision-makers over all T periods. Formally, the market share of product i is given by

$$s^G_i = \frac{1}{N}\frac{1}{T}\sum_{n=1}^{N}\sum_{t=1}^{T} P^G_{nit}. \tag{3.13}$$

Local choice sets Now, we drop the assumption of a global choice set in favour of local choice sets. No longer, all R retailers are necessarily available, and the local choice set of individual n, $\mathcal{J}^n$, only consists of the alternatives offered by retailers which are locally available. The expression for choice probabilities is then

$$P^L_{nit} = \begin{cases} \dfrac{\exp(V_{nit})}{\sum_{j \in \mathcal{J}^n} \exp(V_{njt})} & \forall i \in \mathcal{J}^n \\ 0 & \text{otherwise,} \end{cases} \tag{3.14}$$

where the superscript L indicates local choice sets. The choice probability of all alternatives not available to n is necessarily zero. The sum of the choice probabilities over all available alternatives in $\mathcal{J}^n$ is one.

The expression for the predicted market share of alternative i with local choice sets is given by the following expression:

$$s^L_i = \frac{1}{N}\frac{1}{T}\sum_{n=1}^{N}\sum_{t=1}^{T} P^L_{nit}. \tag{3.15}$$

3.3.2 Diversion Ratios

Having derived expressions for predicted market shares with global and local choice sets, we can compute diversion ratios for both specifications. Recall the expression for the diversion matrix from retailer r to retailer s:

$$\mathbf{DR}_{rs} = -\left(\frac{\partial \mathbf{s}_r(\mathbf{p})^T}{\partial \mathbf{p}_r}\right)^{-1}\left(\frac{\partial \mathbf{s}_s(\mathbf{p})^T}{\partial \mathbf{p}_r}\right) \tag{3.16}$$

The first right-hand side term is a $J^r x J^r$ square matrix of marginal effects of a price change in retailer r's prices on the market shares of its own products. The second right-hand side term is $J^r \times J^s$ matrix of cross marginal effects on the market share of s following a price increase by r.

Global choice sets Using the expression for the predicted market shares derived under the global choice set assumption given in 3.13, the diagonal elements of the square matrix in 3.16 are

$$\frac{\partial s_j^G}{\partial p_j} = \frac{1}{N}\frac{1}{T}\sum_{n=1}^{N}\sum_{t=1}^{T}\frac{\partial P_j^G}{\partial p_j}\forall j \in \mathcal{J}^r \tag{3.17}$$

$$= \frac{1}{N}\frac{1}{T}\sum_{n=1}^{N}\sum_{t=1}^{T}\alpha P_{njt}^G(1 - P_{njt}^G). \tag{3.18}$$

Similarly, the cross-effects between all products held by retailer r are then

$$\frac{\partial s_j^G}{\partial p_k} = \frac{1}{N}\frac{1}{T}\sum_{n=1}^{N}\sum_{t=1}^{T}\frac{\partial P_j^G}{\partial p_k}\forall j, k, j \neq k \in \mathcal{J}^r \tag{3.19}$$

$$= \frac{1}{N}\frac{1}{T}\sum_{n=1}^{N}\sum_{t=1}^{T} -\alpha P_{njt}^G P_{nkt}^G. \tag{3.20}$$

Because predicted choice probabilities P_{nit} are always positive, all the individual direct and indirect effects are non-zero.

The second right-hand side term of 3.16 is a $J^s x J^r$ matrix of indirect marginal effects. As before, they measure the effect of a price change in the column-alternative on the predicted market share of the row-alternative. However, this time, it is the cross effect from price changes in retailer r's initial products on the predicted market share of the product held by the other merging party, i.e. retailer s. Formally:

$$\frac{\partial s_j^G}{\partial p_l} = \frac{1}{N}\frac{1}{T}\sum_{n=1}^{N}\sum_{t=1}^{T}\frac{\partial P_j^G}{\partial p_l}\forall j \in \mathcal{J}^r, l \in \mathcal{J}^s \tag{3.21}$$

$$= \frac{1}{N}\frac{1}{T}\sum_{n=1}^{N}\sum_{t=1}^{T} -\alpha P_{njt}^G P_{nlt}^G \tag{3.22}$$

Individual cross-effects are non-zero for all decision-makers since both P_{njt}^G and P_{nlt}^G are always positive.

Local choice sets Plugging in the expression for predicted market shares with local choice sets given in 3.15, we obtain the following expressions for the diagonal and off-diagonal elements of the first right-hand side matrix in 3.16:

$$\frac{\partial s_j^L}{\partial p_j} = \frac{1}{N}\frac{1}{T}\sum_{n=1}^{N}\sum_{t=1}^{T}\alpha P_{njt}^L(1 - P_{njt}^L)\forall j \in \mathcal{J}^r \tag{3.23}$$

$$\frac{\partial s_j^L}{\partial p_k} = \frac{1}{N}\frac{1}{T}\sum_{n=1}^{N}\sum_{t=1}^{T} -\alpha P_{njt}^L P_{nkt}^L \forall j, k, j \neq k \in \mathcal{J}^r \tag{3.24}$$

First, consider the expression for direct marginal effects. The individual direct marginal effects are no longer necessarily non-zero for all choice situations but only for consumers with access to retailer r. The same holds for the expression for indirect marginal effects. Since the heterogeneity in choice sets is driven by the availability of retailers, P_{nit} and P_{nkt} are both positive for the subset of consumers who have access to retailer r or are both zero for the remaining consumers with no access.

The second right-hand side term of 3.16 is given by

$$\frac{\partial s_j^L}{\partial p_l} = \frac{1}{N}\frac{1}{T}\sum_{n=1}^{N}\sum_{t=1}^{T} -\alpha P_{njt}^L P_{nlt}^L \forall j \in \mathcal{J}^r, l \in \mathcal{J}^s. \tag{3.25}$$

For the cross-effect of a decision maker i to be non-zero, both retailers r and s must be in that decision maker's local choice set. Otherwise, the individual cross-effect is zero. Therefore, the overall cross marginal effect does not only depend on the predicted probabilities for the products considered but also on the number of individuals who have both products j and k in their local choice set and can thus substitute one for the other. In the extreme case that no consumer has access to both retailers r and s, the overall cross-marginal effect is zero.

3.4 Supply Model and Mark-ups

In this section, I establish how retailer mark-ups, the second building block of UPP given in 3.10, can be recovered from the supply model.

Recall the pre-merger first-order conditions describing the Bertrand-equilibrium written in matrix notation given in (3.3):

$$\mathbf{s}_r(\mathbf{p}) + \left(\frac{\partial \mathbf{s_r}(\mathbf{p})^T}{\partial \mathbf{p}_r}\right)(\mathbf{p}_r - \mathbf{c}_r) = \mathbf{0} \tag{3.26}$$

Rearranging yields the vector of r's mark-ups:

$$\mathbf{MU}_r = (\mathbf{p}_r - \mathbf{c}_r) = \left(\frac{\partial \mathbf{s_r}(\mathbf{p})^T}{\partial \mathbf{p}_r}\right)^{-1} \mathbf{s_r}(\mathbf{p}) \tag{3.27}$$

Note that the first right-hand term is identical to the first right-hand side term of the matrix of diversion ratios given in 3.16. Hence, the diagonal and off-diagonal elements are those presented in (3.18) and (3.20) for global choice sets and in (3.23) and (3.24) for their local counterparts. The second right-hand side terms, i.e. the market shares, are given in (3.13) and (3.15) for global and local choice sets, respectively.

3.5 Data

This section describes the two main data sources: the 2010 German GfK (Gesellschaft für Konsumforschung) household panel and the Wer-zu-Wem address dataset for supermarkets in Germany.

3.5.1 GfK Consumer Panel

Retail choice data The GfK consumer panel dataset comprises actual purchases made by the panel households who scan their receipts using home scanning devices. The data contains information on the different products purchased, the actual price paid (including discounts and promotions), the quantity, the retailer[15], the date of purchase and product characteristics. While GfK has information on all products purchased, only data on some selected product categories were available. For this

[15] Only the retailer's name is given. The exact supermarket where the purchase was made is not disclosed.

study, the product category fluid milk was chosen because it is a fast-moving consumer good with high purchasing frequencies and a low number of characteristics that describe the different products offered. These product characteristics are the fat content, the brand, and dummies for UHT, private labels[16] and organically sourced products. To keep the number of products tractable, I consider the 100 most often chosen products. Following the convention in the literature, a product is defined as the combination of a physical product and the retailer where it is sold.[17] That is, two products that are physically identical but sold by different retailers are treated as distinct products.

For reasons explained in what follows, I restrict the analysis to the state of Rhineland-Palatinate in the west of Germany. The total sales of the top 100 products considered in Rhineland-Palatinate is 21,592. This is about 97 per cent of all fluid milk purchases in the dataset; only around 3 per cent of all observations are dropped by restricting the analysis to the top 100 products. The main product characteristics are summarised in Table 3.1 below.

Variable	Mean	Std. Dev.	Min.	Max.
price	69.61	21.302	39	149
privlab (0=brand, 1=private label)	0.690	0.463	0	1
fat (per cent)	2.512	1.17	0.1	3.8
fresh (0=UHT, 1=fresh)	0.560	0.497	0	1
organic (0=conventional, 1=organic)	0.13	0.336	0	1
N		1200		

Table 3.1: Top 100 product characteristics

Of the 100 products considered, 69 are private labels. The average price is approximately 0.70 Euros, however, prices vary strongly ranging from 0.39 to almost 1.50 Euro.[18] The mean fat content is 2.5 cer pent. There are 56 fresh and 44 UHT products. Of the top 100 products, 13 products are organic.

The top 100 products are held by 13 different retailers, 5 being discount stores and the remaining 8 being full-line distributors.[19] Market shares by retailer in terms of sales and the number of products per retailer are displayed in Table 3.2 below.

[16]Private label products are produced for and sold under a retailer's brand.
[17]Technically, the retailer selling the product is treated as a product characteristic.
[18]The GfK consumer panel only reports the price paid of the alternative the household actually purchased. The prices for the non-chosen alternatives in the choice set are monthly averages over the prices paid by other consumers. In very few cases where no price for a product for a given month was observed, yearly averages were computed instead.
[19]Due to the confidentiality agreement with GfK, retailers cannot be named.

	Format	# products	Market share
Retailer 1	D	5	29.78
Retailer 2	FL	12	6.96
Retailer 3	FL	11	6.45
Retailer 4	FL	6	1.10
Retailer 5	FL	9	5.30
Retailer 6	D	6	19.98
Retailer 7	FL	2	.31
Retailer 8	D	10	9.12
Retailer 9	D	4	2.66
Retailer 10	FL	4	4.39
Retailer 11	D	15	7.06
Retailer 12	FL	12	5.94
Retailer 13	FL	4	.95
Total		100	100

Table 3.2: Retail format, number of products and market share by retailer

Retailer market shares range from just 0.3 to almost 30 per cent. The aggregate market share of the five discount supermarkets is approximately 65 per cent and illustrates their strong position in the German grocery retailing market. The number of products per retailer ranges between 2 and 15.

Household locations The GfK consumer panel also contains the five-digit postal code[20] of the household location.[21] Due to the lack of more detailed information, households are assumed to be located at the geographical centre of their respective postcode areas. For urban postcode areas, which tend to be fairly small, the approximation of the actual household location is very precise. However, even in rural postcode areas, we obtain fairly precise approximations of the true location of panel households living in Rhineland-Palatinate. First, Rhineland-Palatinate is the state with the smallest postcode areas in Germany except for the city-states (i.e. Berlin, Hamburg and Bremen). Furthermore, the largest conurbations, in which the majority of people reside, are typically located near the geographical centre of most postcode areas.

Of the 651 postcode areas in Rhineland-Palatinate, 439 (i.e. 67.4 per cent) are represented by panel households. To obtain the coordinates of the geographical centre of those postcode areas, i.e. the longitude and latitude, I use opencagegeo, an OpenStreetMap-based geocoding routine for Stata (Zeigermann 2016).

[20] Five-digit postcodes are the most granular postcode units in Germany.
[21] For reasons of confidentiality, the exact household location is not disclosed.

The locations of the 1,406 panel households are illustrated in Figure 3.1. The size of the symbol indicates the number of households in the respective postcode area.

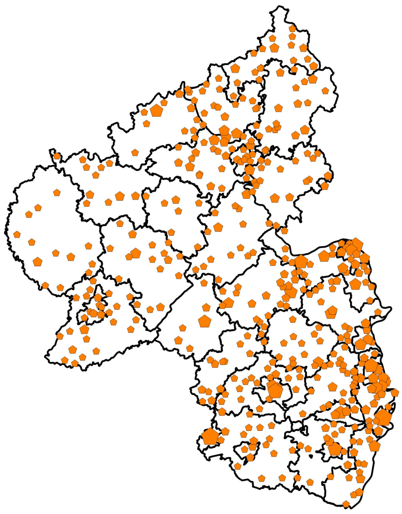


Figure 3.1: Locations of panel households in Rhineland-Palatinate[22]

3.5.2 Retailer Location Data

Address data for supermarket locations is obtained from Wer-zu-Wem GmbH.[23] The dataset comprises around 27,000 retailer addresses of discounters and full-line distributors in Germany as of early 2011.[24] Because completeness is particularly relevant for smaller retailers with few supermarkets, I have cross-checked the data and, in a few cases, added entries from the German telephone book and the Yellow Pages. The total number of supermarkets in Rhineland-Palatinate of the 13 retailers considered is 1,218, however, there is considerable variation between retailers, see Table 3.3 below.

To account for the fact that panel households living in Rhineland-Palatinate may do their grocery shopping in the neighbouring states of Northrhine-Westfalia, Hesse, Baden-Württemberg and Saarland, supermarkets in all 2-digit postcode areas along the border of Rhineland-Palatinate are added. The total number of retailers included in the analysis thereby increases to 3,601.

[22] The shapefile was downloaded from `https://www.suche-postleitzahl.org`.

[23] I am indebted to Stefan A. Duphorn of Wer-zu-Wem GmbH for granting me access to the data.

[24] This dataset has also been used by Burgdorf et al. (2014), Jürgens (2012) and Neumeier (2014).

	Freq.	Percent	Cum.
Retailer 1	190	15.60	15.60
Retailer 2	228	18.72	34.32
Retailer 3	11	0.90	35.22
Retailer 4	13	1.07	36.29
Retailer 5	21	1.72	38.01
Retailer 6	175	14.37	52.38
Retailer 7	3	0.25	52.63
Retailer 8	166	13.63	66.26
Retailer 9	75	6.16	72.41
Retailer 10	100	8.21	80.62
Retailer 11	22	1.81	82.43
Retailer 12	156	12.81	95.24
Retailer 13	58	4.76	100.00

Table 3.3: Number of supermarkets by retail chain

The supermarket locations in Rhineland-Palatinate are displayed in Figure 3.2.[25] The colour indicates the retailer a supermarket belongs to. Supermarkets in neighbouring 2-digit postcode areas bordering Rhineland-Palatinate are not shown.

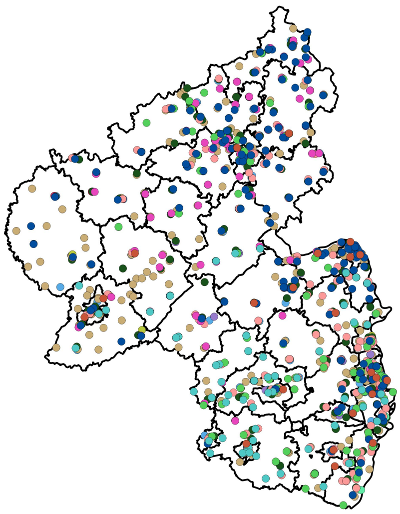


Figure 3.2: Supermarket locations in Rhineland-Palatinate[26]

Comparing Figures 3.2 and 3.1 shows that supermarket density and variety correlate with panel household density.

[25]The retailer addresses were geocoded using opencagegeo (Zeigermann 2016). The exact location of 86.9 per cent of the 3,601 supermarkets was identified. For the remaining 472 (13.1 per cent) supermarkets, the street was found.

[26]The shapefile was downloaded from https://www.suche-postleitzahl.org.

3.6 Local Choice Sets

This section describes the construction of local choice sets and compares them to the global choice set containing all 100 products.

Figure 3.2 illustrates that the retailer landscape is very heterogeneous, exhibiting very strong regional differences. Maintaining the standard assumption of a global choice set, which implicitly assumes that consumers have access to all 13 retailers, ignores that not all retailers have supermarkets within a reasonable distance of all households. To capture the presence (and absence) of retailers and their products when modelling the choice process, I construct choice sets according to a simple 10-kilometre rule: Only the products of retailers with at least one supermarket within a 10-kilometre radius of the approximate household location are included in the choice set.[27]

To identify all retailers within a 10-kilometre radius of a household, I use the Haversine formula for great-circle distances that computes the distances between locations based on their latitudes and longitudes. The all-to-all matrix of distances between the 1,406 panel households and the 3,601 retailer locations in Rhineland-Palatinate and neighbouring 2-digit postcode areas was obtained using the lhaversine library for Mata (Zeigermann 2025).

Constructing local choice sets according to the 10-kilometre rule, I find that approximately 96 per cent of the observed product choices were actually made at retailers that are included in the local choice set. I.e., only 4 per cent of the purchases where made at a retailer not contained in the local choice set. Because the data reveals that the household visited a supermarket of a particular retailer, all other products held by that retailer were added to the local choice set.

The number of retailers and products contained in local choice sets is summarised in Table 3.4 below. The households have, on average, access to 8.5 rather than all 13 retailers. The average number of products to choose from drops from 100 to 68. High standard deviations on both further indicate a strong variation between local choice sets.

The heterogeneity in the local retailing landscape is also reflected in the total number and the percentage of choice occasions a retailer is present, see Table 3.5.

[27]In addition, choice sets with a radius of 5 and 20 kilometres have been constructed. Comparing the results shows that the general effects remain the same, although they stand out most when a small radius is chosen and the differences between the two choice set scenarios are strongest.

	Mean	**Std. Dev.**	**Min.**	**Max.**
# retailers	8.471	2.095	1	13
# products	67.702	17.77	4	100
N		21592		

Table 3.4: Average number of products and retailers in local choice sets

	Freq.	**Percent**
Retailer 1	21219	98.27
Retailer 2	21069	97.58
Retailer 3	6935	32.12
Retailer 4	6856	31.75
Retailer 5	8932	41.37
Retailer 6	21386	99.04
Retailer 7	2793	12.94
Retailer 8	19686	91.17
Retailer 9	16088	74.51
Retailer 10	17250	79.89
Retailer 11	11760	54.46
Retailer 12	18609	86.18
Retailer 13	10315	47.77

Table 3.5: Presence of retailers in local choice sets

The average retailer is present in approximately 65 per cent of the choice occasions. Retailer presence varies strongly and ranges between 13 and 99 per cent. Overall, the average retailer pair is present in only 43.4 per cent of the local choice sets. This value shows substantial variation across retailers, see Table A3.1 in the appendix. The lowest number of choice occasions share retailers 4 and 7. In only 6.5 per cent of the 21,592 choice occasions, decision-makers can choose between both. The largest number of common choice occasions is between retailers 1 and 6 (97.7 per cent).

3.7 Results

In this section, I present the demand estimation results and UPP values for all potential mergers between the 13 retailers under both choice set specifications, and compare the results.

3.7.1 Demand Estimation

The variables entering the logit demand model are the price, the fat content, a dummy indicating if the product is fresh or UHT, and a dummy for organically sourced products. Summary statistics are given in Table 3.1. I also include brand and retailer dummies to account for the variation in popularity of different product brands and retailers. The baseline for the product fixed effects are private labels, and retailer 1, a large discounter, serves as the baseline retailer.

The demand model is estimated conditional on buying, i.e. I model the observed choices, however, do not consider that a household might have opted not to buy any of the fluid milk products at some choice occasions. I refrain from modelling the possibility of not buying, which is commonly referred to as an outside option in the literature, for the following reasons. First, allowing for an outside option adds further complexity, requires strong assumptions about consumers, and raises the requirements regarding the completeness of the data (which is not satisfied by the GfK consumer panel data available for this paper). Second, for fluid milk, which is a low-price fast moving consumer good and staple food, it is not very likely that consumers refrain from buying because of the price. Hence, the exclusion of an outside option seems justified. Third, the general results comparing UPPs under global and local choice sets should not be affected by excluding an outside option.

An additional concern is that estimates might be biased as there can be correlations between observed product characteristics and unobserved product characteristics captured by the error term. For example, the price might be correlated with advertising, which is not observed. This issue is commonly tackled through various forms of instrumental variables (IV), e.g. the control function approach proposed by Petrin and Train (2010). It is, however, often difficult to find valid instrumental variables, and using unsuitable IV can also lead to severely biased estimates. Because the goal of this paper is to analyse relative changes in UPP values under two different choice set specifications rather than computing the absolute price effects resulting from potential mergers, it is in order to estimate the model without using instrumental variables to correct for potentially biased parameter estimates.

The logit demand model results for both specifications are displayed in Table 3.6 below.

	(1)		(2)	
	global CS		local CS	
chosen				
price	-0.0589***	(-49.21)	-0.0590***	(-49.20)
fat	0.0311***	(4.29)	0.0314***	(4.33)
fresh	-0.386***	(-26.17)	-0.386***	(-26.19)
organic	0.436***	(8.57)	0.441***	(8.66)
brand_1	0.620**	(3.08)	0.634**	(3.15)
brand_2	2.537***	(15.57)	2.552***	(15.64)
brand_3	1.174***	(16.17)	1.183***	(16.26)
brand_4	0.963***	(5.33)	0.982***	(5.43)
brand_5	0.634***	(5.78)	0.637***	(5.80)
brand_6	0.124	(1.07)	0.136	(1.18)
brand_7	-0.395	(-1.85)	-0.379	(-1.78)
brand_8	0.375***	(4.46)	0.399***	(4.72)
brand_9	3.253***	(17.76)	3.263***	(17.80)
brand_10	-1.884***	(-11.44)	-1.873***	(-11.37)
brand_11	0.0989	(1.07)	0.104	(1.13)
brand_12	-1.823***	(-8.19)	-1.811***	(-8.13)
ret_2	-1.896***	(-64.36)	-1.923***	(-64.77)
ret_3	-1.862***	(-16.60)	-0.535***	(-4.74)
ret_4	-3.604***	(-47.80)	-2.377***	(-31.21)
ret_5	-2.197***	(-65.78)	-1.175***	(-33.84)
ret_6	-0.146***	(-7.21)	-0.171***	(-8.35)
ret_7	-3.029***	(-24.29)	-0.762***	(-6.04)
ret_8	-1.718***	(-65.40)	-1.614***	(-61.05)
ret_9	-2.382***	(-54.71)	-2.102***	(-47.95)
ret_10	-1.882***	(-54.04)	-1.628***	(-46.38)
ret_11	-2.010***	(-67.18)	-1.291***	(-41.79)
ret_12	-1.819***	(-58.85)	-1.684***	(-53.80)
ret_13	-3.338***	(-46.34)	-2.589***	(-35.73)
N	2159200		1461814	

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3.6: Logit demand estimates with global and local choice sets

In both specifications, almost all variables entering the model are highly statistically significant. The price coefficient is negative, as expected. Consumers value a higher fat content and organically sourced products, and prefer UHT products to fresh milk. All retailer dummies are negative, which indicates that retailer 1, the baseline, is the most popular. Moreover, there is a considerable degree of heterogeneity in the magnitude of the retailer fixed effects, showing strong differences in the popularity of retailers. The brand fixed effects, which are mainly positive, reveal that branded products are generally preferred over private labels which serve as the baseline. Here again, we observe substantial variation between the different brands offered.

Comparing parameter estimates across specifications, we make two observations. First, the coefficients on price, fat, fresh and organic, as well as the brand fixed effects, are almost identical across both specifications.[28] Second, retailer fixed effects decrease in absolute terms and are severely affected by how choice sets are formed. This result makes sense intuitively since the differences in choice sets between both specifications are on the retailer level and thus should be fully captured by the retailer fixed effects.

Closer inspection of the retailer fixed effects reveals that the difference between both choice sets specifications is most pronounced for small retailers that are only contained in a smaller fraction of all local choice sets. With global choice sets, the lower market shares of products of smaller retailers are partly attributed to the lower popularity of those retailers compared to the large discounter, which is the baseline. However, with local choice sets, the increase in the fixed effects shows that the lower market share is not due to lower retailer popularity, but the fact that the products of smaller retailers are available to a considerably smaller number of consumers.

3.7.2 Upward-Pricing Pressure

Using the demand estimates obtained, I compute UPP for all potential mergers. With 13 retailers, this amounts to 78 mergers and consequently 156 UPP values. In what follows, I first describe the outcomes under both choice set specifications and then illustrate the relative differences between the two.

First, consider global choice sets. The average upward-pricing pressure by retailer over all potential mergers with one of the 12 competitors is given in Table 3.7 below. Average UPP values for each retailer pair are provided in Table A3.2 in the appendix.

[28]Re-estimating both specifications with product prices artificially set to the yearly average yields identical parameter estimated for all variables but the retailer fixed effects. This result confirms that the small differences observed are solely driven by slight variations in price over time.

	Mean	**Std. dev.**	**min**	**max**
Retailer 1	1.581	1.583	0.074	6.029
Retailer 2	1.724	2.236	0.056	7.728
Retailer 3	1.723	2.222	0.056	7.686
Retailer 4	1.713	2.060	0.053	7.272
Retailer 5	1.722	2.191	0.055	7.599
Retailer 6	1.695	2.385	0.065	8.986
Retailer 7	1.710	2.033	0.163	7.212
Retailer 8	1.725	2.291	0.057	7.913
Retailer 9	1.716	2.109	0.053	7.387
Retailer 10	1.720	2.162	0.054	7.520
Retailer 11	1.724	2.239	0.056	7.736
Retailer 12	1.722	2.207	0.055	7.643
Retailer 13	1.712	2.054	0.053	7.260

Table 3.7: Average UPP by retailer with global choice sets

UPP values are between close to 0 and almost 9, indicating that some potential mergers are likely to create very little upward pressure on prices while others may potentially lead to substantial price rises. To gain a better understanding of where the strongest price increases are to be expected, it is insightful to relate the UPP values to the market share of both merging parties, see Figure 3.3.

The graph illustrates the predicted upward effect on prices of retailer r's products following a merger with retailer s. The area of the circles is proportional to the UPP values and hence indicates the magnitude of the expected upward pressure on the price. The graph shows that the effect on the price of retailer r does not depend on its own market size but critically hinges on the market share of the merging partner. This result makes sense intuitively. A retailer's incentives to raise its price increase in the fraction of lost sales that are recaptured by the merging party. Hence, when the merging party is big, it will recapture a larger portion of the lost sales, and therefore the scope for profitably raising the prices is largest.

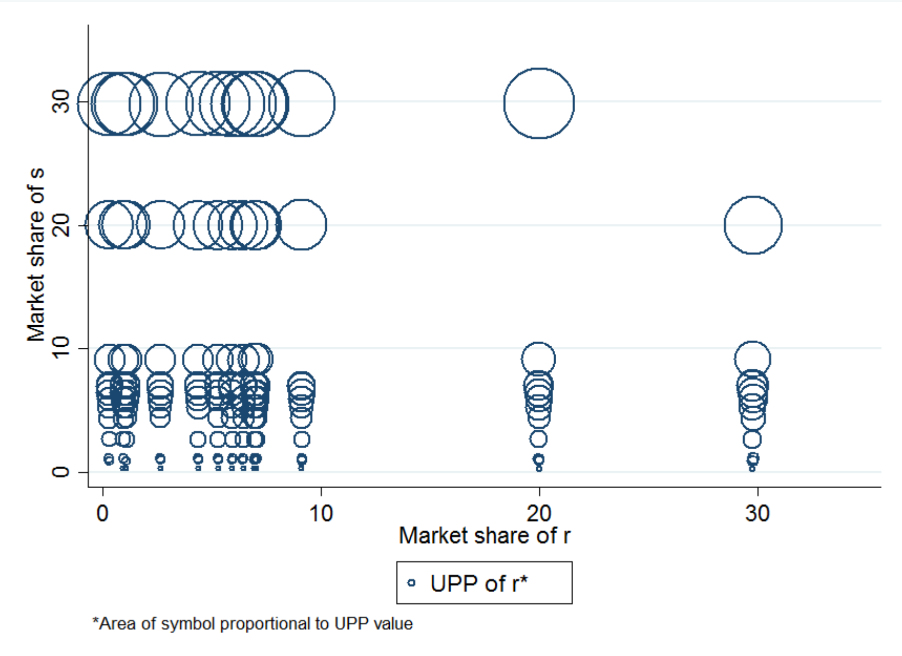


Figure 3.3: Relation between UPP and market shares with global choice sets

Turning to UPP values obtained with local choice sets, the general results look surprisingly similar at first glance. The average UPP values by retailer over the 12 potential mergers with rivals are given in Table 3.8. The average UPP values for each retailer pair are provided in Table A3.3 in the appendix.

	Mean	**Std. dev.**	**min**	**max**
Retailer 1	1.652	1.705	0.064	6.487
Retailer 2	1.799	2.385	0.046	8.155
Retailer 3	1.773	2.214	0.090	7.734
Retailer 4	1.776	2.071	0.062	7.395
Retailer 5	1.779	2.156	0.119	7.590
Retailer 6	1.774	2.574	0.055	9.668
Retailer 7	1.756	1.706	0.217	6.270
Retailer 8	1.802	2.432	0.054	8.430
Retailer 9	1.789	2.232	0.052	7.745
Retailer 10	1.791	2.239	0.060	7.849
Retailer 11	1.783	2.210	0.077	7.738
Retailer 12	1.791	2.274	0.054	7.769
Retailer 13	1.789	2.215	0.070	7.849

Table 3.8: Average UPP by retailer with local choice sets

As with global choice sets, UPP values increase in the market share of the merging partner and do not depend on the retailer's own market size with local choice sets, see Figure 3.4.

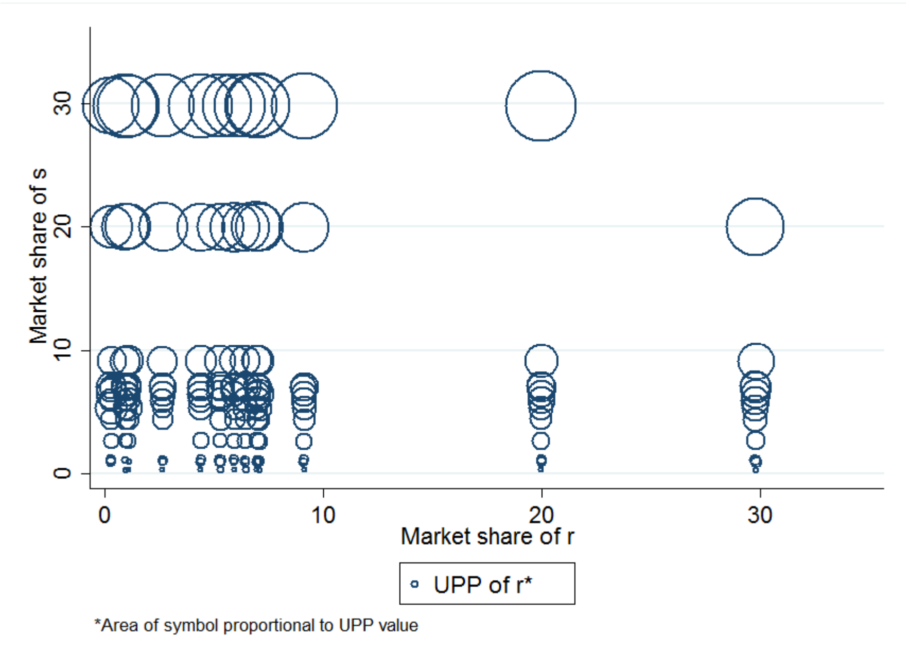


Figure 3.4: Relation between UPP and market shares with local choice sets

Despite the similarities between both specifications, closer inspection reveals significant differences. Average percentage differences in UPP values by retailer are given in Table 3.9. Table A3.4 in the appendix shows the differences between global and local choice sets for all retailer pairs. In some cases, the figures reveal substantial differences between both specifications. The percentage change in UPP from replacing global by local choice sets ranges from a decrease of 47 per cent to an increase of 117 per cent.

	Mean	**Std. dev.**	**min**	**max**
Retailer 1	2.322	5.652	-13.048	8.119
Retailer 2	0.909	8.871	-16.508	18.871
Retailer 3	9.445	19.698	-6.164	61.770
Retailer 4	4.262	21.964	-46.754	51.952
Retailer 5	18.753	34.723	-9.065	116.281
Retailer 6	1.232	6.882	-15.821	8.822
Retailer 7	18.507	39.102	-16.567	115.928
Retailer 8	2.989	3.954	-4.966	7.913
Retailer 9	1.860	6.489	-10.468	8.821
Retailer 10	5.661	3.663	-1.433	10.343
Retailer 11	10.038	12.661	-2.612	37.545
Retailer 12	2.281	8.066	-17.227	18.637
Retailer 13	1.341	19.646	-46.749	33.304

Table 3.9: Percentage change in UPP between choice set specifications by retailer

To get a clearer picture of which UPP values are most severely affected by the construction of choice sets, the percentage changes in UPP are plotted against the market share of both merging partners in Figure 3.5.

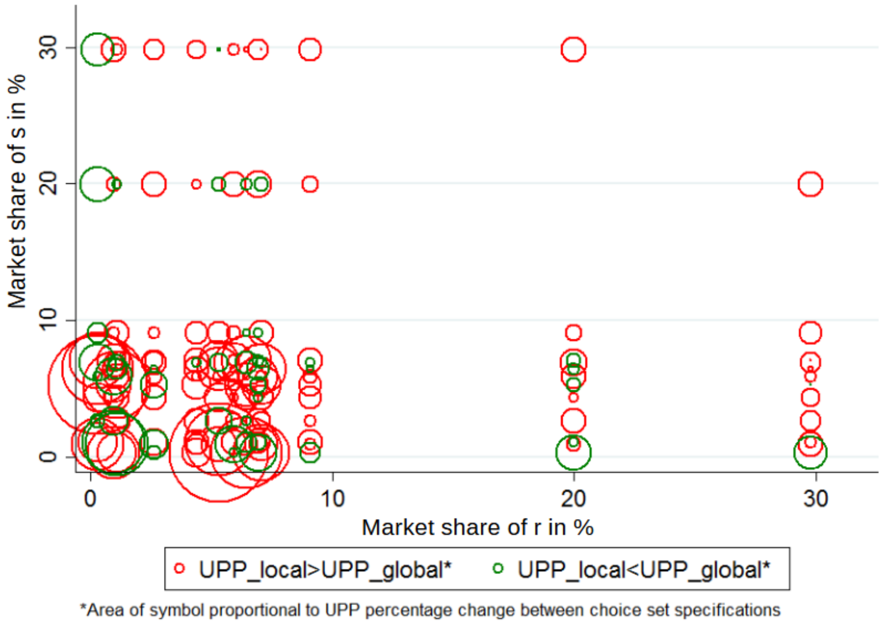


Figure 3.5: Differences in UPP between choice set specifications

The area of the circle indicates the percentage change in absolute values, i.e. circles for a 50 per cent increase and a decrease by 50 per cent are identical in size. The direction of change is marked by the respective colour. All cases where the UPP value increased when the global choice sets were replaced by local choice sets are plotted in red. The others cases, where the UPP value is higher with global choice sets, are displayed in green.

The graph shows that the difference between both specifications is strongest for mergers between smaller retailers with low market shares. Here, we observe the most pronounced differences in both directions. Increases in UPP resulting from local choice sets (red circles) are both larger in magnitude and more frequent than lower UPP values with local choice sets (green circles). For mergers involving at least one large retailer, UPP values with local choice sets usually increase (red circles), but the percentage increase is fairly small.

In a final step, I combine the insights from above, i.e. that the UPP values for a retailer - and hence the anticompetitive effects - increase in the market share of the merging partner and that differences in UPP values between both choice set specifications are largest when both merging parties are small. In Figure 3.6, percentage differences in UPP values between both choice set specifications are plotted against absolute UPP values computed with global choice sets.

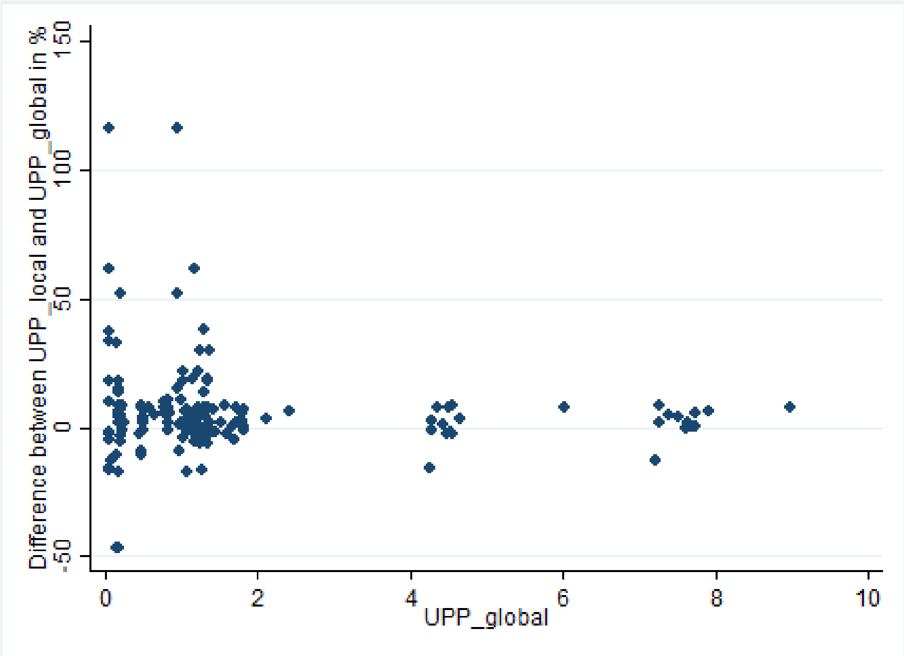


Figure 3.6: Relation between UPP with global choice sets and differences between UPP between choice set specifications

The graph clearly illustrates that differences in UPP between choice set specifications are large where the magnitude of UPP is low, and hence do not raise competitive concerns. The higher the UPP values and, therefore, the risk of rising prices after the merger, the less pronounced are the effects of the choice set specification on UPP values. In other words, using global choice sets can lead to significantly biased UPP values. However, critical cases flag up in UPP merger screening, both with global and local choice sets.

3.8 Conclusion

In this chapter, I show how relaxing the common assumption of a global choice set in favour of local choice sets reflecting the actual retail landscape affects potential price effects of retail mergers. For illustration, I use upward-pricing pressure - or UPP - which is a widely used merger screening tool.

I use data on milk purchases from the 2010 GfK consumer panel for the German federal state of Rhineland-Palatinate and a comprehensive list of supermarket addresses provided by Wer-zu-Wem GmbH. Local choice sets are generated by a simple rule: Only products held by supermarkets within 10 kilometres of the approximate household location are included in the local choice set.

I estimate logit demand with both choice set specifications. Using the demand estimates, I compute 156 UPP values under both specifications for all potential 78 mergers between the 13 retailers in the dataset and compare the results. Both with local and global choice sets, the expected price rise is increasing in the market share of the merging party. Not surprisingly, a retailer's prices rise more strongly when it merges with a large competitor because the lost sales following a potential price increase are captured to a larger degree by additional sales of products by the newly acquired former rival. A pairwise comparison of UPP values for all potential mergers obtained with global and local choice sets reveals pronounced deviations in either direction - the change in UPP values resulting from the introduction of local choice sets is between -46 and 117 per cent.

Closer inspection of these differences between specifications shows that they are strongest for mergers between retailers with smaller market shares. Mergers involving at least one larger retailer are less severely affected.

Combining the two key results, that mergers with larger competitors are potentially more harmful to competition and that UPP values are hardly affected by choice set specifications when at least one large retailer is involved, is great news for practitioners and researchers. Mergers that are expected to be severely anticompetitive are detected, even if the local retail landscape is being ignored. Large relative differences between UPP values only occur when the absolute UPP value is low, and therefore do not raise anti-competitive concerns.

Acknowledgements

I would like to thank Tomaso Duso for fruitful discussions about merger screening under different choice set specifications and his suggestions to illustrate the differences in UPP values between global and local choice sets.

References

Affeldt, P., Filistrucchi, L., and Klein, T. J. (2013). Upward pricing pressure in two-sided markets. *The Economic Journal*, 123(572).

Bailey, E. M., Leonard, G. K., Olley, G. S., and Wu, L. (2010). Merger screens: Market-share based approaches and 'upward pricing pressure'. *The Antitrust Source*.

Bajari, P. and Benkard, L. (2005). Demand estimation with heterogeneous consumers and unobserved product characteristics: A hedonic approach. *Journal of Political Economy*, 113(6):1239–1276.

Baltzopoulos, A., Kim, J., and Mandorff, M. (2015). UPP Analysis in Five Recent Merger Cases. Konkurrensverket Working Paper Series in Law and Economics 2015:3, Konkurrensverket (Swedish Competition Authority).

Berto Villas-Boas, S. (2007). Vertical Relationships between Manufacturers and Retailers: Inference with Limited Data. *The Review of Economic Studies*, 74(2):625–652.

Bonnet, C. and Bouamra-Mechemache, Z. (2016). Organic label, bargaining power, and profit-sharing in the French fluid milk market. *American Journal of Agricultural Economics*, 98(1):113–133.

Bonnet, C. and Dubois, P. (2010). Inference on vertical contracts between manufacturers and retailers allowing for nonlinear pricing and resale price maintenance. *The RAND Journal of Economics*, 41(1):139–164.

Bonnet, C., Dubois, P., Villas Boas, S. B., and Klapper, D. (2013). Empirical Evidence on the Role of Nonlinear Wholesale Pricing and Vertical Restraints on Cost Pass-Through. *The Review of Economics and Statistics*, 95(2):500–515.

Bonnet, C. and Requillart, V. (2013). Impact of cost shocks on consumer prices in vertically-related markets: The case of the French soft drink market. *American Journal of Agricultural Economics*, 95(5):1088–1108.

Bonnet, C., Trachtman, C., Dop, M. V., and Villas-Boas, S. B. (2020). Food markets? structural empirical analysis: a review of methods and topics. *European Review of Agricultural Economics*, 47(2):819–847.

Bronnenberg, B. J. and Vanhonacker, W. R. (1996). Limited choice sets, local price response and implied measures of price competition. *Journal of Marketing Research*, 33(2):163–173.

Budzinski, O. and Ruhmer, I. (2009). Merger simulation in competition policy: A survey. *Journal of Competition Law and Economics*, 6(2):277–319.

Burgdorf, M., Krischausky, G., and Müller-Kleissler, R. (2014). Berechnung und Visualisierung von Indikatoren zur Nahversorgung. In Strobl, J., Blaschke, T., and Griesebner, G., editors, *Angewandte Geoinformatik 2014*. Wichmann Verlag.

Cheung, L. (2016). An empirical comparison between the upward pricing pressure and merger simulation in differentiated product markets. *Journal of Competition Law and Economics*, 12(4):701–734.

Draganska, M. and Klapper, D. (2011). Choice set heterogeneity and the role of advertising: An analysis with micro and macro data. *Journal of Marketing Research*, 48(4):653–669.

Draganska, M., Klapper, D., and Villas-Boas, S. B. (2010). A larger slice or a larger pie? An empirical investigation of bargaining power in the distribution channel. *Marketing Science*, 29(1):57–74.

Dutra, J. and Sabarwal, T. (2020). Antitrust analysis with upward pricing pressure and cost efficiencies. *PLOS ONE*, 15(1):1–31.

Epstein, R. J. and Rubinfeld, D. L. (2010). Understanding UPP. *The B.E. Journal of Theoretical Economics*, 10(1).

Farrell, J. and Shapiro, C. (2010a). Antitrust evaluation of horizontal mergers: An economic alternative to market definition. *The B.E. Journal of Theoretical Economics*, 10(1):1935–1704.

Farrell, J. and Shapiro, C. (2010b). Upward pricing pressure in horizontal merger analysis: Reply to Epstein and Rubinfeld. *The B.E. Journal of Theoretical Economics*, 10(1):0000102202193517041704.

Goeree, M. (2008). Limited information and advertising in the U.S. personal computer industry. *Econometrica*, 76(5):1017–1074.

Haucap, J., Heimeshoff, U., Klein, G. J., Rickert, D., and Wey, C. (2021). Vertical relations, pass-through, and market definition: Evidence from grocery retailing. *International Journal of Industrial Organization*, 74:102693.

Hausman, J., Leonard, G., and Zona, J. D. (1994). Competitive analysis with differenciated products. *Annales d'Economie et de Statistique*, (34):159–180.

Hausman, J., Moresi, S., and Rainey, M. (2011). Unilateral effects of mergers with general linear demand. *Economic Letters*, 111:119–121.

Hickman, W. and Mortimer, J. H. (2016). Demand Estimation with Availability Variation. Socarxiv, Center for Open Science.

Jaffe, S. and Weyl, E. G. (2013). The first-order approach to merger analysis. *American Economic Journal: Microeconomics*, 5(4):188–218.

Jürgens, U. (2012). Standortmuster von Lebensmitteldiscoutern - eine systematische Betrachtung am Beispiel Schleswig-Holsteins. *Berichte des Arbeitskreises Geographische Handelsforschung*, 31:35–41.

Krüger, T., Anders, S., Walther, M., Klein, K., and Segerer, M. (2013). Qualifizierte Nahversorgung im Lebensmitteleinzelhandel - Bericht im Auftrag von HDE und BVL. Technical report, HafenCity University Hamburg and Universität Regensburg.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In *Frontiers in Economics*. Academic Press.

Miller, N. H., Remer, M., Ryan, C., and Sheu, G. (2013). On the first order approximation of counterfactual price effects in oligopoly models. Doj working paper.

Miller, N. H., Remer, M., Ryan, C., and Sheu, G. (2017). Upward pricing pressure as a predictor of merger price effects. *International Journal of Industrial Organization*, 52(C):216–247.

Moresi, S. (2009a). Bidding competition and the UPP test. Technical report, Horizontal Merger Guidelines Review Project - Comment, Project No. P092900.

Moresi, S. (2009b). Cournot competition and the UPP test. Technical report, Horizontal Merger Guidelines Review Project - Comment, Project No. P092900.

Neumeier, S. (2014). Modellierung der Erreichbarkeit von Supermärkten und Discountern: Untersuchung zum regionalen Versorgungsgrad mit Dienstleistungen der Grundversorgung. Thünen Working Papers 16, Johann Heinrich von Thünen Institute, Federal Research Institute for Rural Areas, Forestry and Fisheries.

Petrin, A. and Train, K. (2010). A control function approach to endogeneity in consumer choice models. *Journal of marketing research*, 47(1):3–13.

Reiss, P. C. and Wolak, F. A. (2007). Structural econometric modeling: Rationales and examples from industrial organization. In Heckman, J. J. and Leamer, E. E., editors, *Handbook of Econometrics*, volume 6, pages 4277–4415. Elsevier.

Richards, T. J., Acharya, R. N., and Molina, I. (2011). Retail and wholesale market power in organic apples. *Agribusiness*, 27(1):62–81.

Schmalensee, R. (2009). Should new merger guidelines give UPP market definition? *GCP -The Antitrust Chronicle*.

Shapiro, C. (2010). Unilateral effects calculations. Technical report, University of California at Berkeley.

Simons, J. J. and Coate, M. B. (2010). Upward pressure on price analysis: Issues and implications for merger policy. *European Competition Journal*, 6(2):377–396.

Train, K. (2009). *Discrete choice methods with simulation*. Cambridge University Press.

Werden, G. J. (1996). A robust test for consumer welfare enhancing mergers among sellers of differentiated products. *Journal of Industrial Economics*, 44(4):409–413.

Zeigermann, L. (2016). OPENCAGEGEO: Stata module for geocoding. Statistical Software Components, Boston College Department of Economics.

Zeigermann, L. (2025). LHAVERSINE: Stata module to compute great-circle distances with haversine formula. Statistical Software Components, Boston College Department of Economics.

Appendix

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13
R1	.983	.961	.32	.317	.413	.977	.129	.909	.738	.797	.543	.849	.477
R2	.961	.976	.314	.316	.413	.967	.129	.889	.728	.781	.544	.854	.463
R3	.32	.314	.321	.111	.165	.321	.072	.304	.242	.29	.226	.297	.176
R4	.317	.316	.111	.318	.204	.318	.065	.314	.276	.288	.195	.307	.1
R5	.413	.413	.165	.204	.414	.414	.12	.407	.292	.373	.27	.382	.236
R6	.977	.967	.321	.318	.414	.99	.129	.909	.743	.797	.544	.859	.476
R7	.129	.129	.072	.065	.12	.129	.129	.129	.113	.129	.108	.129	.101
R8	.909	.889	.304	.314	.407	.909	.129	.912	.688	.768	.537	.809	.45
R9	.738	.728	.242	.276	.292	.743	.113	.688	.745	.625	.437	.667	.3
R10	.797	.781	.29	.288	.373	.797	.129	.768	.625	.799	.477	.716	.409
R11	.543	.544	.226	.195	.27	.544	.108	.537	.437	.477	.545	.502	.31
R12	.849	.854	.297	.307	.382	.859	.129	.809	.667	.716	.502	.862	.381
R13	.477	.463	.176	.1	.236	.476	.101	.45	.3	.409	.31	.381	.478

Table A3.1: Share of common choice sets for all retailer pairs

	Ret. 1	Ret. 2	Ret. 3	Ret. 4	Ret. 5	Ret. 6	Ret. 7	Ret. 8	Ret. 9	Ret. 10	Ret. 11	Ret. 12	Ret. 13
Ret. 1	.	1.8052	1.6648	0.2680	1.3527	6.0291	0.0740	2.4226	0.6604	1.1073	1.8329	1.5249	0.2314
Ret. 2	7.7276	.	1.2564	0.2024	1.0144	4.5505	0.0556	1.8292	0.4982	0.8355	1.3876	1.1517	0.1744
Ret. 3	7.6863	1.3565	.	0.2011	1.0141	4.5245	0.0555	1.8186	0.4957	0.8311	1.3757	1.1443	0.1737
Ret. 4	7.2723	1.2835	1.1820	.	0.9575	4.2798	0.0526	1.7244	0.4689	0.7862	1.2968	1.0818	0.1644
Ret. 5	7.5995	1.3344	1.2339	0.1981	.	4.4728	0.0550	1.7967	0.4901	0.8217	1.3538	1.1320	0.1717
Ret. 6	8.9858	1.5838	1.4609	0.2351	1.1866	.	0.0650	2.1255	0.5796	0.9717	1.6092	1.3383	0.2030
Ret. 7	7.2123	1.2668	1.1729	0.1891	0.9534	4.2481	.	1.7059	0.4652	0.7802	1.2924	1.0737	0.1631
Ret. 8	7.9126	1.3975	1.2853	0.2081	1.0442	4.6576	0.0571	.	0.5101	0.8554	1.4141	1.1780	0.1788
Ret. 9	7.3867	1.3017	1.2010	0.1933	0.9757	4.3494	0.0534	1.7473	.	0.7988	1.3227	1.1000	0.1669
Ret. 10	7.5198	1.3252	1.2227	0.1968	0.9933	4.4276	0.0544	1.7788	0.4850	.	1.3465	1.1198	0.1699
Ret. 11	7.7364	1.3662	1.2579	0.2017	1.0167	4.5572	0.0560	1.8290	0.4991	0.8368	.	1.1530	0.1748
Ret. 12	7.6430	1.3472	1.2425	0.1996	1.0109	4.5007	0.0552	1.8075	0.4930	0.8265	1.3706	.	0.1727
Ret. 13	7.2595	1.2783	1.1803	0.1899	0.9585	4.2741	0.0525	1.7172	0.4682	0.7851	1.3003	1.0809	.

Table A3.2: Average UPP between all retailer pairs with global choice sets

	Ret. 1	Ret. 2	Ret. 3	Ret. 4	Ret. 5	Ret. 6	Ret. 7	Ret. 8	Ret. 9	Ret. 10	Ret. 11	Ret. 12	Ret. 13
Ret. 1	.	1.9056	1.6750	0.2726	1.3512	6.4866	0.0644	2.5812	0.6924	1.1557	1.8333	1.5501	0.2502
Ret. 2	8.1547	.	1.1798	0.1961	0.9772	4.9506	0.0465	1.8028	0.5348	0.8236	1.3654	1.3690	0.1819
Ret. 3	7.7342	1.2728	.	0.1903	1.2363	4.4511	0.0898	1.8026	0.4904	0.8942	1.7822	1.1528	0.1847
Ret. 4	7.3948	1.2433	1.1181	.	1.4549	4.2220	0.0620	1.8630	0.5102	0.8470	1.3568	1.1568	0.0875
Ret. 5	7.5902	1.2835	1.5043	0.3009	.	4.3495	0.1188	1.9012	0.4457	0.9069	1.5984	1.1515	0.1971
Ret. 6	9.6678	1.7236	1.4368	0.2319	1.1537	.	0.0547	2.2024	0.6234	0.9839	1.5672	1.4381	0.2084
Ret. 7	6.2704	1.0569	1.8978	0.2231	2.0586	3.5777	.	1.6191	0.4526	0.8604	1.7792	1.0586	0.2171
Ret. 8	8.4299	1.3780	1.2742	0.2246	1.1059	4.8258	0.0543	.	0.5188	0.9059	1.5152	1.2066	0.1823
Ret. 9	7.7449	1.3997	1.1886	0.2103	0.8875	4.6784	0.0519	1.7775	.	0.8594	1.3812	1.1341	0.1495
Ret. 10	7.8485	1.3062	1.3157	0.2119	1.0960	4.4829	0.0600	1.8839	0.5218	.	1.4479	1.1365	0.1789
Ret. 11	7.7379	1.3442	1.6324	0.2108	1.2020	4.4382	0.0770	1.9593	0.5211	0.8994	.	1.1763	0.1986
Ret. 12	7.7691	1.5982	1.2515	0.2134	1.0276	4.8373	0.0544	1.8516	0.5083	0.8390	1.3976	.	0.1429
Ret. 13	7.8493	1.3359	1.2543	0.1011	1.1003	4.3875	0.0700	1.7508	0.4194	0.8264	1.4772	0.8950	.

Table A3.3: Average UPP between all retailer pairs with local choice sets

	Ret. 1	Ret. 2	Ret. 3	Ret. 4	Ret. 5	Ret. 6	Ret. 7	Ret. 8	Ret. 9	Ret. 10	Ret. 11	Ret. 12	Ret. 13
Ret. 1	.	5.564	0.613	1.703	-0.108	7.587	-13.048	6.545	4.848	4.369	0.024	1.650	8.119
Ret. 2	5.527	.	-6.098	-3.125	-3.662	8.792	-16.501	-1.445	7.336	-1.422	-1.604	18.868	4.249
Ret. 3	0.623	-6.164	.	-5.396	21.908	-1.623	61.769	-0.881	-1.072	7.583	29.550	0.743	6.304
Ret. 4	1.685	-3.131	-5.407	.	51.960	-1.352	18.023	8.0389	8.807	7.731	4.627	6.926	-46.754
Ret. 5	-0.121	-3.813	21.913	51.850	.	-2.756	116.278	5.821	-9.065	10.371	18.063	1.718	14.781
Ret. 6	7.590	8.822	-1.646	-1.337	-2.770	.	-15.821	3.620	7.557	1.250	-2.607	7.462	2.662
Ret. 7	-13.060	-16.567	61.808	17.943	115.928	-15.780	.	-5.087	-2.707	10.271	37.673	-1.411	33.071
Ret. 8	6.538	-1.396	-0.864	7.916	5.905	3.611	-4.965	.	1.689	5.901	7.149	2.422	1.963
Ret. 9	4.849	7.533	-1.034	8.820	-9.036	7.564	-2.743	1.729	.	7.583	4.423	3.096	-10.468
Ret. 10	4.371	-1.433	7.612	7.720	10.343	1.249	10.284	5.910	7.584	.	7.529	1.450	5.269
Ret. 11	0.020	-1.616	29.766	4.495	18.225	-2.612	37.548	7.122	4.403	7.487	.	2.021	13.602
Ret. 12	1.650	18.636	0.727	6.929	1.658	7.478	-1.505	2.443	3.100	1.512	1.974	.	-17.227
Ret. 13	8.124	4.508	6.272	-46.749	14.786	2.653	33.304	1.955	-10.428	5.266	13.605	-17.198	.

Table A3.4: Average percentage difference in UPP between local choice and global choice sets for all retailer pairs

Chapter 4

Mixlelast: Stata Module for Mixed Logit Elasticities and Marginal Effects

4.1 Introduction

Mixed logit models have become very popular in discrete choice analysis. This is due to their greater flexibility and the more realistic substitution patterns between alternatives compared to simpler discrete choice models. Thanks to the excellent user-written `mixlogit` by Hole (2007), it is fairly simple for the researcher to fit mixed logit models in Stata. As there exists no straightforward interpretation for the estimated parameters beyond their signs, researchers often compute elasticities or marginal effects. In contrast to simpler logit models, computing elasticities and marginal effects for mixed logit models is not trivial and requires simulation.

This chapter describes `mixlelast`, a post-estimation command after `mixlogit`, which allows the user to obtain various forms of mixed logit elasticities and marginal effects and to compute bootstrapped standard errors and confidence intervals.[1]

The remainder of this chapter is structured as follows: Section 4.2 briefly outlines the mixed logit model. The various sorts of mixed logit elasticities and marginal effects are discussed in sections 4.3 and 4.4. Section 4.5 deals with sample elasticities and marginal effects in the case of heterogeneous choice sets.[2] A bootstrap procedure for standard errors and confidence intervals is presented in section 4.6. Finally, the syntax of `mixlelast` is described in section 4.7, and section 4.8 provides some examples.

4.2 Mixed Logit Model[3]

As in Revelt and Train (1998), I assume N decision-makers to choose among J alternatives on T choice occasions. The utility for individual n from choosing alternative j on choice occasion t is given by

$$U_{njt} = \beta_n' x_{njt} + \varepsilon_{njt},$$

where β_n is a vector of individual-specific coefficients for decision-maker n and x_{njt} is a vector of attributes. The error term, ε_{njt}, is assumed to be independently and identically distributed extreme value I. The density of β is denoted $f(\beta|\theta)$, where θ are the parameters of the distribution, i.e. the mean and the (co)variance.

[1] `mixlelast` is published in the the Boston College Statistical Software Components (SSC) archive, see Zeigermann (2024).

[2] I.e. the size and composition of the choice set differ between choice occasions.

[3] This section draws heavily from Hole (2007). See Revelt and Train (1998) and Train (2009) for an in-depth discussion of the model.

The conditional probability, i.e. conditional on β_n, for decision-maker n of choosing alternative i on choice occasion t is

$$L_{nit}(\beta_n) = \frac{exp(\beta_n' x_{nit})}{\sum_{j=1}^{J} exp(\beta_n' x_{njt})}. \quad (4.1)$$

Then, the conditional probability of the observed sequences of choices made by decision-maker n is given by

$$S_n(\beta_n) = \prod_{t=1}^{T} L_{ni(n,t)t}(\beta_n),$$

where $i(n,t)$ denotes the alternative chosen by n on choice occasion t. The *unconditional* probability of the observed sequence of choices is the conditional probability integrated over the distribution of β:

$$P_n(\theta) = \int S_n(\beta|\theta) f(\beta) d\beta$$

The researcher specifies the form of the density function f and estimates its parameters θ by maximum likelihood. The log-likelihood function is given by

$$LL(\theta) = \sum_{n=1}^{N} \ln P_n(\theta). \quad (4.2)$$

As no analytical solution exists for this expression, it needs to be approximated using simulation techniques, see Train (2009). The simulated equivalent of (4.2) is given by

$$SLL(\theta) = \sum_{n=1}^{N} \ln \left[\frac{1}{R} \sum_{r=1}^{R} S_n(\beta^r) \right],$$

where R is the number of draws and β^r is the rth draw from $f(\beta|\theta)$.

The estimated parameters $\hat{\theta}$ have no straightforward interpretation beyond their sign. Therefore, to obtain meaningful and interpretable results for discrete choice models, researchers often compute elasticities or marginal effects of various kinds.

4.3 Elasticities

Elasticities are a unitless measure describing the relationship between a percentage change in an attribute of an alternative and the percentage change in the choice probability, ceteris paribus. Elasticities can be differentiated on the basis of following three categories, see Hensher et al. (2015).

First, we distinguish between *direct* and *cross* elasticities. The former quantify the percentage change in the choice probability of an alternative resulting from a percentage change in an attribute of the *same* alternative. A prominent example of direct elasticities is own-price elasticities, i.e. the change in the choice probability of an alternative when its own price changes. The latter, in contrast, describes the change in the choice probability of one alternative with respect to a percentage change in an attribute of a *competing* alternative. Here, cross-price elasticities are a common example describing the change in the choice probability resulting from a price change of a rival alternative.

Second, we differentiate between two calculation methods, namely the *point* and the *arc* method. Point elasticities use differential calculus and evaluate a marginal change in an attribute of one alternative. The arc method calculates the choice probabilities before and after a (non-marginal) attribute change and evaluates the difference. This allows the researcher to compute the effects of a discrete change in an attribute, e.g. a price increase of 20 per cent or of a dummy variable change.

Third, there is a distinction between *individual* and *sample* elasticities. Data on individual choices of decision-makers allows us to compute individual elasticities for every alternative on every choice occasion. However, researchers are most often more interested in some form of average elasticities over the entire sample. As sample elasticities are (unweighted or weighted) averages of individual elasticities, the latter will be discussed before turning to different ways of aggregation.

4.3.1 Individual Elasticities

Point elasticities

Point elasticities describe the percentage change in the choice probability given a small change in one attribute. Recalling the s-shape of the logit curve makes it clear why point elasticities are only valid for marginal changes at a specific point on the logit curve. With a non-marginal change in an attribute, we move along the logit curve and may encounter a considerably different slope compared to that at the original attribute value. Ignoring the impact of the non-linearity of the logit curve can give rise to substantially biased results.

It should be noted that calculating point elasticities only makes sense for continuous variables. Although it is mathematically possible to compute point elasticities for integer or dummy variables, it is meaningless. The appropriate way to evaluate the impact of a discrete change in an attribute is to use the arc method described later.

Direct elasticities Stated formally, the direct point elasticity of alternative i of decision-maker n for a marginal change in the mth attribute of the attribute vector x is given by[4]

$$E_{nix_{ni}^m} = \frac{x_{ni}^m}{P_{ni}} \frac{\partial P_{ni}}{\partial x_{ni}^m}, \tag{4.3}$$

where

$$P_{ni} = \int L_{ni}(\beta) f(\beta) d\beta. \tag{4.4}$$

Recall that L_{ni} is the conditional probability of decision-maker n of choosing alternative i as given in (4.1). Taking the derivative of (4.4) with respect to x_{ni}^m and plugging into (4.3) yields

$$E_{nix_{ni}^m} = \frac{x_{ni}^m}{P_{ni}} \int \beta^m L_{ni}(\beta)(1 - L_{ni}(\beta)) f(\beta) d\beta, \tag{4.5}$$

where β^m is the mth element of vector β.

The integral in (4.5) cannot be solved analytically and is therefore approximated using simulation techniques. The simulated version is given by

$$E_{nix_{ni}^m} \approx \frac{x_{nj}^m}{P_{ni}} \left(\frac{1}{R} \sum_{r=1}^{R} \beta^{mr} L_{ni}(\beta^r)(1 - L_{ni}(\beta^r)) \right),$$

where β^r denotes the rth draw from the distribution of β, see Train (2009).

Cross elasticities The cross point elasticity of alternative i for individual n given a marginal change in the mth attribute of the rival alternative j is then

$$E_{nix_{nj}^m} = \frac{x_{nj}^m}{P_{ni}} \frac{\partial P_{ni}}{\partial x_{nj}^m}.$$

Plugging in the first derivative of (4.4) with respect to x_{nj}^m yields

$$E_{nix_{nj}^m} = -\frac{x_{nj}^m}{P_{ni}} \int \beta^m L_{ni}(\beta) L_{nj}(\beta) f(\beta) d\beta. \tag{4.6}$$

[4]For ease of notation, I drop T in what follows. I.e., I assume each of the N decision-makers to choose only once.

The simulated equivalent of (4.6) is then given by

$$E_{nix_{nj}^m} \approx -\frac{x_{nj}^m}{P_{ni}}\left(\frac{1}{R}\sum_{r=1}^{R}\beta^{mr}L_{ni}(\beta^r)L_{nj}(\beta^r)\right).$$

Arc elasticities

In contrast to point elasticities, arc elasticities consider a discrete change in one attribute rather than a marginal change.[5] Here, we calculate expected probabilities both with the original attribute value and with the changed attribute value and then evaluate the difference. This approach allows us to handle the non-linearity of the logit curve, which affects our results once we consider non-marginal attribute changes.

Direct elasticities Formally, the direct arc elasticity of alternative i for individual n for a discrete change in the mth attribute is given by

$$E_{nix_{ni}^m} = \frac{\bar{x}_{ni}^m}{\bar{P}_{ni}}\frac{\Delta P_{ni}}{\Delta x_{ni}^m}. \tag{4.7}$$

We first consider the second right-hand side term of (4.7). The numerator, ΔP_{ni}, is the difference in the choice probability for the attribute value before the change, x_{ni}, and after the change $\hat{x}_{ni}$, i.e.

$$\Delta P = P_{ni}(x_n) - P_{ni}(\hat{x}_n).$$

The denominator, Δx_{ni}^m, is simply the difference between the original and the changed attribute value:

$$\Delta x_{nj} = x_{nj}^m - \hat{x}_{nj}^m$$

The first right-hand side term comprises the average choice probability before and after the attribute change[6]

$$\bar{P}_{ni} = (P_{ni}(x_n) + P_{ni}(\hat{x}_n))/2$$

[5]The term arc elasticity is not used consistently in the literature. Sometimes, researchers refer to arc elasticities when the first derivative is not multiplied by x/P, but by the average over the value before and after the attribute change, i.e. $\Delta x/\Delta P$. This approach can be seen as a linear approximation that ignores the non-linearity of the logit curve.

[6]The approach followed here, i.e. taking the mean probability and mean attribute value, is sometimes referred to as the midpoint method. Two alternatives not considered multiply by either the original or the changed values and the respective predicted probabilities.

and the attribute mean

$$\bar{x}^m = (x^m_{ni} + \hat{x}^m_{ni})/2.$$

For integer variables, we can now specify $\hat{x} = x + a$, where a can take any integer value. By the same token, elasticities for dummy variables are calculated by setting $\hat{x} = 0$ if $x = 1$ and vice versa.

Cross elasticities Equivalently, cross arc elasticities are given by

$$E_{nix^m_{nj}} = \frac{\bar{x}^m_{nj}}{\bar{P}_{ni}} \frac{\Delta P_{ni}}{\Delta x^m_{nj}},$$

where ΔP_{ni} now denotes the difference between the choice probabilities for alternative i with the original attributes of j, x^m_{nj}, and the changed attribute m, $\hat{x}^m_{nj}$.

Arc vs. point elasticities

Having established the two different methods, the question arises as to which is the correct one to use. For dummy and integer variables, the answer is obvious, as marginal changes are meaningless. In these cases, we always use arc elasticities. For continuous variables, the answer depends strongly on the question we seek to answer. If we want to evaluate the decision-makers' sensitivity to attribute changes and establish general substitution patterns across alternatives, the point method is appropriate. If, in contrast, we want to evaluate how a substantial change in an attribute, e.g. a price increase by 20 per cent, affects choice behaviour, we need to compute arc elasticities.

Calculating arc elasticities for very small discrete attribute changes should generate results that are very close to point elasticities. Without any post-estimation command available, this was a practicable solution to approximate point elasticities.[7] With `mixlelast`, the researcher is strongly advised to use the point method rather than calculating arc elasticities for very small changes. First, arc elasticities for small changes are only an approximation of true point elasticities. Second, calculating arc elasticities is more computationally intensive, which can result in significantly longer computation times, particularly when working with large data sets.

[7] Cameron and Trivedi (2010) advocate calculating arc elasticities for a discrete change of a one-thousandth of the standard deviation of the attribute to approximate point elasticities.

Interpretation

Elasticities can take any value between infinity and minus infinity. Positive direct elasticities indicate that an increase in an attribute of an alternative leads to an increase in the choice probability of that alternative. For cross elasticities, positive values indicate that an increase in an attribute leads to a higher choice probability of the rival alternative. For negative elasticities, the effects are reversed.

If a one per cent change in an attribute results in a one per cent change in the choice probability, we speak of unit elasticity. If the percentage in the choice probability is smaller or larger than one, we speak about relatively inelastic and relatively elastic elasticities, respectively. Two extreme cases are perfect inelasticity and perfect elasticity. In the former case, an attribute change leaves the choice probability unchanged. In the latter case, the choice probability will drop to zero following an increase of one per cent in the attribute. The various types of elasticities are summarised in Table 4.1 below.

	Absolute value	**Direct elasticity**	**Cross elasticity**
Perfectly inelastic	$E = 0$	no change in P_i	no change in P_j
Relatively inelastic	$0 < E < 1$	change in P_i less than one percent	change in P_j less than one percent
Unit elastic	$E = 1$	one percent change in P_i	one percent change in P_j
Relatively elastic	$1 < E < \infty$	change in P_i larger than one percent	change in P_j larger than one percent
Perfectly elastic	$E = \infty$	change in P_i is ∞	change in P_j is ∞

Table 4.1: Changes in the choice probability given a one per cent change in an attribute of alternative i, adapted from Hensher et al. (2015)

Mixed vs. conditional logit elasticities

Having established expressions for individual direct and cross-point elasticities, we can now compare mixed logit elasticities with those from standard conditional logit models.

The conditional logit formula (McFadden 1973) is

$$P_{nit} = \frac{exp(\beta' x_{nit})}{\sum_{j=1}^{J} exp(\beta' x_{njt})}.$$

In contrast to the mixed logit model, the β coefficients are assumed to be identical for all decision-makers. Now, direct elasticities are given by

$$E_{nix_{ni}^m} = \frac{x_{ni}^m}{P_{ni}} \frac{\partial P_{ni}}{\partial x_{ni}^m}. \tag{4.8}$$

Taking the derivative in (4.8) yields

$$E_{nix_{ni}^m} = \frac{x_{ni}}{P_{ni}} \beta P_{ni}(1 - P_{ni}),$$

which simplifies to

$$E_{nix_{ni}^m} = x_{ni}\beta(1 - P_{ni}). \tag{4.9}$$

Contrary to the complex mixed logit elasticity in (4.5) containing integrals that need to be simulated, direct logit elasticities given in (4.9) are a simple function of the attribute value, the estimated coefficient and the predicted choice probability.

More insightful, however, is the comparison of cross elasticities. Conditional logit cross elasticities are given by

$$\begin{aligned} E_{nix_{nj}^m} &= \frac{x_{nj}^m}{P_{ni}} \frac{\partial P_{ni}}{\partial x_{nj}^m} \\ &= -\frac{x_{nj}^m}{P_{ni}} \beta P_{ni} P_{nj}. \end{aligned}$$

As before, P_{ni} cancels out:

$$E_{nix_{nj}^m} = -x_{nj}^m \beta P_{nj} \tag{4.10}$$

As we see in (4.10), cross elasticities for a marginal change in one attribute are identical for all i. Hence, a change in an attribute of j gives rise to the exact same change in the choice probability of all rival alternatives. This property of logit elasticities is sometimes referred to as "proportionate shifting".[8] While this may be realistic for some choice situations, it is not in many others. If some alternatives are similar to each, while others are very different, we would expect substitution between the former to be stronger than between the latter.

[8] See Train (2009) for an in-depth discussion of the properties of logit elasticities.

Here, the strength of mixed logit models comes into play. Recall the mixed logit cross elasticities given in (4.6):

$$E_{nix^m_{nj}} = -\frac{x^m_{nj}}{P_{ni}} \int \beta^m L_{ni}(\beta) L_{nj}(\beta) f(\beta) d\beta$$

This expression is different for each alternative i. The percentage change in the probability P_{ni} depends on the correlation of L_{ni} and L_{nj} over different realisations of β. For alternatives with positively correlated x^m, substitution will be stronger than for alternatives with no or negative correlation, see Train (2009). Therefore, similar alternatives will exhibit much stronger substitution than those which are very different.

The comparison also illustrates that using the more flexible mixed logit framework does not come without costs. While there are simple closed-form solutions for the logit model and its elasticities, simulation techniques are required for both the estimation of the mixed logit model and the calculation of elasticities.

4.3.2 Sample Elasticities

The elasticities above are for one individual decision-maker on one particular choice occasion. With J alternatives, this gives J^2 values on a single choice occasion: J direct elasticities and $(J-1) \cdot J$ cross elasticities. For N decision-makers, this amounts to $N \cdot J^2$ values.[9]

Thus, the first obvious reason why researchers usually want to compute sample elasticities is to reduce the amount of output produced and to make it readable. Taking the sample averages reduces the number of values obtained by a factor of N to J^2. These can be practically displayed in a JxJ matrix with direct elasticities on the diagonal. A second and equally important argument in favour of sample elasticities is that the underlying model is estimated on a sample of choice data. Therefore, meaningful elasticities can only be calculated for the sample as a whole, rather than for each decision-maker separately, see Hensher et al. (2015).

Unweighted sample average

The most straightforward way of obtaining sample averages is to simply calculate the mean over all decision-makers in the sample. Sample direct and cross elasticities are then given by

[9] If we also consider T choice occasion, the number of values rises to $T \cdot N \cdot J^2$.

$$E_{ix_i^m} = \frac{1}{N}\sum_{n=1}^{N} E_{nix_i^m}$$

and

$$E_{ix_j^m} = \frac{1}{N}\sum_{n=1}^{N} E_{nix_j^m}$$

respectively, where $E_{ix_i^m}$ and $E_{ix_j^m}$ can either be individual point or arc elasticities.

Probability weighted sample average

In the simple approach above, all elasticities enter with the same weight, regardless of the contribution to the choice outcome of each alternative. Now, suppose an alternative has a very high choice probability for one decision-maker, but a low one for another.[10] Simply averaging over the elasticities from those two choice occasions ignores this and is sometimes referred to as "naive pooling", see Louviere et al. (2000). To account for variations in individual choice probabilities for the same alternative between different decision-makers, we can compute probability-weighted sample averages using the decision-maker's choice probabilities as weights.[11]

Formally, probability-weighted sample elasticities are then

$$wE_{ix_i^m} = \left(\sum_{n=1}^{N} P_{ni}E_{nix_i^m}\right) \bigg/ \sum_{n=1}^{N} P_{ni}$$

and

$$wE_{ix_j^m} = \left(\sum_{n=1}^{N} P_{ni}E_{nix_j^m}\right) \bigg/ \sum_{n=1}^{N} P_{ni}$$

where $E_{ix_i^m}$ and $E_{ix_j^m}$ can be calculated using the point or the arc method.

4.4 Marginal Effects

In contrast to elasticities, marginal effects describe a *unit* change in the choice probability given a *unit* change in an attribute.

[10] This situation can arise if the attribute values for the alternatives are different for the two decision-makers or if the choice sets are not identical for both. The latter case is discussed in detail in section 4.5.

[11] For a discussion of these issues (in the context of the conditional logit model) see Louviere et al. (2000).

The direct marginal effect for decision-maker n and alternative i with respect to a unit change in x_{ni}^m is given by

$$ME_{nix_{ni}^m} = \frac{\partial P_{ni}}{\partial x_{ni}^m}. \tag{4.11}$$

Comparing (4.11) and the expression for direct point elasticities given in (4.3), the difference between the two becomes obvious. It is the second term in (4.7), x_{ni}^m/P_{ni}, which transforms a marginal effect into an elasticity.

Equivalently, the cross marginal effect is given by

$$ME_{nix_{nj}^m} = \frac{\partial P_{ni}}{\partial x_{nj}^m}.$$

For the same reason as for elasticities, we need to consider discrete changes for discrete and dummy variables. In this context, is it more meaningful to refer to incremental rather than marginal effects, as the impact considered here is not marginal. However, to avoid further complication of the notation, I continue using ME in the following equations describing incremental effects.

Direct and cross-incremental effects for discrete and dummy variables using the arc method are given by

$$ME_{nix_{ni}^m} = \frac{\Delta P_{ni}}{\Delta x_{ni}}$$

and

$$ME_{nix_{nj}^m} = \frac{\Delta P_{ni}}{\Delta x_{nj}^m},$$

respectively.

The same methods of aggregation discussed in section 4.3.2 can be applied.

4.5 Heterogeneous Choice Sets

So far, we have implicitly assumed choice sets to be homogeneous across individuals, i.e. $J^n = J$. With homogeneous choice sets, we get a full JxJ matrix of direct and cross effects for all N decision-makers. However, it is possible that the availability of alternatives differs across decision-makers.[12] That is, only a subset J^n of the J alternatives is available to decision-maker n. Then, cross and direct effects are only

[12] The availability of alternatives might even differ across choice occasions faced by the same decision-maker if we deal with panel data where individuals choose repeatedly.

defined for the alternatives contained in J^n. That is, we get J^n direct effects and $(J^n - 1) * J^n$ cross effects between all available alternatives.

Considering the entire sample of all N decision-makers, we obtain between one and N individual direct effects for an alternative: N if the respective alternative is contained in all choice sets and one if it is available on only a single choice occasion. For cross effects between two alternatives to be defined, both need to be available at the same time. The number of choice occasions for which this holds can range between zero and N; it is zero if there is no choice set containing both alternatives, and it is N if both are always available.

Now, the question arises of how to calculate average elasticities or marginal effects over a sample of decision-makers that are heterogeneous regarding the alternatives to choose from. In the subsequent section, I first discuss the aggregation of direct effects before turning to cross effects.

4.5.1 Direct Elasticities

Two different ways of calculating sample averages for direct effects exist are described below.

First, we consider all N decision-makers in the sample, ignoring whether the respective alternative is available or not. That is, we also include those decision-makers that do not react to the attribute change because the affected alternative is not available to them and set their undefined choice probability to zero[13]. This type of aggregation shall be denoted **type I**.

Second, we restrict our attention to those who can actually choose the alternative and will hence be affected by an attribute change. This shall be referred to as aggregation of **type II** in what follows.

Unweighted sample averages Formally, unweighted sample averages of type I are given by

$$E^I_{ix_i^m} = \frac{1}{N} \sum_{n=1}^{N} E_{nix_i^m}.$$

[13]In this case, setting the choice probability to zero makes sense intuitively as the decision-maker's choice probability is zero before and after the attribute change.

Type II unweighted sample elasticities are

$$E^{II}_{ix^m_i} = \frac{1}{N^i} \sum_{n \in N^i} E_{nix^m_i},$$

where N^i denotes the subset of decision-makers that hold alternative i in their choice set.

Probability weighted sample averages Applying the same logic of aggregation to probability-weighted sample elasticities, we get the following expression for type I:

$$wE^{I}_{ix^m_i} = \sum_{n=1}^{N} P_{ni} E_{nix^m_i} \bigg/ \sum_{n=1}^{N} P_{ni}$$

The observant reader will have noticed that the choice probability P_{ni} is not defined if alternative i is not in n's choice set. For convenience, we set it to zero, indicating that alternative i is chosen with zero probability. For this reason, type I weighted sample averages are identical to their type II counterparts given by

$$wE^{II}_{ix^m_i} = \sum_{n \in N^i} P_{ni} E_{nix^m_i} \bigg/ \sum_{n \in N^i} P_{ni}.$$

4.5.2 Cross Elasticities

Regarding cross elasticities, we need to distinguish three cases denoted by **type I**, **IIa** and **IIb** in what follows. First, **type I** aggregation considers all N decision-makers no matter if alternative i and/or j is present in the choice set. Second, **type IIa** aggregates over all decision-makers with alternative i in their choice sets regardless of whether alternative j is also available. Third, **type IIb** only considers decision-makers with both i and j in their choice sets.

Unweighted sample averages Formally, unweighted average sample elasticities of type I are given by

$$E^{I}_{ix^m_j} = \frac{1}{N} \sum_{n=1}^{N} E_{nix^m_j}.$$

Aggregation by type IIa and IIb yields

$$E^{IIa}_{ix^m_j} = \frac{1}{N^i} \sum_{n \in N^i} E_{nix^m_j}$$

and

$$E^{IIb}_{ix^m_j} = \frac{1}{N^{ij}} \sum_{n \in N^{ij}} E_{nix^m_j},$$

respectively, where N^{ij} denotes the subset of N where both alternatives are available.

Probability weighted sample averages Probability-weighted sample elasticities of type I are given by

$$E^{I}_{ix^m_j} = \sum_{n=1}^{N} P_{ni} E_{nij} \Big/ \sum_{n=1}^{N} P_{ni}.$$

Again, we set the choice probability and hence the weight to zero for all decision-makers who do not have alternative i in their choice set. For this reason, type I and type IIa aggregation given below are identical:

$$E^{IIa}_{ix^m_j} = \sum_{n \in N^i} P_{ni} E_{ix^m_j} \Big/ \sum_{n \in N^i} P_{ni}$$

Sample cross elasticities of type IIb are finally given by

$$E^{IIb}_{ix^m_j} = \sum_{n \in N^{ij}} P_{ni} E_{nix^m_i} \Big/ \sum_{n \in N^{ij}} P_{ni}.$$

4.5.3 Choice of Aggregation Type

Having established different ways of aggregation, we need to determine which is the correct one to use. Once again, the answer very much depends on the question we are asking. If we are interested in a measure of how the entire sample of all decision-makers - being affected or not - reacts to an attribute change, we need to use type I. If, however, we care about those who can potentially choose alternative i, type II is to be used instead. For cross-effects, we further need to decide if all individuals holding alternative i in their choice sets are the sub-sample of interest (i.e. type IIa) or if we want to restrict our attention to those who could also choose the rival alternative j (type IIb) and are therefore affected by an attribute change in j.

4.6 Standard Errors and Confidence Intervals

Recall that when estimating a mixed logit model, we technically estimate the parameters of the distribution denoted θ, i.e. the mean and the (co)variance. In what we have discussed so far, we simply take the point estimates $\hat{\theta}$ of θ to simulate elasticities and marginal effects as described above. This gives us sample effects and standard deviations describing how much individual effects vary across the sample. However, we have ignored the sampling variance $\widehat{W}$ of the estimated $\hat{\theta}$. To obtain a measure of precision of our average sample effects, i.e. standard errors or confidence intervals, we need to incorporate the sampling variance of the estimated coefficients, $\widehat{W}$, into our calculation of elasticities and marginal effects.

We can generally think of three different methods for generating standard errors and confidence intervals. These are the non-parametric bootstrap, the parametric bootstrap (Krinsky-Robb method) and the delta-method. In what follows, I present the solution proposed by Krinsky and Robb (1986, 1990) as it does not require re-estimating the model and is fairly easy to implement.[14]

The Krinsky-Robb-method involves the following four steps:

1. Compute the Cholesky-decomposition of the covariance matrix $\widehat{W}$, yielding a lower triangular matrix L such that $LL' = \widehat{W}$, see e.g. Greene (2008).

2. For $t = 1, \dots, T$ create a $Kx1$ vector of random draws from a standard normal distribution, where K is the length of vector θ and label it η^t. Then create $\theta^t = \hat{\theta} + L\eta^t$.

3. For $t = 1, \dots, T$ use θ^t to simulate elasticities or marginal effects as described above.

4. Calculate the mean sample effect by averaging over the T draws. To obtain standard errors, compute the variance and take the square root. For the 95 per cent confidence interval, take the 0.025 and 0.975 percentile.

[14] For the non-parametric bootstrap, we would need to re-estimate the model multiple times. This is often far too time-consuming, and there is a risk that the model will not converge for some subsamples. The delta method produces confidence intervals that are symmetric about the mean by construction.

4.7 Mixlelast

This section describes the syntax of `mixlelast`.

Syntax

`mixlelast` [*if*] [*in*] , `alternatives(`*varname*`)` [`for(`*varname*`)` `marginaleffect` `percentchange(#)` `absolutechange(#)` `dummy` `weighted` `nosd` `nrep(#)` `burn(#)` `hettype()` `krobb(#)` `krse` `krlevel(#)` `kruserdraws` `krburn(#)` `quietly`]

Options

`alternatives(`*varname*`)` is required and identifies the alternatives. *varname* must be numeric.

`for(`*varname*`)` specifies the variable for which the elasticities/marginal effects are calculated. The default is the first variable specified in `mixlogit`'s `rand` option.

`percentchange(#)` allows the user to compute arc elasticities/marginal effects for an attribute change of # per cent. This option cannot be combined with `absolutechange` and `dummy`.

`absolutechange(#)` allows the user to compute arc elasticities/marginal effects for an absolute attribute change of #. This option cannot be combined with `percentchange` and `dummy`.

`dummy` allows the user to compute arc elasticities/marginal effects for dummy variables. This option cannot be combined with `percentchange` and `absolutechange`.

`marginaleffect` allows the user to obtain marginal effects. The default is elasticities.

`weighted` specifies that probability-weighted sample averages are computed. The default is unweighted sample averages.

`nosd` prevents `mixlelast` from computing and displaying standard deviations.

`nrep(#)` specifies the number of Halton draws used for simulation. The default is the number specified for `mixlogit`, which has a default of 50.

`burn(#)` specifies the number of initial sequence elements to drop when creating the Halton sequences. The default is the number specified for `mixlogit`, which has a default of 15.

`hettype()` allows the user to specify the type of aggregation when choice sets are heterogeneous. The default is I for unweighted and IIa for probability-weighted sample averages. `hettype()` cannot be specified if choice sets are homogeneous.

`krobb(#)` allows the user to obtain Krinsky-Robb-standard errors or confidence intervals using # draws.

`krse` specifies that standard errors are computed. The default is confidence intervals.

`krlevel(#)` specifies the confidence level. The default is 95 per cent.

`krburn(#)` specifies the number of initial sequence elements to drop when creating the Halton sequences for the Krinsky-Robb procedure.

`kruserdraws` allows the user to provide draws for the Krinsky-Robb parametric bootstrap in a Mata matrix `mixl_userdraws`. The number of rows of the matrix `mixl_userdraws` must equal the number of coefficients in the model, and the number of columns must equal the number of choice occasions times the number of repetitions.[15] If `userdraws` is not specified, `mixlelast` will use Halton draws if the number of coefficients is less or equal to 10 and standard normal draws if it exceeds 10.[16]

`quietly` suppresses the output.

4.8 Examples

This section provides several illustrative examples of `mixlelast` both for homogeneous and heterogeneous choice sets.

[15]Here, the number of coefficients is the total number of coefficients estimated in `mixlogit`, i.e. the variable(s) specified in the `random()` option count twice as the mean and the standard deviation are two separate coefficients.

[16]For low dimensions, Halton sequences should be used. With dimensions above 10, pseudo-random points outperform Halton draws, see Drukker and Gates (2006). More advanced procedures, such as scrambled Halton sequences (Kolenikov 2012), can be implemented manually via the `kruserdraws` option.

4.8.1 Homogeneous Choice Sets

To illustrate `mixlelast` with homogeneous, I use an artificial data set of 100 individuals making 10 repeated purchasing decisions.

In the example, three attributes enter the choice model:

- price
- quality
- brand dummy

Following the notation of Hole (2007), `gid` identifies the alternatives in a choice occasion, and `pid` identifies the choice occasion faced by a decision-maker. The dependent variable `y` is 1 for the chosen alternative and 0 otherwise. The first 10 observations are listed below:

```
. use example_hom.dta
. list in 1/10,sepby(gid)
```

	alt	y	price	quality	brand	gid	pid
1.	Prod 1	0	1.308418	3	Brand A	1	1
2.	Prod 2	0	2.357423	5	Brand A	1	1
3.	Prod 3	0	3.380476	8	Brand A	1	1
4.	Prod 4	1	4.414344	11	Brand B	1	1
5.	Prod 5	0	8.617756	22	Brand B	1	1
6.	Prod 1	0	.8786653	3	Brand A	2	1
7.	Prod 2	0	1.99105	5	Brand A	2	1
8.	Prod 3	0	3.536737	8	Brand A	2	1
9.	Prod 4	0	4.295895	11	Brand B	2	1
10.	Prod 5	1	8.679978	22	Brand B	2	1

We start by fitting a mixed logit model. The coefficient on price is assumed to be normally distributed; those on quality and the brand dummy are fixed.

```
. mixlogit y brand quality, rand(price) group(gid) id(pid)
Iteration 0:   log likelihood = -1386.9157
  (output omitted)
Iteration 4:   log likelihood = -1218.6785
Mixed logit model                               Number of obs     =      5,000
                                                LR chi2(1)        =     635.40
Log likelihood = -1218.6785                     Prob > chi2       =     0.0000
```

y	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
Mean						
brand	.8389558	.1664776	5.04	0.000	.5126657	1.165246
quality	.7326987	.0890225	8.23	0.000	.5582178	.9071797
price	-1.766318	.2272046	-7.77	0.000	-2.21163	-1.321005
SD						
price	.5507636	.0545198	10.10	0.000	.4439067	.6576205

```
The sign of the estimated standard deviations is irrelevant: interpret them as
being positive
```

The SD coefficient is significant, indicating that preference heterogeneity is present.

We now compute unweighted sample point price elasticities.

```
. mixlelast, alternatives(alt)
Mixed logit sample elasticities
```

		Prod 1	Prod 2	Prod 3	Prod 4	Prod 5
Prod 1						
	Mean	-1.583582	1.20702	1.234225	.5806123	1.160138
	SD	.5321263	.35009	.4233442	.2133211	.2015333
Prod 2						
	Mean	.8620516	-2.993063	1.256074	.6522315	1.689042
	SD	.1310344	.6221173	.3964381	.226443	.2854372
Prod 3						
	Mean	.6431518	.9099626	-4.794224	.7198147	2.664615
	SD	.1075382	.2245667	.6208259	.2306577	.4073909
Prod 4						
	Mean	.4741855	.7339106	1.113647	-6.85574	3.730741
	SD	.0869767	.1709876	.2885865	.470882	.4400616
Prod 5						
	Mean	.1097656	.2170911	.4650955	.4201686	-3.954723
	SD	.0208497	.0434097	.0906856	.1006895	.1834887

```
Calculated for a marginal change in price
```

The output table displays the mean direct and cross elasticities and their standard deviations. The elements on the diagonal are the direct elasticities, which are all negative. Given a one per cent increase in price, the decrease in the choice probability ranges from 1.584 to 6.856 per cent. The off-diagonal elements are the cross elasticities, where each entry describes the percentage change in the row alternative given a one per cent change in an attribute of the column alternative. For example, the choice probability of Prod 1 increases by 1.207 per cent when the price of Prod

2 rises by 1 per cent. All cross-price elasticities are positive, indicating that the demand for rival alternatives rises when the price of one alternative is increased.

Alternatively, we could compute probability-weighted sample averages. This time, we use the `nosd` option that omits standard deviations.

```
. mixlelast, alternatives(alt) weighted nosd
Mixed logit probability weighted sample elasticities
             |    Prod 1     Prod 2     Prod 3     Prod 4     Prod 5
-------------+-------------------------------------------------------
      Prod 1 | -1.437372   1.129342   1.178512   .5706147   1.147615
      Prod 2 |  .8389906  -2.795044   1.216929   .6372329    1.64468
      Prod 3 |  .6416162   .8983047  -4.591341   .7086891   2.560194
      Prod 4 |  .4834713   .7372868   1.113011  -6.706311   3.620954
      Prod 5 |  .1108344   .2183731   .4632295   .4174095  -3.941572
Calculated for a marginal change in price
```

Instead of calculating price elasticities, we could also consider a change in the quality variable. As price was the first (and only) variable set to be random in `mixlogit`, we did not have to specify it in `mixlelast`. To obtain elasticities w.r.t. quality, we need to use the `elastvar` option.

Quality is an integer variable, so computing a marginal change does not make much sense. `mixlelast` will still produce results, but it detects that quality is an integer variable and issues a warning. More meaningful, however, is to look at a discrete change in the quality variable, e.g. by one.

```
. mixlelast, alternatives(alt) for(quality) absolutechange(1)
Mixed logit sample elasticities
             |    Prod 1     Prod 2     Prod 3     Prod 4     Prod 5
-------------+-------------------------------------------------------
Prod 1       |
        Mean |  1.388182  -1.396422  -1.524345  -.7976167  -1.520772
          SD |  .2607879   .4197096   .5221947   .2983788   .2449431
-------------+-------------------------------------------------------
Prod 2       |
        Mean | -1.005672   2.624413  -1.615071   -.929526  -2.236142
          SD |  .2040373    .388189    .513181   .3318357   .3566263
-------------+-------------------------------------------------------
Prod 3       |
        Mean | -.8210379   -1.18665   4.448009  -1.079768   -3.60011
          SD |  .1746568   .3130168    .459669   .3588448    .523858
-------------+-------------------------------------------------------
Prod 4       |
        Mean | -.6550803  -1.018255  -1.592809   7.037799  -5.179757
          SD |  .1483684   .2581882   .4276922   .3251452   .5807602
-------------+-------------------------------------------------------
Prod 5       |
        Mean | -.1922135   -.367616  -.7905778  -.7807326   4.480929
          SD |  .0462874   .0848141   .1714049   .2014827   .2330713
Calculated for an absolute change of 1 in quality
```

Contrary to price elasticities, we find positive direct and negative cross elasticities for quality. This is due to the positive sign on the quality coefficient and makes sense intuitively as an increase in quality makes a product more attractive.

Finally, we could also look at the change in the brand variable, i.e. evaluate the counterfactual situation where a product is sold under the rival's brand. In this case, we would use the `dummy` option to simulate hypothetical changes of the dummy variable *brand* from zero to one and vice versa.

To obtain incremental effects instead of elasticities, we use the `marginaleffect` option.

```
. mixlelast, alternatives(alt) marginal
Mixed logit sample marginal effects
```

		Prod 1	Prod 2	Prod 3	Prod 4	Prod 5
Prod 1						
	Mean	-.2878743	.1365916	.0887652	.0314141	.0311034
	SD	.0313022	.0341118	.0318789	.0144679	.0097729
Prod 2						
	Mean	.1365916	-.2766283	.0749002	.0286624	.0364741
	SD	.0341118	.0560073	.0303258	.0131542	.0110592
Prod 3						
	Mean	.0887652	.0749002	-.2383686	.0268862	.0478171
	SD	.0318789	.0303258	.0612841	.0126498	.0115994
Prod 4						
	Mean	.0314141	.0286624	.0268862	-.1184339	.0314712
	SD	.0144679	.0131542	.0126498	.0382954	.0084627
Prod 5						
	Mean	.0311034	.0364741	.0478171	.0314712	-.1468658
	SD	.0097729	.0110592	.0115994	.0084627	.0136522

```
Calculated for a marginal change in price
```

Here, the columns (and rows) add up to zero, as the decrease in the choice probability of one alternative must necessarily be compensated by an increase in the sum of probabilities of choosing rival alternatives.

Finally, we look at a very small discrete change of 0.001 per cent and recompute the marginal effects from above using the arc method.

```
. mixlelast, alternatives(alt) marginal percentchange(0.001)
Mixed logit sample incremental effects
```

		Prod 1	Prod 2	Prod 3	Prod 4	Prod 5
Prod 1						
	Mean	-.2878735	.1365904	.0887634	.031413	.0311024
	SD	.0313027	.0341119	.0318786	.0144675	.0097726
Prod 2						
	Mean	.1365913	-.2766259	.0748988	.0286615	.0364732
	SD	.0341118	.0560079	.0303255	.0131538	.0110589
Prod 3						
	Mean	.0887649	.0748996	-.2383644	.0268853	.0478164
	SD	.0318789	.0303257	.0612842	.0126495	.0115992
Prod 4						
	Mean	.031414	.0286622	.0268857	-.1184302	.0314711
	SD	.0144679	.0131541	.0126497	.0382946	.0084626
Prod 5						
	Mean	.0311032	.0364737	.0478164	.0314704	-.1468632
	SD	.0097729	.0110592	.0115994	.0084626	.0136522

```
Calculated for a change of .001 per cent in price
```

As expected, the results for sample averages and their standard deviations are practically identical. Nonetheless, users are advised to only use the arc method when necessary, as the computation is slower. The difference in computation time is increasing in the number of alternatives and the number of choice occasions.

4.8.2 Heterogeneous Choice Sets

In the second data set, decision-makers live in one of three locations that differ according to the set of products available. Individuals 1 to 50 live in location 1, where all five products of both brands A and B are available. In location 2, where individuals 51 to 75 live, only products of brand A are available, i.e. products 1, 2 and 3. Finally, only products 4 and 5 of brand B are in the choice sets of decision-makers 75 to 100 living in location 3. Each decision-maker chooses ten times among the available alternatives.

A representative choice set for each location is displayed below:

```
. use example_het.dta
. list if gid == 1 | gid == 501 | gid == 751, sepby(gid)
```

	alt	y	price	quality	brand	gid	pid	location
1.	Prod 1	0	1.308418	3	Brand A	1	1	1
2.	Prod 2	0	2.357423	5	Brand A	1	1	1
3.	Prod 3	0	3.380476	8	Brand A	1	1	1
4.	Prod 4	1	4.414344	11	Brand B	1	1	1
5.	Prod 5	0	8.617756	22	Brand B	1	1	1
2501.	Prod 1	0	1.356619	3	Brand A	501	51	2
2502.	Prod 2	1	2.180957	5	Brand A	501	51	2
2503.	Prod 3	0	3.14988	8	Brand A	501	51	2
3251.	Prod 4	1	4.398738	11	Brand B	751	76	3
3252.	Prod 5	0	8.765503	22	Brand B	751	76	3

As before, we estimate the mixed logit model.

```
. mixlogit y brand quality, rand(price) group(gid) id(pid)
Iteration 0:   log likelihood = -1094.6025
  (output omitted)
Iteration 5:   log likelihood = -930.11945
Mixed logit model                               Number of obs     =      3,750
                                                LR chi2(1)        =     559.92
Log likelihood = -930.11945                     Prob > chi2       =     0.0000
```

y	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
Mean						
brand	.9619626	.2485758	3.87	0.000	.474763	1.449162
quality	.8084889	.1005983	8.04	0.000	.6113199	1.005658
price	-1.926808	.2566438	-7.51	0.000	-2.429821	-1.423795
SD						
price	.6458248	.0680799	9.49	0.000	.5123906	.779259

```
The sign of the estimated standard deviations is irrelevant: interpret them as
being positive
```

First, we compute the unweighted sample elasticities of type I, i.e. over the entire sample of 100 decision-makers.

```
. mixlelast, alternatives(alt) hettype(I) nosd
Mixed logit sample elasticities (type I)
```

	Prod 1	Prod 2	Prod 3	Prod 4	Prod 5
Prod 1	-1.249679	1.027456	1.100893	.2595871	.5574325
Prod 2	.7087478	-2.334087	1.184107	.3049821	.8680116
Prod 3	.4862719	.7379635	-3.355797	.3496448	1.460553
Prod 4	.2535825	.3996399	.6194909	-4.395615	3.301949
Prod 5	.0487804	.1000686	.2244125	.6515905	-2.76727

```
Calculated for a marginal change in price
```

Now, we restrict our attention to individuals with the product i, i.e., the row product in the table below, in their choice set (type IIa).

```
. mixlelast, alternatives(alt) hettype(IIa) nosd
Mixed logit sample elasticities (type IIa)
```

	Prod 1	Prod 2	Prod 3	Prod 4	Prod 5
Prod 1	-1.666239	1.369941	1.467857	.3461161	.7432433
Prod 2	.9449971	-3.112116	1.57881	.4066428	1.157349
Prod 3	.6483625	.9839513	-4.474397	.4661931	1.947404
Prod 4	.33811	.5328532	.8259879	-5.86082	4.402598
Prod 5	.0650405	.1334248	.2992166	.8687874	-3.689693

```
Calculated for a marginal change in price
```

Comparing both tables shows that direct and cross elasticities have increased in absolute values. This result is not surprising. For type I, we consider all individuals, including those with zero individual effects. For IIa, we ignore all individuals who could not choose the relevant product. For products 1, 2 and 3, we now look at decision-makers in location 1 and location 2 only. Likewise, only decision-makers in location 1 and location 3 are considered for products 4 and 5. That means, we no longer consider individuals with zero individual direct effects, so direct sample elasticities must rise in absolute value compared to type IIb.

For cross elasticities, we also reduce the number of decision-makers with zero individual cross effects by excluding those who could not choose product i. Note, however, that we still consider those instances of zero individual cross effects where product i is available, but product j, i.e. the column product, is not.

Finally, aggregation of type IIb considers only those individuals who are actually affected by an attribute change, i.e. both alternatives i and j are present in their choice sets. These are only the decision-makers living in location 3.

```
. mixlelast, alternatives(alt) hettype(IIb) nosd
Mixed logit sample elasticities (type IIb)
```

	Prod 1	Prod 2	Prod 3	Prod 4	Prod 5
Prod 1	-1.666239	1.369941	1.467857	.5191742	1.114865
Prod 2	.9449971	-3.112116	1.57881	.6099641	1.736023
Prod 3	.6483625	.9839513	-4.474397	.6992897	2.921106
Prod 4	.507165	.7992798	1.238982	-5.86082	4.402598
Prod 5	.0975608	.2001371	.4488249	.8687874	-3.689693

```
Calculated for a marginal change in price
```

Comparing the results for type IIa and type IIb, we make three observations. First, direct elasticities are identical. This is because the subset of individual direct effects taken into consideration does not differ between type IIa and type IIb. Second, cross elasticities between alternatives of the same brand also remain unchanged. Recalling that the heterogeneity was on the brand level, there are no cases in which one product of a certain brand is available and another product of the same brand is not. Hence, regarding substitution patterns between alternatives within the same brand, there is no difference between type IIa and type IIb either. Third, for cross effects across brands, only those decision-makers who have both products available are considered. All individuals with zero cross effects are now excluded, which is why the sample effects must increase in absolute value.

4.8.3 Standard Errors and Confidence Intervals

Finally, we want to compute confidence intervals and standard errors using the Krinsky-Robb parametric bootstrap. For demonstration, we use the same data set as in the first example (see subsection 4.8.1).

We now compute Krinsky-Robb standard errors with 100 replications using the `krse` option.

```
. mixlelast, alternatives(alt) kr(100) krse
Mixed logit sample elasticities
              |    Prod 1     Prod 2     Prod 3     Prod 4     Prod 5
--------------+-------------------------------------------------------
Prod 1        |
         Mean | -1.588732   1.209058   1.235619   .5851914   1.168598
           SE |  .1741421   .1293064     .14276   .0947962   .2206916
--------------+-------------------------------------------------------
Prod 2        |
         Mean |  .8656952  -3.006101   1.258633    .657472   1.695867
           SE |  .1059331   .3317349   .1474431   .1033878    .283632
--------------+-------------------------------------------------------
Prod 3        |
         Mean |  .6456213   .9128968  -4.814247   .7260494   2.669903
           SE |   .076321   .1053874   .5710928   .1146381   .4180427
--------------+-------------------------------------------------------
Prod 4        |
         Mean |  .4762443   .7370268   1.118481  -6.881942    3.73549
           SE |  .0566474   .0891121   .1443512   .8692364   .5892882
--------------+-------------------------------------------------------
Prod 5        |
         Mean |  .1115721   .2204908   .4716262   .4269033  -4.017563
           SE |  .0207407   .0412111   .0887768   .0860485   .7422291
Calculated for a marginal change in price
Means and standard errors by Krinsky-Robb parametric bootstrap with
100 repetitions
```

Alternatively, we could compute confidence intervals. The default is a 95 per cent confidence interval. To set the confidence level, e.g. to 99 per cent, we use the `krlevel` option.

```
. mixlelast, alternatives(alt) kr(100) krlevel(99)
Mixed logit sample elasticities
             |    Prod 1     Prod 2     Prod 3     Prod 4     Prod 5
-------------+-------------------------------------------------------
Prod 1       |
        Mean | -1.588732   1.209058   1.235619   .5851914   1.168598
    CI_lower | -2.036889    .885817   .9136785    .390892   .7357463
    CI_upper | -1.162032   1.540968   1.617532   .8230219   1.778616
-------------+-------------------------------------------------------
Prod 2       |
        Mean |  .8656952  -3.006101   1.258633    .657472   1.695867
    CI_lower |   .594959  -3.751526   .9098775   .4391683   1.131964
    CI_upper |  1.192836  -2.167427   1.635249   .9178194   2.445996
-------------+-------------------------------------------------------
Prod 3       |
        Mean |  .6456213   .9128968  -4.814247   .7260494   2.669903
    CI_lower |  .4488157   .6570522  -6.078871   .4858516   1.739164
    CI_upper |  .8719096   1.181786  -3.407677   1.006442   3.725657
-------------+-------------------------------------------------------
Prod 4       |
        Mean |  .4762443   .7370268   1.118481  -6.881942    3.73549
    CI_lower |  .3345948   .5277493   .7775908   -8.96048   2.333426
    CI_upper |  .6314298   .9583172   1.492383  -4.728254   5.122632
-------------+-------------------------------------------------------
Prod 5       |
        Mean |  .1115721   .2204908   .4716262   .4269033  -4.017563
    CI_lower |  .0712387    .141619   .2904997   .2830188  -6.543079
    CI_upper |  .1802272   .3584905   .7616104   .6880237  -2.611262
Calculated for a marginal change in price
Means and 99% confidence intervals by Krinsky-Robb parametric bootstrap
with 100 repetitions
```

As we can see, none of the confidence intervals contains zero, i.e. all direct- and cross-elasticities are statistically significantly different from zero at the 1 per cent level. However, to obtain more robust results, we should use a much higher number of draws.

4.9 Saved Results

`mixlogit` stores the following results to r():

Scalars

`r(N)`	number of observations	`r(burn)`	initial elements to drop when creating Halton sequences
`r(N_group)`	number of choice occasions	`r(nrep)`	number of Halton draws
`r(N_id)`	number of decision-makers	`r(krobb)`	number of Krinsky-Robb repetitions
`r(N_alt)`	number of alternatives	`r(krlevel)`	Krinsky-Robb confidence level
`r(het)`	1 if heterogeneous choice sets, 0 otherwise	`r(krse)`	1 if Krinsky-Robb standard erros, 0 if confidence intervals
`r(marginal)`	1 if marginal effects, 0 otherwise	`r(krburn)`	initial elements to drop when Halton sequence for `krobb`
`r(weighted)`	1 if probability weighted effects, 0 otherwise	`r(kruser)`	1 if user-provided random numbers, 0 otherwise
`r(nosd)`	1 if no standard deviations, 0 otherwise		

Macros

`r(cmd)`	`mixlelast`	`r(title)`	title in output table
`r(subtitle)`	subtitle in output table	`r(altid)`	name of altid() variable
`r(subtitle2)`	subtitle of output for `krobb`	`r(hettype)`	type of aggregation for heterogeneous choice sets
`r(elastvar)`	name of elastvar() variable	`r(group)`	name of group() variable
`r(id)`	name of id() variable	`r(method)`	type of calculation method

Matrices

`r(mean)`	sample elasticities/marginal effects	`r(sd)`	standard deviations
`r(KRSE)`	Krinsky-Robb standard errors	`r(KR_lower)`	Krinsky-Robb lower confidence bounds
`r(KR_upper)`	Krinsky-Robb upper confidence bounds		

Acknowledgements

`mixlelast` builds on `mixlogit`'s `mixlpred` option for calculating predicted probabilities written byHole (2007).
The idea for `mixlelast` was born during a joint project with Anna Lu, whose contribution to `mixlelast` is highly appreciated. Also, I thank Arne Risa Hole for his feedback and helpful suggestions.

References

Cameron, A. C. and Trivedi, P. K. (2010). *Microeconometrics Using Stata*. Stata Press.

Drukker, D. M. and Gates, R. (2006). Generating halton sequences using mata. *Stata Journal*, 6(2):214–228.

Greene, W. H. (2008). *Econometric Analysis*. Prentice Hall.

Hensher, D. A., Rose, J. M., and Greene, W. H. (2015). *Applied Choice Analysis*. Cambridge University Press.

Hole, A. R. (2007). Estimating mixed logit models using maximum simulated likelihood. *Stata Journal*, 7(3):388–401.

Kolenikov, S. (2012). Scrambled halton sequences in mata. *Stata Journal*, 12(1):29–44.

Krinsky, I. and Robb, A. L. (1986). On approximating the statistical properties of elasticities. *The Review of Economics and Statistics*, 68(4):715–719.

Krinsky, I. and Robb, A. L. (1990). On approximating the statistical properties of elasticities: a correction. *The Review of Economics and Statistics*, 72(1):189–90.

Louviere, J. J., Hensher, D. A., and Swait, J. (2000). *Stated Choice Methods: Analysis and Applications in Marketing, Transportation and Environmental Evaluation*. Cambridge University Press.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In Zarembka, P., editor, *Frontiers in Economics*, pages 105–142. Academic Press.

Revelt, D. and Train, K. (1998). Mixed logit with repeated choices: Households' choices of appliance efficiency level. *The Review of Economics and Statistics*, 80(4):647–657.

Train, K. (2009). *Discrete choice methods with simulation.* Cambridge University Press.

Zeigermann, L. (2024). MIXLELAST: Stata module to compute mixed logit sample elasticities and marginal effects. Statistical Software Components, Boston College Department of Economics.

Chapter 5

Opencagegeo: Stata Module for Geocoding

5.1 Introduction

Geocoded data is extensively used by researchers and practitioners in numerous fields, and geocoding (and reverse geocoding) has become a common exercise.[1] Geocoding is the process of converting an address into geographic coordinates. Reverse geocoding uses geographic coordinates to retrieve the postal address.

To facilitate geocoding for Stata users, several user-written geocoding routines have been made available. All have in common that they use application programming interfaces or APIs to access online geocoding services. These routines are (the outdated) `geocode` (Ozimek and Miles 2011), its (no longer available) successor `geocode3` (Bernhard 2013) and the immediate command `gcode` (Ansari 2015) using the Google Maps API, `geocodeopen` (Anderson 2013) using MapQuest's API and `geocodehere` (Hess 2015) using the API for Here Maps. However, the terms of use of Google Maps, MapQuest, and Here Maps are very restrictive and prohibit data storage, which render them unusable for many purposes.

This chapter describes *opencagegeo*, an alternative geocoding command for Stata that uses the geocoding API of OpenCage Data which has very flexible terms of use.[2] All data is jointly licensed under the ODbL and CC-BY-SA licenses[3] and OpenCage Data does not restrict the use of geocodes.[4]

The syntax of `opencagegeo` is described in Section 5.2. Its immediate version `opencagegeoi` is presented in Section 5.3. Two examples for illustration are given in Section 5.4.

5.2 Opencagegeo

5.2.1 Syntax

`opencagegeo` [*if*] [*in*] [, `key(`*string*`)` `number(`*varname*`)` `street(`*varname*`)`

`postcode(`*varname*`)` `city(`*varname*`)` `county(`*varname*`)` `state(`*varname*`)`

`country(`*varname*`)` `fulladdress(`*varname*`)` `latitude(`*varname*`)` `longitude(`*varname*`)`

---

[1] For some interesting examples in the field of economics, see CESifo (2018).

[2] `opencagegeo` is published in the Boston College Statistical Software Components (SSC) archive, see Zeigermann (2016).

[3] The ODbL and CC-BY-SA licences are available at `http://opendatacommons.org/licenses/odbl/summary/` and `https://creativecommons.org/licenses/by-sa/2.0/`.

[4] OpenCage Data's terms of use are available at `http://geocoder.opencagedata.com/faq.html#legal`.

coordinates(*varname*) countrycode(*varname* or *string*) language(*varname* or *string*) replace resume]

5.2.2 Options

General

`key` (*string*) is required and takes the OpenCage Data API key as input.

`countrycode` (*varname* `or` *string*) specifies the country code. `countrycode()` allows either a string variable or a string as input.

`language` (*varname* `or` *string*) specifies the language in which the results are returned. `language()` allows either a string variable or a string as input. The default is *en* for English.

`resume` specifies that the process is continued after the rate limit was exceeded the day before. `resume` may not be combined with `replace`.

Geocoding

`number` (*varname*) specifies the variable containing the house number.

`street` (*varname*) specifies the variable containing the street. *varname* must be string.

`postcode` (*varname*) specifies the variable containing the post-code.

`city` (*varname*) specifies the variable containing the city, town or village. *varname* must be string.

`county` (*varname*) specifies the variable containing the county. *varname* must be string.

`state` (*varname*) specifies the variable containing the state. *varname* must be string.

`country` (*varname*) specifies the variable containing the country. *varname* must be string.

`fulladdress` (*varname*) specifies a single variable containing some or all of the above options. *varname* must be string. `fulladdress()` may not be combined with any of the options above.

Reverse Geocoding

`latitude(`*varname*`)` specifies the variable containing the latitude. Values must lie between -90 and 90.

`longitude(`*varname*`)` specifies the variable containing the longitude. Values must lie between -180 and 180.

`coordinates(`*varname*`)` specifies the variable containing latitude longitude pairs. Latitudes must lie between -90 and 90 and must be stated first. Longitudes must lie between -180 and 180. Both values must be separated by a comma. *varname* must be string. `coordinates()` may not be combined with `latitude()` and `longitude()`.

5.2.3 Remarks

General

`opencagegeo` requires an OpenCage Data API key, which can be obtained by signing up at `https://geocoder.opencagedata.com/users/sign_up`. The user can choose among several customer plans with different daily rate limits. The free trial plan allows 2.500 requests per day. If the rate limit is hit, `opencagegeo` will issue an error message and exit. To continue the task the following day, the user may add the `resume` option to the original specification. `opencagegeo` will automatically detect which observations are still to be geocoded. The user is strongly advised not to make any changes to the data until geocoding of all observations is completed.[5] A new day begins at 00:00:00 Coordinated Universal Time (UTC).

If the output variables created by `opencagegeo` already exist, the user can specify `replace` to overwrite existing observations.[6] If either `in` or `if` is specified, only the selected observations will be replaced. `resume` and `replace` may not be combined.

The `countrycode()` option allows the user to specify the country of the location. Providing a country code will restrict the results to the given country. If all locations lie in the same country, the user may enter the country code directly into `countrycode()`. Alternatively, if the locations are in more than just one country, a variable containing the respective country codes must be specified. The coun-

[5] The `resume` option evaluates the output variable `g_quality` to select observations which have not yet been geocoded.

[6] The variables created by `opencagegeo` are described in subsection 5.2.4.

try code is a two-letter code defined by the ISO 3166-1 Alpha 2 standard.[7] If any country code is invalid, `opencagegeo` will issue an error message and exit.

The `language()` option allows the user to specify the language in which the results are returned. As with `countrycode()`, either a string or a string variable can be specified. The language must be entered in IETF format language code.[8] The default is English. If the language is set to *native*, the results will be returned in the native language of the location, provided the underlying OpenStreetMap (OSM) data is available in that language. Users of Stata 13 or earlier versions are advised to carefully use the `language()` option as many languages other than English contain special characters that cannot be displayed correctly.

`opencagegeo` requires two user-written Stata libraries, `insheetjson` and `libjson` (Lindsley 2012a,b), which are available at the Statistical Software Components (SSC) archive.

Geocoding

Generally, there are two different ways of feeding addresses into `opencagegeo`. First, by components using the `street()`, `number()`, `postcode()`, `city()`, `county()`, `state()` and/or `country()` options. This is recommended if the location address data is provided in separate variables. Not all options need to be specified at the same time; they can be combined in any meaningful way. E.g. to obtain geocodes of cities, you may use `city()`, `state()` and `country()`.

Second, if the address is contained in a single string variable, the `fulladdress()` option must be used. The address should follow country-specific conventions, although the OpenCage Geocoder allows some flexibility regarding address formats. E.g. the two following well-formatted examples should give the same results:

"number street, city post-code, country"

"street number, post-code city, country"

Generally, the OpenCage Geocoder is not case-sensitive and can deal with the most commonly used abbreviations. Running `opencagegeo` in Stata 14 (or newer) allows the address variables to be in Unicode (UTF-8), i.e. address names may contain accented characters, symbols and non-Latin characters. For older releases, the

[7] A comprehensive list of all ISO 3166-1 Alpha 2 codes is available at `https://en.wikipedia.org/wiki/ISO_3166-1_alpha-2`.

[8] A comprehensive list of all IETF language codes is available at `http://www.iana.org/assignments/language-subtag-registry/language-subtag-registry`.

input variables must not contain any special characters.[9] Otherwise, `opencagegeo` will issue an error message and exit. `opencagegeo` is sensitive to spelling mistakes and will return empty strings for misspelt addresses.

Reverse geocoding

As with geocoding, two options are available for reverse geocoding. If longitudes and latitudes are contained in two separate variables, the `latitude()` and `longitude()` options should be used. Alternatively, latitude-longitude pairs contained in one variable may be fed into the `coordinates()` option. Latitudes must be stated first, and both values must be separated by a comma.

Latitudes must be between -90 and 90, and longitudes must be between -180 and 180. Otherwise, `opencagegeo` will issue an error message and exit.

5.2.4 Output Variables

Running `opencagegeo` generates a set of 12 variables, all having `g_` prefixes.[10]

`g_latitude` and `g_longitude` contain the latitudes and longitudes retrieved. The variables `g_number`, `g_street`, `g_postcode`, `g_city`, `g_county`, `g_state` and `g_country` contain the respective information returned. If the information is missing (e.g. the house number is not known) or not requested (`number()` was not specified, or no information on the house number was contained in address variable fed into the `fulladdress()` option), an empty string for the respective variable will be returned. `g_formatted` contains a well-formatted place name generated by the OpenCage Geocoder. In addition to the postal address, it might also have additional information on the name of the building, institution, shop, etc., at that location.

Finally, `g_confidence` and `g_quality` are generated. The former provides a measure of precision of the match generated by the OpenCage Geocoder. The confidence value is calculated as the distance in kilometres between the south-east and the north-west corners of the bounding box.

[9] Only ASCII printable characters, i.e. character codes 32-127, may be used in the input variables.

[10] The JSON paths used to extract the results returned from the OpenCage Geocoder are given in Table A5.1 in the appendix.

Confidence levels are defined as follows:

$$\begin{aligned}
0 &= \text{unable to determine bounding box}\\
1 &= 25 \text{ km or more}\\
2 &= \text{less than } 25 \text{ km}\\
3 &= \text{less than } 20 \text{ km}\\
4 &= \text{less than } 15 \text{ km}\\
5 &= \text{less than } 10 \text{ km}\\
6 &= \text{less than } 7.5 \text{ km}\\
7 &= \text{less than } 5 \text{ km}\\
8 &= \text{less than } 1 \text{ km}\\
9 &= \text{less than } 0.5 \text{ km}\\
10 &= \text{less than } 0.25 \text{ km}
\end{aligned}$$

The variable `g_quality` contains the accuracy level of the returned results and is defined as follows:

$$\begin{aligned}
0 &= \text{location not found}\\
1 &= \text{country}\\
2 &= \text{state}\\
3 &= \text{county}\\
4 &= \text{city}\\
5 &= \text{postcode}\\
6 &= \text{street}\\
7 &= \text{number}
\end{aligned}$$

If, for example, the user provides the full address, including street name and house number, but the Opencage geocoder can only identify the postcode, the quality level will be 5 (*postcode*). The highest quality level to be reached is determined by the inputs; e.g. if the `number()` is not specified or no information on the house number is contained in the variable fed into `fulladdress()`, the highest quality level to be achieved is 6 (*street*).

5.3 Opencagegeoi

5.3.1 Syntax

```
opencagegeoi #location
```

5.3.2 Remarks

The immediate command `opencagegeoi` takes its inputs from what is typed as arguments rather than from data stored in the memory. Instead of creating output variables, it immediately displays the results in the output window.

`opencagegeoi` requires the API key to be stored in a global macro *mykey*. The location to be geocoded is then directly entered after `opencagegeoi`. For geocoding an address, the input should be well-formatted as described above. For reverse geocoding, `opencagegeoi` accepts a latitude-longitude pair as input. The latitude is entered first, followed by the longitude. Both values must be separated by a comma.

The user may define a global macro *language* containing the language code in which the result shall be returned. The default is English. Setting the global macro *language* to *native*, the results are returned in the native language of the location. Users of Stata 13 or older should be aware that special characters cannot be displayed correctly.

5.3.3 Saved Results

In addition to displaying the well-formatted address returned from OpenCage and the latitude and longitude values, `opencagegeoi` saves the following results to `r()`.

Macros

`r(input)`	address as typed by user	`r(formatted)`	well-formatted address returned from OpenCage Geocoder
`r(lat)`	latitude	`r(lon)`	longitude
`r(conf)`	confidence level of result		

5.4 Examples

5.4.1 Opencagegeo

Geocoding

For example, consider the addresses of five Goethe Institutes in Europe, Mexico and the United States.[11] The data is displayed below.

```
. list

     +------------------------------------------------------------+
     |        CITY   POSTCODE            STREET   NUMBER   COUNTRY |
     |------------------------------------------------------------|
  1. |      London    SW7 2PH   Exhibition Road       50        UK |
  2. | Mexico City      06700            Tonala       43    Mexico |
  3. |    New York      10003     Irving Place       30       USA |
  4. |       Paris      75116     Avenue d´Iena       17    France |
  5. |        Rome      00198        Via Savoia       15     Italy |
     +------------------------------------------------------------+
```

To geocode the five addresses given above, specify `opencagegeo` as follows:

```
. opencagegeo, key(YOUR-KEY-HERE) street(STREET) number(NUMBER) postcode(POSTCODE)
> country(COUNTRY)

OpenCage geocoded 1 of 5
  (output omitted)
OpenCage geocoded 5 of 5

  g_quality |      Freq.     Percent        Cum.
------------+-----------------------------------
     street |          1       20.00       20.00
     number |          4       80.00      100.00
------------+-----------------------------------
      Total |          5      100.00

Data generated is jointly licensed under the ODbL and CC-BY-SA licenses.
```

The output table produced by `opencagegeo` illustrates that four out of five addresses were identified at the highest quality level, i.e. *number*. The exact house number was not found for one location, and the resulting quality level is *street*.

[11] The Goethe Institute is a not-for-profit German cultural organisation that promotes the German language; the addresses were obtained from `https://www.goethe.de/en/wwt.html`.

Reverse Geocoding

Now, the geocoding process is reversed to translate the latitude-longitude pairs generated above back into addresses. First, rename the variables and then specify `opencagegeo` using `latitude()`, `longitude()` and the `replace` option as follows:

```
. rename g_lat LATITUDE
. rename g_lon LONGITUDE
. opencagegeo, key(YOUR-KEY_HERE) latitude(LATITUDE) longitude(LONGITUDE) > replace

OpenCage geocoded 1 of 5
  (output omitted)
OpenCage geocoded 5 of 5
```

g_quality	Freq.	Percent	Cum.
street	1	20.00	20.00
number	4	80.00	100.00
Total	5	100.00	

```
Data generated is jointly licensed under the ODbL and CC-BY-SA licenses.
```

Not surprisingly, four latitude longitude pairs are identified at the *number* and one at the *street* level.

5.4.2 Opencagegeoi

Geocoding

If the user wants to geocode a single location rather than a data set containing various geographic locations, the immediate version `opencagegeoi` should be used. Note, however, that it is necessary to define a global macro *mykey* containing the OpenCage key first.[12]

```
. global mykey YOUR-KEY-HERE
```

To obtain the geographic coordinates of e.g. the Goethe Institute London, specify `opencagegeoi` as follows:

[12] The global macro *mykey* needs to be specified only once when a new Stata window is opened.

```
. opencagegeoi 50 Exhibition Road, London SW7 2PH, UK

*********************************
*** OpenCage Geocoder Results ***
*********************************
Formatted address: Goethe Institute, 50-51 Exhibition Road, London SW7 1BF, UK
Latitude: 51.4994811
Longitude: -0.174013268370617
```

Reverse Geocoding

To retrieve the postal address from geographic coordinates - in this example of the Goethe Institute in London - type:

```
. opencagegeoi 51.4994811,-0.174013268370617

*********************************
*** OpenCage Geocoder Results ***
*********************************
Formatted address: Goethe Institute, 50-51 Exhibition Road, London SW7 1BF, UK
Latitude: 51.4994811
Longitude: -0.174013268370617
```

Acknowledgements

I would like to thank Ed Freyfogle of OpenData for his support, and Achim Ahrens for testing `opencagegeo` and helping to improve it. The immediate version was added at Kit Baum's suggestion. Parts of `opencagegeo` build on `geocode` (Ozimek and Miles 2011).

References

Anderson, M. (2013). Geocodeopen: Stata module to geocode addresses using mapquest open geocoding services and open street maps. `https://ideas.repec.org/c/boc/bocode/s457733.html`.

Ansari, M. R. (2015). Gcode: Stata module to download google geocode data. `https://ideas.repec.org/c/boc/bocode/s457969.html`.

Bernhard, S. (2013). Geocode3: Stata module to retrieve coordinates or addresses from google geocoding api version 3.

CESifo (2018). Special issue: On the use of geo-coded data in economic research. *CESifo Economic Studies*, 64(2):123–338.

Hess, S. (2015). Geocodehere: Stata module to provide geocoding relying on nokia's here maps api. `https://ideas.repec.org/c/boc/bocode/s458048.html`.

Lindsley, E. (2012a). Insheetjson: Stata module for importing tabular data from json sources on the internet. `https://ideas.repec.org/c/boc/bocode/s457407.html`.

Lindsley, E. (2012b). Libjson: Stata module to provide mata class library for obtaining and parsing json strings into object trees. `https://ideas.repec.org/c/boc/bocode/s457406.html`.

Ozimek, A. and Miles, D. (2011). Stata utilities for geocoding and generating travel time and travel distance information. *The Stata Journal*, 11(1):pp. 106–119.

Zeigermann, L. (2016). OPENCAGEGEO: Stata module for geocoding. Statistical Software Components, Boston College Department of Economics.

Appendix

The OpenCage Geocoder API returns its results in JSON (JavaScript Object Notation) format. The respective information is then extracted and stored in the variables described above (or displayed immediately for `opencagegeoi`). As the underlying data (provided by OpenStreetMap (OSM) and other open data sources) is maintained by many contributors, keys are not uniquely defined. The key leading to e.g. the street name is not always *street*, but may be *street_name*, *road* etc. instead. Unfortunately, no exhaustive list of all these aliases exists. However, `opencagegeo` uses the most common ones to generate its output variables.[13] The so-called JSON paths used for extracting the data are listed in Table A5.1 below.

Variable	JSON path (aliases)
g_latitude	results:1:geometry:lat
g_longitude	results:1:geometry:lng
g_street	results:1:components:street (street_name, road, residential, footway, pedestrian)
g_number	results:1:components:house_number
g_postcode	results:1:components:postcode
g_city	results:1:components:city (town, village, hamlet)
g_county	results:1:components:county
g_state	results:1:components:state
g_country	results:1:components:country
g_formatted	results:1:components:formatted
g_confidence	results:1:components:confidence

Table A5.1: JSON paths for extracting Opencage Data geocoding results

[13] If none of the keys used by `opencagegeo` leads to the requested information, an empty string in the respective variable will be returned. However, the information will still be part of the well-formatted address in `g_formatted`.

Chapter 6

Lhaversine: Mata Library for Distance Functions

6.1 Introduction

This chapter describes `lhaversine`[1], a Mata library containing functions to compute distance matrices (Section 6.2) and matrices indicating if a locations within a specified distance from eachother (Section 6.3).[2]

All functions of `lhaversine` use the haversine formula that approximates geographic distances between two locations. It is a special case of a more general formula in spherical trigonometry, the law of haversines, that calculates the great-circle distance between two points from their latitudes and longitudes.[3]

Formally:

$$d = 2r \arcsin\left(\sqrt{\begin{array}{c}\sin^2\left(\frac{(\rho_2-\rho_1)\times dr}{2}\right) + \cos(\rho_1 \times dr)\times \\ \cos(\rho_2 \times dr) \times \sin^2\left(\frac{(\lambda_2-\lambda_1)\times dr}{2}\right)\end{array}}\right)$$

where

$$\begin{aligned} d &= \text{distance} \\ r &= \text{radius} \\ \rho_1 &= \text{latitude of point 1 in degrees} \\ \rho_2 &= \text{latitude of point 2 in degrees} \\ \lambda_1 &= \text{longitude of point 1 in degrees} \\ \lambda_2 &= \text{longitude of point 2 in degrees} \\ dr &= \pi/180 \end{aligned}$$

6.2 Mata Functions for Distance Matrices

6.2.1 Syntax

real matrix `hav_dist(`*real colvector lat_A*`,` *real colvector lng_A*`,` *real colvector lat_B*`,` *real colvector lng_B* [, *real scalar radius*]

[1]`lhaversine` is publised in the Boston Statistical Software Components (SSC) archive, see (Zeigermann 2025).

[2]`lhaversine` draws heavily from `geodist` and `geonear`, two user-written Stata module to compute geographical distances, see (Picard 2010a,b). Another related Stata module is `geocircles`, see (Picard 2015).

[3]For an illustrative derivation and explanation of the haversine formula, see van Brummelen (2013).

The syntax of `hav_distm()` is analogue.

6.2.2 Description

`hav_dist(`*lat_A*`,` *lng_A*`,` *lat_B*`,` *lng_B*`,` *radius*`)` returns a matrix of distances between each of the A locations specified by the vectors of latitudes (*lat_A*) and longitudes (*lng_A*) and all B locations specified by *lat_B* and *lng_B*. The results are in kilometres. The argument *radius* is optional and allows the user to set the earth radius. The default is 6,371 km.[4]

`hav_distm()` is analogue, but the results and *radius* are in miles. The default radius is 3,958.76 miles.

6.2.3 Remarks

The input vectors (*lat_A, lng_A, lat_B, lng_B*) must be in degrees and allow values between -90 and 90 for latitudes and values between -180 and 180 for longitudes.[5] The default radius is the mean earth radius.

6.2.4 Conformability

`hav_dist(`*lat_A*`,` *lng_A*`,` *lat_B*`,` *lng_B*`,` *radius*`)`

input:

lat_A: $n \times 1$

lng_A: $n \times 1$

lat_B: $k \times 1$

lng_B: $k \times 1$

radius: 1×1 (optional)

output:

result: $n \times k$

`hav_distm()` is analogue.

[4]The earth radius ranges from 6,357 km (3,950 miles) at the poles to 6,378 km (3,963 miles) at the equator (`https://en.wikipedia.org/wiki/Earth_radius`).

[5]Latitudes and longitudes in radians can be easily transformed to degrees by multiplying by $\frac{180}{\pi}$.

6.2.5 Examples

For demonstration, define four column vectors containing the latitudes and longitudes of two A locations and three B locations:

```
: lat_A = 53.1205306\53.2056589
: lng_A = 8.101857\8.2530387
: lat_B = 53.1479943\53.2450662\51.2082624
: lng_B = 7.8664467\6.8423613\8.2001266
```

To obtain a matrix of the distances between the two A and the three B locations, type:

```
: hav_dist(lat_A,lng_A,lat_B,lng_B)
                  1              2              3
    +----------------------------------------------+
  1 |   15.99849446    85.06021874    212.7400776  |
  2 |   26.55005308    94.00798103     222.129603  |
    +----------------------------------------------+
```

To obtain an all-to-all distance matrix for the B locations, type:

```
: hav_dist(lat_B,lng_B,lat_B,lng_B)
[symmetric]
                  1              2              3
    +----------------------------------------------+
  1 |             0                                |
  2 |   69.06623017              0                 |
  3 |   216.8843668    244.6240856              0  |
    +----------------------------------------------+
```

6.3 Mata Functions Indicating if Location is in Range

6.3.1 Syntax

real matrix `hav_inrange(`*real colvector lat_A*, *real colvector lng_A*, *real colvector lat_B*, *real colvector lng_B*, *real scalar range*) [, *real scalar radius*]

The syntax of `hav_inrangem` is analogue.

real colvector `hav_ninrange(`*real colvector lat_A*, *real colvector lng_A*, *real colvector lat_B*, *real colvector lng_B* , *real scalar range*) [, *real scalar radius*]

The syntax of `hav_ninrangem` is analogue.

6.3.2 Description

`hav_inrange(`*lat_A*, *lng_A*, *lat_B*, *lng_B*, *range*) returns a matrix of ones and zeros indicating which of the B locations are within a kilometre range specified by *range* of each A location. The argument *radius* is optional. The default is 6,371 km.

`hav_inrangem(`*lat_A*, *lng_A*, *lat_B*, *lng_B*, *range*) is analogue but for miles.

`hav_ninrange(...)` returns a column vector containing the number of B locations within a kilometre range specified by *range* of each A location. The argument *radius* is optional. The default is 6,371 km.

`hav_ninrangem(...)` is analogue but for miles.

6.3.3 Conformability

`hav_inrange(`*lat_A*, *lng_A*, *lat_B*, *lng_B*, *range*, *radius*)

input:

lat_A: $n \times 1$

lng_A: $n \times 1$

lat_B: $k \times 1$

lng_B: $k \times 1$

range: 1×1

radius: 1×1 (optional)

output:

result: $n \times k$

`hav_inrangem()` is analogue.

`hav_ninrange(`*lat_A*, *lng_A*, *lat_B*, *lng_B*, *range*, *radius*) *input*:

lat_A: $n \times 1$

lng_A: $n \times 1$

lat_B: $k \times 1$

lng_B: $k \times 1$

range: 1×1

radius: 1×1 (optional)

output:

result: $n \times 1$

`hav_ninrangem()` is analogue.

6.3.4 Examples

A matrix indicating which of the B locations lie within 90 km of each A location is obtained by typing:

```
: range = 90
: hav_inrange(lat_A,lng_A,lat_B,lng_B,range)
        1   2   3
    +-------------+
  1 |   1   1   0 |
  2 |   1   0   0 |
    +-------------+
```

To obtain a count of B locations which lie within the given range of each A location, we type:

```
: hav_ninrange(lat_A,lng_A,lat_B,lng_B,range)
        1
    +-----+
  1 |   2 |
  2 |   1 |
    +-----+
```

References

Picard, R. (2010a). GEODIST: Stata module to compute geographical distances. Statistical Software Components, Boston College Department of Economics.

Picard, R. (2010b). GEONEAR: Stata module to find nearest neighbors using geodetic distances. Statistical Software Components, Boston College Department of Economics.

Picard, R. (2015). GEOCIRCLES: Stata module to create circles defined by geographic coordinates. Statistical Software Components, Boston College Department of Economics.

van Brummelen, G. (2013). *Heavenly Mathematics: The Forgotten Art of Spherical Trigonometry*. Princeton University Press.

Zeigermann, L. (2025). LHAVERSINE: Stata module to compute great-circle distances with Haversine formula. Statistical Software Components, Boston College Department of Economics.

Chapter 7

Conclusion

This dissertation consists of five chapters. The first two illustrate the impact of relaxing the common global choice set assumption in favour of local choice sets reflecting the retailer landscape on demand estimation and retail merger screening results. The remaining three describe routines written for the statistical software package Stata.

Chapter 2, **Estimating Retail Demand and Price Elasticities with Local Choice Sets**, analyses the effect of relaxing the common - and often implicit - assumption of a global choice set in favour of more realistic local choice sets. Local choice sets reflect the retailer landscape by only considering products of retailers with a supermarket within a reasonable distance of the household. The results show that both the logit demand estimates and the resulting elasticities are severely biased if the differences in the local availability of retailers are ignored. This is particularly true for smaller retailers operating a small number of supermarkets.

Chapter 3, **Merger Screening with Local Choice Sets**, builds on the previous. It shows how the effects of the simplifying assumption of a global choice set for all consumers carry over to retail merger effects approximated by upward pricing pressure (UPP) for all possible mergers between the 13 retailers in the dataset. The results illustrate that UPP is severely affected by the choice set assumption, and that the potential merger effects are overestimated in some cases and underestimated in others when we ignore the local retailer landscape. However, the results also show that for mergers with potentially large anticompetitive effects, the relative difference in UPP between choice set specifications is fairly low.

Chapter 4, **Mixlelast: Stata Module for Mixed Logit Elasticities and Marginal Effects**, describes a post-estimation command for `mixlogit`, a user-written estimation command for mixed logit models discrete choice models. It enables the user to obtain various forms of elasticities and marginal effects, including standard errors and confidence intervals, without having to manually implement complex simulation techniques required for their computation.

Chapter 5, **Opencagegeo: Stata Module for Geocoding**, presents a routine for (reverse) geocoding. Geocoding is the process of converting an address into coordinates, i.e. latitudes and longitudes. Reverse geocoding is the process of converting latitudes and longitudes into addresses. The major advantage of this routine over available alternatives is the very flexible terms of the web-geocoding service offered by OpenCageData, which `opencagegeo` accesses via an application programming interface (API).

Chapter 6, **Lhaversine: Mata Library for Distance Functions**, describes a library of distance functions for Stata's matrix language. All functions use the haversine formula for great circle distances. They allow the user to obtain a distance matrix between many locations, and indicators of how many and which locations are within a specified range. `hlaversine` takes latitudes and longitudes as inputs.

Summing up, the contribution of this dissertation is twofold. The Stata routines provide researchers and practitioners with useful tools for making full use of geospacial data and for computing various forms of mixed logit elasticities and marginal effects (including standard errors), for which advanced programming skills would otherwise be required. The chapters on retail demand and supply with local choice sets serve as a warning, encouraging both researchers and practitioners to carefully consider how choice sets should be constructed within the context of (grocery) retailing. At the same time, the results are also reassuring for those who have relied on the common global choice set assumption in their merger analysis. For critical mergers that could potentially lead to significant anticompetitive effect, the two choice set specifications produce fairly similar UPP values. Large differences between choice set specifications only arise for potential mergers that do not raise serious concerns.

However, this final result critically hinges on the additional, often implicit, assumption that retailers set uniform prices that do not reflect the local retail structure. If this does not hold, the above result is no longer valid, raising a new question for further research: How can we advance our structural models of demand and supply to reflect local price setting?

Heinrich Heine
Universität
Düsseldorf

Eidesstattliche Versicherung

Ich, Lars Marten Zeigermann, versichere an Eides statt, dass die vorliegende Dissertation von mir selbstständig und ohne unzulässige fremde Hilfe unter Beachtung der Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf erstellt worden ist.

Bonn, der 5. September 2025

Unterschrift